Chapter 9
Consciousness, Decision-Making and Neural Computation

Edmund T. Rolls

Abstract Computational processes that are closely related to conscious processing and reasoning are described. Evidence is reviewed that there are two routes to action, the explicit, conscious, one involving reasoning, and an implicit, unconscious route for well-learned actions to obtain goals. Then a higher order syntactic thought (HOST) computational theory of consciousness is described. It is argued that the adaptive value of higher order syntactic thoughts is to solve the credit assignment problem that arises if a multi-step syntactic plan needs to be corrected. It is then suggested that it feels like something to be an organism that can think about its own linguistic and semantically based thoughts. It is suggested that qualia, raw sensory and emotional feels, arise secondarily to having evolved such a HOST processing system, and that sensory and emotional processing feels like something because it would be unparsimonious for it to enter the planning, HOST system and *not* feel like something. Neurally plausible models of decision-making are described, which are based on noise-driven and therefore probabilistic integrate-and-fire attractor neural networks, and it is proposed that networks of this type are involved when decisions are made between the explicit and implicit routes to action. This raises interesting issues about free will. It has been argued that the confidence one has in one's decisions provides an objective measure of awareness, but it is shown that two coupled attractor networks can account for decisions based on confidence estimates from previous decisions. In analyses of the implementation of consciousness, it is shown that the threshold for access to the consciousness system is higher than that for producing behavioural responses. The adaptive value of this may be that the systems in the brain that implement the type of information processing involved in conscious thoughts are not interrupted by small signals that could be noise in sensory pathways. Then oscillations are argued to not be a necessary part of the implementation of consciousness in the brain.

E.T. Rolls (✉)
Oxford Centre for Computational Neuroscience, Oxford, UK
e-mail: Edmund.Rolls@oxcns.org

V. Cutsuridis et al. (eds.), *Perception-Action Cycle: Models, Architectures, and Hardware*, Springer Series in Cognitive and Neural Systems 1,
DOI 10.1007/978-1-4419-1452-1_9,

9.1 Introduction

In the perception–reason–action cycle, the reasoning step may produce an error. For example, in a reasoned plan with several steps to the plan, if there is an error, how do we know which step has the error? There is a credit assignment problem here, for the occurrence of an error does not tell us which step had a fault. I argue that in this situation, thoughts about the steps of the plan, that is thoughts about thoughts, namely "higher order thoughts," can help us to detect which was the weak step in the plan, with perhaps weak premises, so that we can correct the plan and try again. I argue that having thoughts about our previous thoughts, reflecting on them, may be a computational process that, when it occurs, is accompanied by feelings of consciousness, that it feels like something, and this general approach to the phenomenal aspects of consciousness is shared by a number of philosophers and scientists (Rosenthal 1990, 1993, 2004, 2005; Weiskrantz 1997). I then go on to argue that when this "higher order syntactic thought" (HOST) brain processor is actively processing sensory states or emotions (about which we may sometimes need to reason), then conscious feelings about these sensory states or emotions, called qualia, are present. I argue that this reasoning (i.e. rational) explicit (conscious) system is propositional and uses syntactic binding of symbols.

I contrast this rational or reasoning, explicit, system with an implicit (unconscious) system that uses gene-specified goals [e.g. food reward when hungry, water when thirsty, social interaction (Rolls 2005b, 2011)] for actions and can perform arbitrary actions to obtain the genotypically defined goals. The explicit system, because of the reasoning, can perform multiple-step planning which might lead to goals that are in the interest of the phenotype but not of the genotype (e.g. not having children in order to devote oneself to the arts, philosophy, or science). I argue that when decisions are made between the implicit and explicit computational systems (the genotype vs. the phenotype), then noise produced by neuronal spiking can influence the decision-making in an attractor network. Decisions made that are based on our subjective confidence in our earlier decisions are also influenced by noise in the brain (Rolls and Deco 2010). The computations involved in the implicit vs. the explicit system, and the effects of noise in decision-making, raise the issues of free will and of determinism. Finally, I consider some related computational issues, such as the role of oscillations and stimulus-dependent synchrony in consciousness, and why the threshold for consciousness is set to be at a higher level than the threshold for sensory processing and some implicit behavioural responses.

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 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, and many approaches have been described (Dennett 1991, 2005; Chalmers 1996;

Block 2005; Rosenthal 2005; Davies 2008). They are at the heart of the problem of consciousness, for why it should feel like something at all is the great mystery. Other aspects of consciousness, such as the fact that often when we "pay attention" to events in the world, we can process those events in some better way, that is process or access as opposed to phenomenal aspects of consciousness, may be easier to analyze (Allport 1988; Block 1995; Chalmers 1996).

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 what follows in this paper is to address the issue of consciousness, especially of qualia, what is written cannot be regarded as being establishable by the normal methods of scientific enquiry. Accordingly, I emphasize that the 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 should not be taken to have practical implications.

9.2 A Higher Order Syntactic Thought Theory of Consciousness

9.2.1 Multiple Routes to Action

A starting point is that much perception and action can be performed relatively automatically without apparent conscious intervention. An example sometimes given is driving a car. Another example is the identification of a visual stimulus that can occur without conscious awareness as described in Sect. 9.6. Another example is much of the sensory processing and actions that involve the dorsal stream of visual processing to the parietal cortex, such as posting a letter through a box at the correct orientation even when one may not be aware of what the object is (Milner and Goodale 1995; Goodale 2004; Milner 2008). Another example is blindsight, in which humans with damage to the visual cortex may be able to point to objects even when they are not aware of seeing an object (Weiskrantz 1997, 1998). Similar evidence applies to emotions, some of the processing for which can occur without conscious awareness (De Gelder et al. 1999; Phelps and LeDoux 2005; LeDoux 2008; Rolls 2005b, 2008a, b). Further, there is evidence that split-brain patients may not be aware of actions being performed by the "non-dominant" hemisphere (Gazzaniga and LeDoux 1978; Gazzaniga 1988, 1995; Cooney and Gazzaniga 2003). Further evidence consistent with multiple including non-conscious routes to action is that patients with focal brain damage, for example to the prefrontal cortex, may perform actions, yet comment verbally that they should not be performing those actions (Rolls et al. 1994a; Rolls 1999a, 2005b; Hornak et al. 2003, 2004). The actions, which appear to be performed implicitly, with surprise expressed later by the explicit system, include making behavioural responses to a no-longer rewarded visual stimulus in a visual discrimination reversal

(Rolls et al. 1994a; Hornak et al. 2004). 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 and LeDoux 1978; Gazzaniga 1988; Rolls et al. 1994a; Gazzaniga 1995; Rolls 2005b; LeDoux 2008).

This evidence (see further Sect. 9.3.1) suggests that there are multiple routes to action, only some of which involve conscious processing (Rolls 2005a, b). It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region such as those involved in some types of implicit 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. 9.1). (It is of interest that mathematicians may get a hunch that something

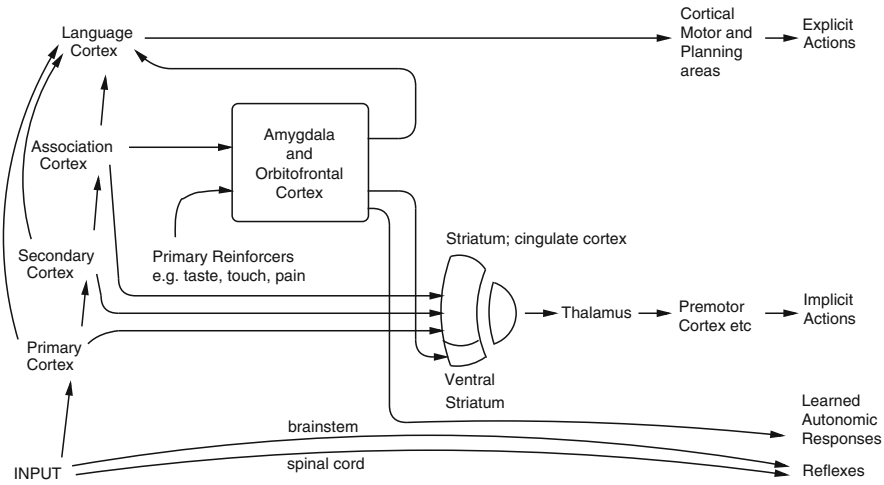

Fig. 9.1 Dual routes to the initiation of action in response to rewarding and punishing stimuli. The inputs from different sensory systems to brain structures such as the orbitofrontal cortex and amygdala allow these brain structures to evaluate the reward- or punishment-related value of incoming stimuli or of remembered stimuli. The different sensory inputs enable evaluations within the orbitofrontal cortex and amygdala based mainly on the primary (unlearned) reinforcement value for taste, touch and olfactory stimuli and on the secondary (learned) reinforcement value for visual and auditory stimuli. In the case of vision, the "association cortex" which outputs representations of objects to the amygdala and orbitofrontal cortex is the inferior temporal visual cortex. One route for the outputs from these evaluative brain structures is via projections directly to structures such as the basal ganglia (including the striatum and ventral striatum) to enable implicit, direct behavioural responses based on the reward or punishment-related evaluation of the stimuli to be made. The second route is via the language systems of the brain, which allow explicit decisions involving multi-step syntactic planning to be implemented

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 an indication of a close association between linguistic processing and consciousness. The linguistic processing need not involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements and to perform syntactic operations on the semantic representations of states in the world would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g. I think that he thinks that...), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others.[1] This capability for HOSTs 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. [For an earlier view which is close to this part of the argument see Humphrey (1980).] The examples of behaviour from non-humans that may reflect planning may reflect much more limited and inflexible planning. For example, the dance of the honey-bee to signal to other bees the location of food may be said to reflect planning, but the symbol manipulation is not arbitrary. There are likely to be interesting examples of non-human primate behaviour, perhaps in the great apes, that reflect the evolution of an arbitrary symbol-manipulation system that could be useful for flexible planning, cf. Cheney and Seyfarth (1990). It is important to state that the language ability referred to here is not necessarily human verbal language (though this would be an example). What it is suggested is important to planning is the syntactic manipulation of symbols, and it is this syntactic manipulation of symbols which is the sense in which language is defined and used here.

I understand *reasoning*, and *rationality*, to involve syntactic manipulations of symbols in the way just described. Reasoning thus typically may involve multiple steps of "if. then" conditional statements, all executed as a one-off or one-time

[1] Second order thoughts are thoughts about thoughts. Higher order thoughts refer to second order, third order etc. thoughts about thoughts... (A thought may be defined briefly as an intentional mental state, that is a mental state that is about something. Thoughts include beliefs, and are usually described as being propositional (Rosenthal DM (2005) Consciousness and Mind. Oxford: Oxford University Press). An example of a thought is "It is raining". A more detailed definition is as follows. A thought may be defined as an occurrent mental state (or event) that is intentional - that is a mental state that is about something - and also propositional, so that it is evaluable as true or false. Thoughts include occurrent beliefs or judgements. An example of a thought would be an occurrent belief that the earth moves around the sun/ that Maurice's boat goes faster with two sails/ that it never rains in southern California.)

process (see below), and is very different from associatively learned conditional rules typically learned over many trials, such as "if yellow, a left choice is associated with reward".

9.2.2 A Computational Hypothesis of Consciousness

It is next suggested that this arbitrary symbol-manipulation using important aspects of language processing and used for planning but not in initiating all types of behaviour is close to what consciousness is about. In particular, consciousness may *be* the state which arises in a system that can think about (or reflect on) its own (or other peoples') thoughts, that is in a system capable of second or higher order thoughts (HOTs) (Rosenthal 1986, 1990, 1993, 2004, 2005; Dennett 1991; Rolls 1995, 1997a, b, 1999b, 2004b, 2005b, 2007c, 2008a; Carruthers 1996; Gennaro 2004). 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 HOSTs 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 (and mine) 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 which 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*, which amounts effectively for him in humans to saying that consciousness is what we can report (verbally) about.] 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 HOSTs 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 processing is being performed by this particular neural system that is capable of second or HOSTs.

It might be objected that this captures some of the process aspects of consciousness, what is being performed in the relevant 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 HOSTs to be thinking about its own first or lower order thoughts. But the return point (discussed further below) 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 HOSTs 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.)

As a point of clarification, I note that according to this theory, a language processing system (let alone a working memory, LeDoux 2008) is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have HOSTs, and a first order language processor (that might be perfectly competent at language) would not be conscious, in that it could not think about its own or others' thoughts. One can perfectly well conceive of a system that obeyed the rules of language (which is the aim of some 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.]

A 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 (or one-time) 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.

9.2.3 Adaptive Value of Processing in the System That Is Related to Consciousness

It is suggested that part of the evolutionary *adaptive significance* of this type of HOST 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 is shown elsewhere that the hippocampus may be a system for such "declarative" recall of recent memories (Rolls 2008b). Its close relation to "conscious" processing in humans [Squire and Zola (1996) have 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 (Rolls and Kesner 2006; Rolls 2008b). Another part of the adaptive value of a HOST

system may be that by thinking about its own thoughts in a given situation, it may be able to better understand the thoughts of another individual in a similar situation and therefore predict that individual's behaviour better (cf. Humphrey 1980, 1986; Barlow 1997).

In line with the argument on the adaptive value of HOSTs and thus consciousness given above, which they are useful for correcting lower order thoughts, I now suggest that correction using HOSTs 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, "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. It is partly because complex lower order thoughts such as these which involve syntax and language would benefit from correction by HOSTs that I suggest that there is a close link between this reflective consciousness and language. The *computational hypothesis* is that by thinking about lower order thoughts, the HOSTs can discover what may be weak links in the chain of reasoning at the lower order level, and having detected the weak link, might alter the plan, to see if this gives better success. In our example above, if it transpired that C could not do Z, how might the plan have failed? Instead of having to go through endless random changes to the plan to see if by trial and error some combination does happen to produce results, what I am suggesting is that by thinking about the previous plan, one might, for example using knowledge of the situation and the probabilities that operate in it, guess that the step where the plan failed was that B did not in fact do Y. So by thinking about the plan (the first or lower order thought), one might correct the original plan, in such a way that the weak link in that chain, that "B will probably do Y", is circumvented.

I draw a parallel with neural networks: there is a "*credit assignment*" problem in such multi-step 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 connections; see Rumelhart et al. (1986), Rolls and Deco (2002) and Rolls (2008b).] The suggestion is that this is the function of HOSTs and is why systems with HOSTs 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.

9.2.4 Symbol Grounding

A further point in the argument should be emphasized for clarity. The system that is having syntactic thoughts about its own syntactic thoughts (higher order syntactic thoughts or HOSTs) would have to have its symbols grounded in the real world for it to feel like something to be having HOSTs. 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 by Rolls (2005b) . The symbols (or symbolic representations) are symbols in the sense that they can take part in syntactic processing. The symbolic representations are grounded in the world in that they refer to events in the world. The symbolic representations must have a great deal of information about what is referred to in the world, including the quality and intensity of sensory events, emotional states, etc. The need for this is that the reasoning in the symbolic system must be about stimuli, events and states, and remembered stimuli, events and states, and for the reasoning to be correct, all the information that can affect the reasoning must be represented in the symbolic system, including, for example, just how light or strong the touch was, etc. Indeed, it is pointed out in *Emotion Explained* (Rolls 2005b) that it is no accident that the shape of the multidimensional phenomenal (sensory, etc.) space does map so clearly onto the space defined by neuronal activity in sensory systems, for if this were not the case, reasoning about the state of affairs in the world would not map onto the world and would not be useful. Good examples of this close correspondence are found in the taste system, in which subjective space maps simply onto the multi-dimensional space represented by neuronal firing in primate cortical taste areas. In particular, if a three-dimensional space reflecting the distances between the representations of different tastes provided by macaque neurons in the cortical taste areas is constructed, then the distances between the subjective ratings by humans of different tastes are very similar (Yaxley et al. 1990; Smith-Swintosky et al. 1991; Kadohisa et al. 2005). Similarly, the changes in human subjective ratings of the pleasantness of the taste, smell and sight of food parallel very closely the responses of neurons in the macaque orbitofrontal cortex (see *Emotion Explained*).

The representations in the first order linguistic processor that the HOSTs process include beliefs (e.g. “Food is available”, or at least representations of this), and the HOST system would then have available to it the concept of a thought [so that it could represent “I believe (or there is a belief) that food is available”]. However, as argued by Rolls (1999b, 2005b), representations of sensory processes and emotional states must be processed by the first order linguistic system, and HOSTs may be about these representations of sensory processes and emotional states capable of taking part in the syntactic operations of the first order

linguistic processor. Such sensory and emotional information may reach the first order linguistic system from many parts of the brain, including those such as the orbitofrontal cortex and amygdala implicated in emotional states (see Fig. 9.1 and *Emotion Explained,* Fig. 10.3). When the sensory information is about the identity of the taste, the inputs to the first order linguistic system must come from the primary taste cortex, in that the identity of taste, independently of its pleasantness (in that the representation is independent of hunger), must come from the primary taste cortex. In contrast, when the information that reaches the first order linguistic system is about the pleasantness of taste, it must come from the secondary taste cortex, in that there the representation of taste depends on hunger (Rolls and Grabenhorst 2008).

9.2.5 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. 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 used (or are available to be used) in this type of (linguistically based) HOST processing, and this is what I propose.

The explanation for perceptual, emotional and motivational subjective feelings or qualia that this discussion has led towards is thus that they should be felt as conscious because they enter into a specialized linguistic symbol-manipulation system that is part of a HOST 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 HOST system) becoming felt qualia. The qualia are thus accounted for by the evolution of the linguistic system that can reflect on and correct its own lower order processes and thus has adaptive value.

This account implies that it may be especially animals with a higher order belief and thought system and with linguistic symbol manipulation that have qualia. It may be that much non-human animal behaviour, provided that it does not require flexible linguistic planning and correction by reflection, could take place according

to reinforcement-guidance [using e.g. stimulus-reinforcement association learning in the amygdala and orbitofrontal cortex (Rolls 2004a, 2005b, 2008b)] and rule-following [implemented e.g. using habit or stimulus-response learning in the basal ganglia (Rolls 2005b)]. Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals 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.

It is of interest to comment on how the evolution of a system for flexible planning might affect emotions. Consider grief which may occur when a reward is terminated and no immediate action is possible [see Rolls (1990, 2005b)]. It may be adaptive by leading to a cessation of the formerly rewarded behaviour and thus facilitating the possible identification of other positive reinforcers in the environment. In humans, grief may be particularly potent because it becomes represented in a system which can plan ahead and understand the enduring implications of the loss. (Thinking about or verbally discussing emotional states may also in these circumstances help, because this can lead towards the identification of new or alternative reinforcers and of the realization that, for example, negative consequences may not be as bad as feared.)

9.2.6 Pathways

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 (see Fig. 9.1). 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 2005b, 2006). Now the quality of the tastant (sweet, salt, etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety (Rolls 2005b). The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex) and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it).

The same argument arises for reward in general and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is reflected in the responses of neurons to visual stimuli (Rolls 2005b, 2006). 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 2005b, 2006). I suggest that this correspondence arises because pleasure is the subjective state that represents in the conscious system a signal that is positively reinforcing (rewarding), and that inconsistent behaviour would result if the representations did not correspond to a signal for positive reinforcement in both the conscious and the non-conscious processing systems.

Do these arguments mean that the conscious sensation of, for example, taste quality (i.e. identity and intensity) is represented or occurs in the primary taste cortex and of the pleasantness of taste in the secondary taste cortex, and that activity in these areas is sufficient for conscious sensations (qualia) to occur? I do not suggest this at all. Instead the arguments I have put forward above suggest that we are only conscious of representations when engage a system capable of HOSTs. The implication then is that pathways must connect from each of the brain areas in which information is represented about which we can be conscious to the system that has the HOSTs, which as I have argued above requires a brain system capable of HOSTs. 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 2005b). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have HOSTs about the pleasantness of the representation in the secondary taste cortex. There would also need to be pathways from the hippocampus, implicated in the recall of declarative memories, back to the language areas of the cerebral cortex (at least via the cortical areas which receive backprojections from the amygdala, orbitofrontal cortex and hippocampus, see Fig. 9.1, which would in turn need connections to the language areas).

9.2.7 Consciousness and Causality

One question that has been discussed is whether there is a causal role for consciousness [e.g. Armstrong and Malcolm (1984)]. The position to which the above arguments lead is that indeed conscious processing does have a causal role in the elicitation of behaviour, but only under the set of circumstances when HOSTs 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 HOST is causally involved in correcting the lower order thought, and that it is a property of the HOST 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 (see further Rolls (2003, 2005a) for relations between implicit and explicit processing). There are many brain processing routes to output regions, and only one of these involves conscious, verbally represented processing which can later be recalled (see Fig. 9.1).

I suggest that these concepts may help us to understand what is happening in experiments of the type described by Libet and many others (Libet 2002), in which consciousness appears to follow with a measurable latency the time when a decision was taken. This is what I predict, if the decision is being made by an implicit perhaps reward/emotion or habit-related process, for then the conscious processor confabulates an account of or commentary on the decision, so that inevitably the conscious account follows the decision. On the other hand, I predict that if the rational (multi-step, reasoning) route is involved in taking the decision, as it might be during planning, or a multi-step task such as mental arithmetic, then the conscious report of when the decision was taken, and behavioural or other objective evidence on when the decision was taken, would correspond much more. Under those circumstances, the brain processing taking the decision would be closely related to consciousness, and it would not be a case of just confabulating or reporting on a decision taken by an implicit processor. It would be of interest to test this hypothesis in a version of Libet's task (Libet 2002) in which reasoning was required. The concept that the rational, conscious, processor is only in some tasks involved in taking decisions is extended further in the section on dual routes to action below.

9.2.8 Consciousness, a Computational System for Higher Order Syntactic Manipulation of Symbols, and a Commentary or Reporting Functionality

I now consider some clarifications of the present proposal, and how it deals with some issues that arise when considering theories of the phenomenal aspects of consciousness.

First, the present proposal has as its foundation the type of computation that is being performed and suggests that it is a property of a HOST system used for correcting multi-step plans with its representations grounded in the world that it would feel like something for a system to be doing this type of processing. To do this type of processing, the system would have to be able to recall previous multi-step plans and would require syntax to keep the symbols in each step of the plan separate. In a sense, the system would have to be able to recall and take into consideration its earlier multi-step plans, and in this sense *report* to itself, on those earlier plans. Some approaches to consciousness take the ability to report on or make a *commentary* on events as being an important marker for consciousness (Weiskrantz 1997), and the computational approach I propose suggests why there should be a close relation between consciousness and the ability to report or provide a commentary, for the ability to report is involved in using HOSTs to correct a multi-step plan.

Second, the implication of the present approach is that the type of linguistic processing or reporting need not be verbal, using natural language, for what is required to correct the plan is the ability to manipulate symbols syntactically, and this could be implemented in a much simpler type of mentalese or syntactic system (Fodor 1994; Jackendoff 2002; Rolls 2004b) than verbal language or natural language which implies a universal grammar.

Third, this approach to consciousness suggests that the information must be being processed in a system capable of implementing HOSTs for the information to be conscious and in this sense is more specific than global workspace hypotheses (Baars 1988; Dehaene and Naccache 2001; Dehaene et al. 2006). Indeed, the present approach suggests that a workspace could be sufficiently global to enable even the complex processing involved in driving a car to be performed, and yet the processing might be performed unconsciously, unless HOST (supervisory, monitory, correcting) processing was involved.

Fourth, the present approach suggests that it just is a property of HOST computational processing with the representations grounded in the world that it feels like something. There is to some extent an element of mystery about why it feels like something, why it is phenomenal, but the explanatory gap does not seem so large when one holds that the system is recalling, reporting on, reflecting on and reorganizing information about itself in the world in order to prepare new or revised plans. In terms of the physicalist debate (see for a review Davies 2008), an important aspect of my proposal is that it is a *necessary* property of this type of (HOST) computational processing that it feels like something (the philosophical description is that this is an absolute metaphysical necessity), and given this view, then it is up to one to decide whether this view is consistent with one's particular view of physicalism or not (Rolls 2008a). Similarly, the possibility of a zombie is inconsistent with the present hypothesis, which proposes that it is by virtue of performing processing in a specialized system that can perform higher order syntactic processing with the representations grounded in the world that phenomenal consciousness is necessarily present.

An implication of these points is that my theory of consciousness is a computational theory. It argues that it is a property of a certain type of computational processing that it feels like something. In this sense, although the theory spans many levels from the neuronal to the computational, it is unlikely that any particular neuronal phenomena such as oscillations are necessary for consciousness, unless such computational processes happen to rely on some particular neuronal properties not involved in other neural computations but necessary for higher order syntactic computations. It is these computations and the system that implements them that this computational theory argues are necessary for consciousness.

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.

9.3 Selection Between Conscious vs. Unconscious Decision-Making and Free Will

9.3.1 Dual Routes to Action

According to the present formulation, there are two types of route to action performed in relation to reward or punishment in humans (see also Rolls 2003, 2005b). Examples of such actions include emotional and motivational behaviour.

The first route is via the brain systems that have been present in non-human primates such as monkeys, and to some extent in other mammals, for millions of years. These systems include the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. These systems control behaviour in relation to previous associations of stimuli with reinforcement. The computation which controls the action thus involves assessment of the reinforcement-related value of a stimulus. This assessment may be based on a number of different factors. One is the previous reinforcement history, which involves stimulus-reinforcement association learning using the amygdala, and its rapid updating especially in primates using the orbitofrontal cortex. This stimulus-reinforcement association learning may involve quite specific information about a stimulus, for example of the energy associated with each type of food, by the process of conditioned appetite and satiety (Booth 1985). A second is the current motivational state, for example whether hunger is present, whether other needs are satisfied, etc. A third factor which affects the computed reward value of the stimulus is whether that reward has been received recently. If it has been received recently but in small quantity, this may increase the reward value of the stimulus. This is known as incentive motivation or the "salted peanut" phenomenon. The adaptive value of such a process is that this positive feedback of reward value in the early stages of working for a particular reward tends to lock the organism onto behaviour being performed for that reward. This means that animals that are, for example, almost equally hungry and thirsty will show hysteresis in their choice of action rather than continually switching from eating to drinking and back with each mouthful of water or food. This introduction of hysteresis into the reward evaluation system makes action selection a much more efficient process in a natural environment, for constantly switching between different types of behaviour would be very costly if all the different rewards were not available in the same place at the same time. (For example, walking half a mile between a site where water was available and a site where food was available after every mouthful would be very inefficient.) The amygdala is one structure that may be involved in this increase in the reward value of stimuli early on in a series of presentations, in that lesions of the amygdala (in rats) abolish the expression of this reward incrementing process which is normally evident in the increasing rate of working for a food reward early on in a meal (Rolls 2005b). A fourth factor is the computed absolute value of the reward or punishment expected or being obtained from a stimulus, e.g. the sweetness of the stimulus (set by evolution so that sweet stimuli will tend to be rewarding, because they are generally associated with energy sources), or the pleasantness of

touch (set by evolution to be pleasant according to the extent to which it brings animals of the opposite sex together, and depending on the investment in time that the partner is willing to put into making the touch pleasurable, a sign which indicates the commitment and value for the partner of the relationship). After the reward value of the stimulus has been assessed in these ways, behaviour is then initiated based on approach towards or withdrawal from the stimulus. A critical aspect of the behaviour produced by this type of system is that it is aimed directly towards obtaining a sensed or expected reward by virtue of connections to brain systems such as the basal ganglia and cingulate cortex (Rolls 2009), which are concerned with the initiation of actions (see Fig. 9.1). The expectation may of course involve behaviour to obtain stimuli associated with reward, which might even be present in a chain.

Now part of the way in which the behaviour is controlled with this first route is according to the reward value of the outcome. At the same time, the animal may only work for the reward if the cost is not too high. Indeed, in the field of behavioural ecology, animals are often thought of as performing optimally on some cost-benefit curve [see e.g. Krebs and Kacelnik (1991)]. This does not at all mean that the animal thinks about the rewards and performs a cost-benefit analysis using a lot of thoughts about the costs, other rewards available and their costs, etc. Instead, it should be taken to mean that in evolution, the system has evolved in such a way that the way in which the reward varies with the different energy densities or amounts of food and the delay before it is received can be used as part of the input to a mechanism, which has also been built to track the costs of obtaining the food (e.g. energy loss in obtaining it, risk of predation, etc.), and to then select given many such types of reward and the associated cost, the current behaviour that provides the most "net reward". Part of the value of having the computation expressed in this reward-minus-cost form is that there is then a suitable "currency", or net reward value, to enable the animal to select the behaviour with currently the most net reward gain (or minimal aversive outcome).

The second route in humans involves a computation with many "if. . . then" statements to implement a plan to obtain a reward. In this case, the reward may actually be *deferred* as part of the plan, which might involve working first to obtain one reward, and only then to work for a second more highly valued reward, if this was thought to be overall an optimal strategy in terms of resource usage (e.g. time). In this case, syntax is required, because the many symbols (e.g. names of people) that are part of the plan must be correctly linked or bound. Such linking might be of the form: "if A does this, then B is likely to do this, and this will cause C to do this. . . ". The requirement of syntax for this type of planning implies that an output to language systems in the brain is required for this type of planning (see Fig. 9.1). Thus the explicit language system in humans may allow working for deferred rewards by enabling use of a one-off, individual, plan appropriate for each situation. Another building block for such planning operations in the brain may be the type of short-term memory in which the prefrontal cortex is involved. This short term memory may be, for example, in non-human primates of where in space a response has just been made. A development of this type of short term response memory system in humans to enable multiple short term memories to be held in place correctly,

preferably with the temporal order of the different items in the short term memory coded correctly, may be another building block for the multiple step "if. . . . then" type of computation in order to form a multiple step plan. Such short term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of non-human primates and humans (Goldman-Rakic 1996; Petrides 1996; Rolls 2008b) and may be part of the reason why prefrontal cortex damage impairs planning (Shallice and Burgess 1996).

Of these two routes (see Fig. 9.1), it is the second which I have suggested above is related to consciousness. The hypothesis is that consciousness is the state which arises by virtue of having the ability to think about one's own thoughts, which has the adaptive value of enabling one to correct long multi-step syntactic plans. This latter system is thus the one in which explicit, declarative, processing occurs. Processing in this system is frequently associated with reason and rationality, in that many of the consequences of possible actions can be taken into account. The actual computation of how rewarding a particular stimulus or situation is or will be probably still depends on activity in the orbitofrontal and amygdala, as the reward value of stimuli is computed and represented in these regions, and in that it is found that verbalized expressions of the reward (or punishment) value of stimuli are dampened by damage to these systems. (For example, damage to the orbitofrontal cortex renders painful input still identifiable as pain, but without the strong affective, "unpleasant", reaction to it.) This language system which enables long-term planning may be contrasted with the first system in which behaviour is directed at obtaining the stimulus (including the remembered stimulus) which is currently most rewarding, as computed by brain structures that include the orbitofrontal cortex and amygdala. There are outputs from this system, perhaps those directed at the basal ganglia, which do not pass through the language system, and behaviour produced in this way is described as implicit, and verbal declarations cannot be made directly about the reasons for the choice made. When verbal declarations are made about decisions made in this first system, those verbal declarations may be confabulations, reasonable explanations or fabrications, of reasons why the choice was made. These reasonable explanations would be generated to be consistent with the sense of continuity and self that is a characteristic of reasoning in the language system.

The question then arises of how decisions are made in animals such as humans that have both the implicit, direct reward-based, and the explicit, rational, planning systems (see Fig. 9.1) (Rolls 2008b). One particular situation in which the first, implicit, system may be especially important is when rapid reactions to stimuli with reward or punishment value must be made, for then the direct connections from structures such as the orbitofrontal cortex to the basal ganglia may allow rapid actions (Rolls 2005b). Another is when there may be too many factors to be taken into account easily by the explicit, rational, planning, system, when the implicit system may be used to guide action. In contrast, when the implicit system continually makes errors, it would then be beneficial for the organism to switch from automatic, direct, action based on obtaining what the orbitofrontal cortex system decodes as being the most positively reinforcing choice currently available to the explicit conscious control system which can evaluate with its long-term planning algorithms

what action should be performed next. Indeed, it would be adaptive for the explicit system to regularly be assessing performance by the more automatic system and to switch itself in to control behaviour quite frequently, as otherwise the adaptive value of having the explicit system would be less than optimal.

There may also be a flow of influence from the explicit, verbal system to the implicit system, in that the explicit system may decide on a plan of action or strategy, and exert an influence on the implicit system which will alter the reinforcement evaluations made by and the signals produced by the implicit system (Rolls 2005b).

It may be expected that there is often a conflict between these systems, in that the first, implicit, system is able to guide behaviour particularly to obtain the greatest immediate reinforcement, whereas the explicit system can potentially enable immediate rewards to be deferred and longer-term, multi-step, plans to be formed. This type of conflict will occur in animals with a syntactic planning ability, that is in humans and any other animals that have the ability to process a series of "if…then" stages of planning. This is a property of the human language system, and the extent to which it is a property of non-human primates is not yet fully clear. In any case, such conflict may be an important aspect of the operation of at least the human mind, because it is so essential for humans to correctly decide, at every moment, whether to invest in a relationship or a group that may offer long-term benefits or whether to directly pursue immediate benefits (Rolls 2005b, 2008b).

The thrust of the argument (Rolls 2005b, 2008b) thus is that much complex animal including human behaviour can take place using the implicit, non-conscious, route to action. We should be very careful not to postulate intentional states (i.e. states with intentions, beliefs and desires) unless the evidence for them is strong, and it seems to me that a flexible, one-off, linguistic processing system that can handle propositions is needed for intentional states. What the explicit, linguistic, system does allow is exactly this flexible, one-off, multi-step planning ahead type of computation, which allows us to defer immediate rewards based on such a plan.

This discussion of dual routes to action has been with respect to the behaviour produced. There is of course in addition a third output of brain regions, such as the orbitofrontal cortex and amygdala involved in emotion, that is directed to producing autonomic and endocrine responses (see Fig. 9.1). Although it has been argued by Rolls (2005b) that the autonomic system is not normally in a circuit through which behavioural responses are produced (i.e. against the James–Lange and related somatic theories), there may be some influence from effects produced through the endocrine system (and possibly the autonomic system, through which some endocrine responses are controlled) on behaviour or on the dual systems just discussed that control behaviour.

9.3.2 The Selfish Gene vs. the Selfish Phene

I have provided evidence in Sect. 9.3.1 that there are two main routes to decision-making and action. The first route selects actions by gene-defined goals for action

and is closely associated with emotion. The second route involves multi-step planning and reasoning which requires syntactic processing to keep the symbols involved at each step separate from the symbols in different steps. (This second route is used by humans and perhaps by closely related animals.) Now the "interests" of the first and second routes to decision-making and action are different. As argued very convincingly by Richard Dawkins in The Selfish Gene (Dawkins 1989), and by others (Hamilton 1964, 1996; Ridley 1993), many behaviours occur in the interests of the survival of the genes, not of the individual (nor of the group), and much behaviour can be understood in this way. I have extended this approach by arguing that an important role for some genes in evolution is to define the goals for actions that will lead to better survival of those genes; that emotions are the states associated with these gene-defined goals; and that the defining of goals for actions rather that actions themselves is an efficient way for genes to operate, as it leaves flexibility of choice of action open until the animal is alive (Rolls 2005b). This provides great simplification of the genotype as action details do not need to be specified, just rewarding and punishing stimuli, and also flexibility of action in the face of changing environments faced by the genes. Thus the interests that are implied when the first route to action is chosen are those of the "selfish genes" and not those of the individual.

However, the second route to action allows, by reasoning, decisions to be taken that might not be in the interests of the genes, might be longer term decisions and might be in the interests of the individual. An example might be a choice not to have children, but instead to devote oneself to science, medicine, music or literature. The reasoning, rational, system presumably evolved because taking longer-term decisions involving planning rather than choosing a gene-defined goal might be advantageous at least sometimes for genes. But an unforeseen consequence of the evolution of the rational system might be that the decisions would, sometimes, not be to the advantage of any genes in the organism. After all, evolution by natural selection operates utilizing genetic variation like a Blind Watchmaker (Dawkins 1986). In this sense, the interests when the second route to decision-making is used are at least sometimes those of the "selfish phenotype". Indeed, we might euphonically say that the interests are those of the "selfish phene" (where the etymology is Gk phaino, "appear", referring to appearance, hence the thing that one observes, the individual). Hence the decision-making referred to in Sect. 9.3.1 is between a first system where the goals are gene-defined and a second rational system in which the decisions may be made in the interests of the genes, or in the interests of the phenotype and not in the interests of the genes. Thus we may speak of the choice as sometimes being between the "Selfish Genes" and the "Selfish Phenes".

Now what keeps the decision-making between the "Selfish Genes" and the "Selfish Phenes" more or less under control and in balance? If the second, rational, system chose too often for the interests of the "Selfish Phene", the genes in that phenotype would not survive over generations. Having these two systems in the same individual will only be stable if their potency is approximately equal, so that sometimes decisions are made with the first route and sometimes with the second route. If the two types of decision-making, then, compete with approximately equal

potency, and sometimes one is chosen, and sometimes the other, then this is exactly the scenario in which stochastic processes in the decision-making mechanism are likely to play an important role in the decision that is taken. The same decision, even with the same evidence, may not be taken each time a decision is made, because of noise in the system.

The system itself may have some properties that help to keep the system operating well. One is that if the second, rational, system tends to dominate the decision-making too much, the first, gene-based emotional system might fight back over generations of selection and enhance the magnitude of the reward value specified by the genes, so that emotions might actually become stronger as a consequence of them having to compete in the interests of the selfish genes with the rational decision-making process.

Another property of the system may be that sometimes the rational system cannot gain all the evidence that would be needed to make a rational choice. Under these circumstances, the rational system might fail to make a clear decision, and under these circumstances, basing a decision on the gene-specified emotions is an alternative. Indeed, Damasio (1994) argued that under circumstances such as this, emotions might take an important role in decision-making. In this respect, I agree with him, basing my reasons on the arguments above. He called the emotional feelings gut feelings, and, in contrast to me, hypothesized that actual feedback from the gut was involved. His argument seemed to be that if the decision was too complicated for the rational system, then send outputs to the viscera, and whatever is sensed by what they send back could be used in the decision-making and would account for the conscious feelings of the emotional states. My reading of the evidence is that the feedback from the periphery is not necessary for the emotional decision-making, or for the feelings, nor would it be computationally efficient to put the viscera in the loop given that the information starts from the brain, but that is a matter considered elsewhere (Rolls 2005b).

Another property of the system is that the interests of the second, rational, system, although involving a different form of computation, should not be too far from those of the gene-defined emotional system, for the arrangement to be stable in evolution by natural selection. One way that this could be facilitated would be if the gene-based goals felt pleasant or unpleasant in the rational system and in this way contributed to the operation of the second, rational, system. This is something that I propose is the case.

9.3.3 Decision-Making Between the Implicit and Explicit Systems

Decision-making as implemented in neural networks in the brain is now becoming understood and is described in Sect. 9.4. As shown there, two attractor states, each one corresponding to a decision, compete in an attractor single network with the evidence for each of the decisions acting as biases to each of the attractor states. The non-linear dynamics, and the way in which noise due to the random spiking

of neurons makes the decision-making probabilistic, makes this a biologically plausible model of decision-making consistent with much neurophysiological and fMRI data (Wang 2002; Deco and Rolls 2006; Deco et al. 2009; Rolls and Deco 2010).

I propose (Rolls 2005b, 2008b) that this model applies to taking decisions between the implicit (unconscious) and explicit (conscious) systems in emotional decision-making, where the two different systems could provide the biasing inputs λ_1 and λ_2 to the model. An implication is that noise will influence with probabilistic outcomes which system takes a decision.

When decisions are taken, sometimes 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 and LeDoux 1978; Gazzaniga 1988, 1995; Rolls 2005b; LeDoux 2008). 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.

9.3.4 Free Will

These thoughts raise the issue of free will in decision-making.

First, we can note that in so far as 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 in this book. [Philosophers may wish to argue about different senses of the term deterministic, but is it being used here in a precise, scientific and quantitative way, which has been clearly defined (Rolls and Deco 2010).]

Second, do we have free will when both the implicit and the explicit systems have made the choice? Free will would in Rolls' view (2005b) 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 reason why confabulation occurs.

The interesting view we are led to is thus that when probabilistic choices influenced by stochastic dynamics 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.

9.4 Decision-Making and "Subjective Confidence"

Animals including humans can not only take decisions, but they can then make further decisions based on their estimates of their confidence or certainty in the decision just taken. It has been argued that the ability to make confidence estimates "objectively measures awareness" (Koch and Preuschoff 2007; Persaud et al. 2007). This process is sometimes called "subjective confidence", referring to the fact that one can report on the confidence one has in one's decisions. But does estimating the confidence in a decision really provide a measure of consciousness or require it? The process of confidence estimation has been described in animals including monkeys and rodents, who may, for example, terminate a trial if they estimate that a wrong decision may have been made so that they can get on to the next trial (Hampton 2001; Hampton et al. 2004; Kepecs et al. 2008). Does this really imply subjective (i.e. conscious) awareness (Heyes 2008)?

We have now developed an understanding of how probabilistic decision-making may be implemented in the brain by a single attractor network, and how adding a second attractor network allows the system to take a decision based on the confidence in the decision that emerges in the neuronal firing during the decision-making in the first network (Insabato et al. 2010). We describe these developments next. We note that there is no reason to believe that a system with two attractor networks is consciously aware, yet this system can account for confidence estimation and decisions made based on this, that is, what is described as "subjective confidence".

9.4.1 Neural Networks for Decision-Making That Reflect "Subjective Confidence" in Their Firing Rates

In spite of the success of phenomenological models for accounting for decision-making performance (Smith and Ratcliff 2004), a crucial problem that they present is the lack of a link between the model variables and parameters and the biological substrate. Recently, a series of biologically plausible models, motivated and constrained by neurophysiological data, have been formulated to establish an explicit link between behaviour and neuronal activity (Wang 2002; Deco and Rolls 2006; Wong and Wang 2006; Rolls and Deco 2010; Rolls et al. 2010a, b). The way in which these integrate-and-fire neuronal network models operate is as follows.

An attractor network of the type illustrated in Fig. 9.2a is set up to have two possible high firing rate attractor states, one for each of the two decisions. The evidence for each decision (1 vs. 2) biases each of the two attractors via the external inputs λ_1 and λ_2. The attractors are supported by strengthened synaptic connections in the recurrent collateral synapses between the (e.g. cortical pyramidal) neurons activated when λ_1 is applied or when λ_2 is applied. (This is an associative or Hebbian process set up during a learning stage by a process like long-term potentiation.) Inhibitory interneurons (not shown in Fig. 9.2a) receive inputs from the pyramidal neurons and make negative feedback connections onto the pyramidal cells to control their activity. When inputs λ_1 and λ_2 are applied, there is positive feedback via the recurrent collateral connections and competition implemented through the inhibitory interneurons so that there can be only one winner. The network starts in a low spontaneous state of firing. When λ_1 and λ_2 are applied, there is competition between the two attractors, each of which is pushed towards a high firing rate state, and eventually, depending on the relative strength of the two inputs and the noise in the network caused by the random firing times of the neurons, one of the attractors will win the competition, and it will reach a high firing rate state, with the firing of the neurons in the other attractor inhibited to a low firing rate. The process is illustrated in Fig. 9.3. The result is a binary decision, with one group of neurons due to the positive feedback firing at a high firing rate, and the neurons corresponding to the other decision firing with very low rates. Because it is a non-linear positive feedback system, the final firing rates are in what is effectively a binary decision state, of high firing rate or low firing rate, and do not reflect the exact relative values of the two inputs λ_1 and λ_2 once the decision is reached. The noise in the network due to the random spiking of the neurons is important to the operation of the network, because it enables the network to jump out of a stable spontaneous rate of firing to a high firing rate and to do so probabilistically, depending on whether on a particular trial there is relatively more random firing in the neurons of one attractor than the other attractor. This can be understood in terms of energy landscapes, where each attractor (the spontaneous state and the two high firing rate attractors) is a low energy basin, and the spiking noise helps the system to jump over an energy barrier into another energy minimum, as illustrated in Fig. 9.2c. If λ_1 and λ_2 are equal, then the decision that is taken is random and probabilistic, with the noise in each attractor determining which decision is taken on a particular trial. If one of the inputs is larger than

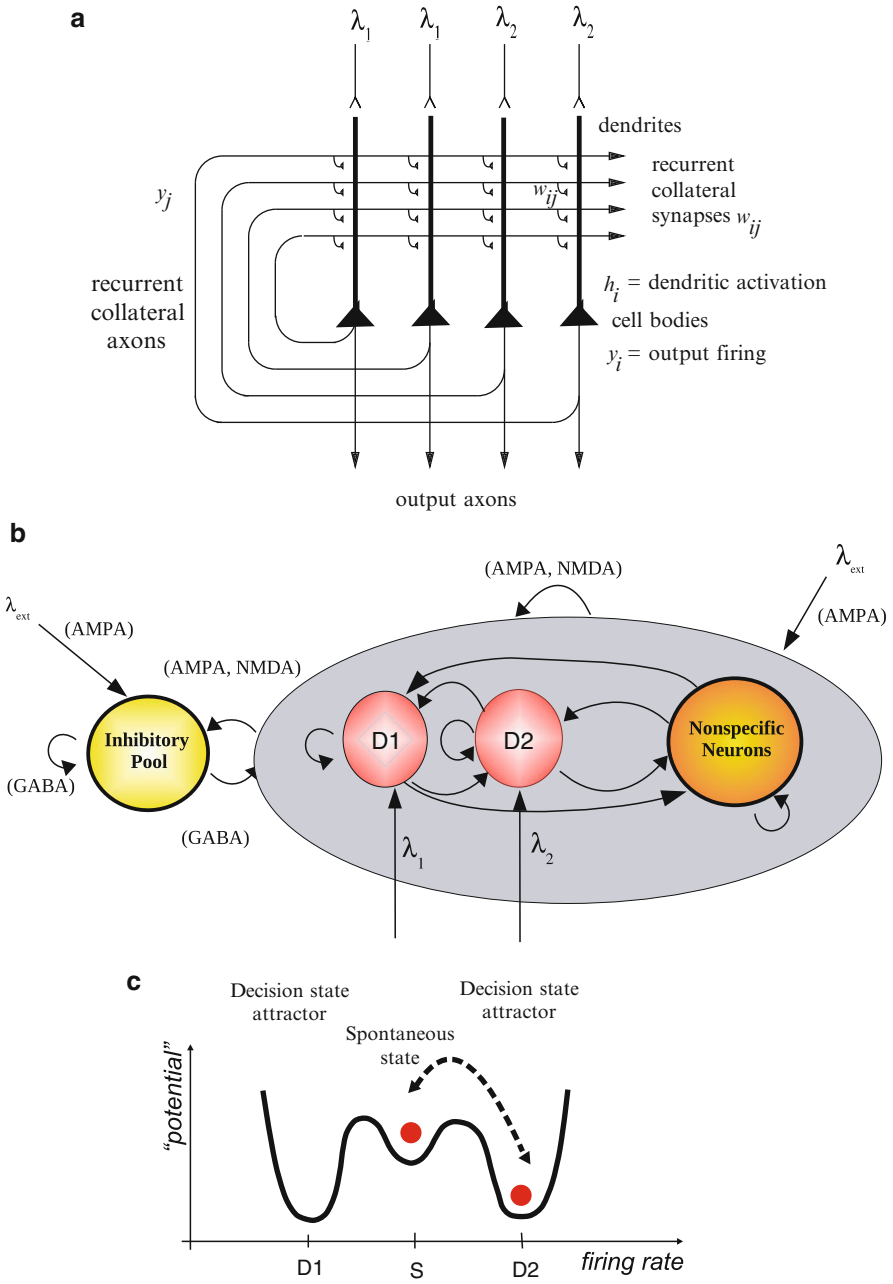

Fig. 9.2 (**a**) Attractor or autoassociation single network architecture for decision-making. The evidence for decision 1 is applied via the λ_1 inputs and for decision 2 via the λ_2 inputs. The synaptic weights $w_i j$ have been associatively modified during training in the presence of λ_1 and at a different time of λ_2.

the other, then the decision is biased towards it but is still probabilistic. Because this is an attractor network, it has short term memory properties implemented by the recurrent collaterals, which tend to promote a state once it is started, and these help it to maintain the firing once it has reached the decision state, enabling a suitable action to be implemented even if this takes some time.

In the multi-stable regime investigated by Deco and Rolls (2006), the spontaneous firing state is stable even when the decision cues are being applied, and noise-related fluctuations are essential for decision-making (Deco et al. 2009; Rolls and Deco 2010).

The process of decision-making in this system, and how the certainty or confidence of the decision is represented by the firing rates of the neurons in the network, is illustrated in Fig. 9.3. Figure 9.3a, e shows the mean firing rates of the two neuronal populations D1 and D2 for two trial types, easy trials ($\Delta I = 160$ Hz) and difficult trials ($\Delta I = 0$) (where ΔI is the difference in spikes/s summed across all synapses to each neuron between the two inputs, λ_1 to population D1 and λ_2 to population D2). The results are shown for correct trials, that is, trials on which the D1 population won the competition and fired with a rate for >10 spikes/s for the last 1,000 ms of the simulation runs. Figure 9.3b shows the mean firing rates of the four populations of neurons on a difficult trial, and Fig. 9.3c shows the rastergrams for the same trial. Figure 9.3d shows the firing rates on another difficult trial ($\Delta I = 0$) to illustrate the variability shown from trial to trial, with on this trial prolonged competition between the D1 and D2 attractors until the D1 attractor finally won after approximately 1,100 ms. Figure 9.3f shows firing rate plots for the four neuronal populations on an example of a single easy trial ($\Delta I = 160$), Fig. 9.3g shows the synaptic currents in the four neuronal populations on the same trial, and Fig. 9.3h shows rastergrams for the same trial.

Three important points are made by the results shown in Fig. 9.3. First, the network falls into its decision attractor faster on easy trials than on difficult trials. Reaction times are thus shorter on easy than on difficult trials. Second, the mean

Fig. 9.2 (continued) When λ_1 and λ_2 are applied, each attractor competes through the inhibitory interneurons (not shown), until one wins the competition, and the network falls into one of the high firing rate attractors that represents the decision. The noise in the network caused by the random spiking of the neurons means that on some trials, for given inputs, the neurons in the decision 1 (D1) attractor are more likely to win, and on other trials the neurons in the decision 2 (D2) attractor are more likely to win. This makes the decision-making probabilistic, for, as shown in (**c**), the noise influences when the system will jump out of the spontaneous firing stable (low energy) state S, and whether it jumps into the high firing state for decision 1 (D1) or decision 2 (D2). (**b**) The architecture of the integrate-and-fire network used to model decision-making (see text). (**c**) A multi-stable "effective energy landscape" for decision-making with stable states shown as low "potential" basins. Even when the inputs are being applied to the network, the spontaneous firing rate state is stable, and noise provokes transitions into the high firing rate decision attractor state D1 or D2 (see Rolls and Deco 2010)

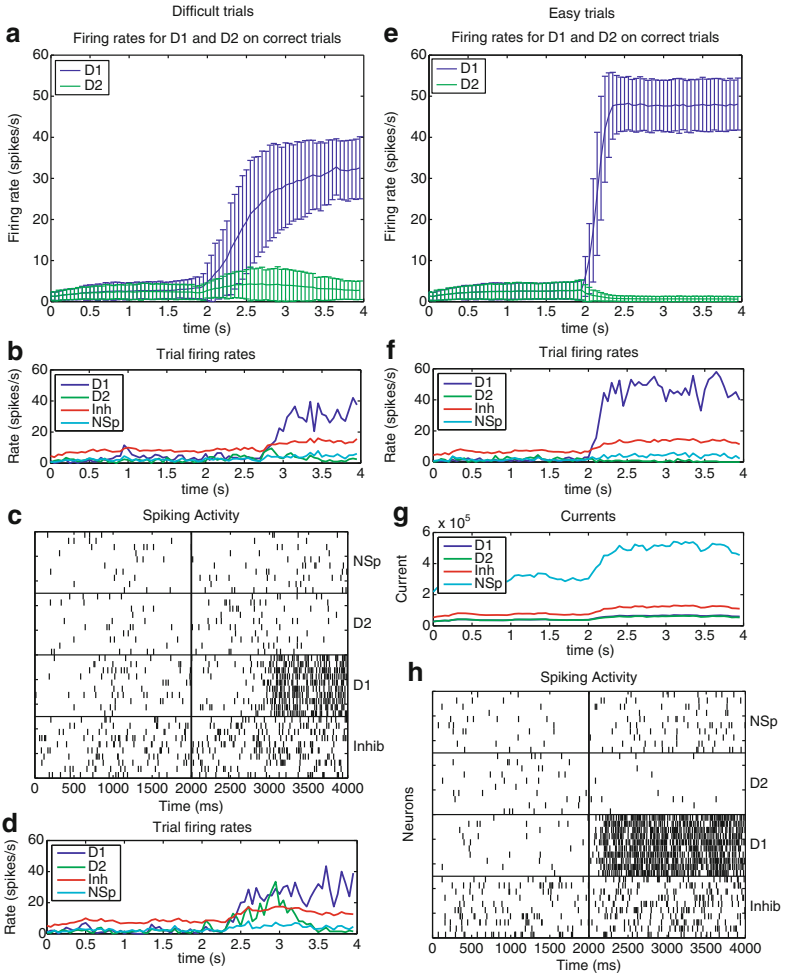

Fig. 9.3 (**a** and **e**) Firing rates (mean ± sd) for difficult ($\Delta I = 0$) and easy ($\Delta I = 160$) trials. The period 0–2 s is the spontaneous firing, and the decision cues were turned on at time = 2 s. The mean was calculated over 1,000 trials. D1: firing rate of the D1 population of neurons on correct trials on which the D1 population won. D2: firing rate of the D2 population of neurons on the correct trials on which the D1 population won. A correct trial was one in which in which the mean rate of the D1 attractor averaged >10 spikes/s for the last 1,000 ms of the simulation runs. (Given the attractor nature of the network and the parameters used, the network reached one of the attractors on >90% of the 1,000 trials, and this criterion clearly separated these trials, as indicated by the mean rates and standard deviations for the last s of the simulation as shown.) (**b**) The mean firing rates of the four populations of neurons on a difficult trial. Inh is the inhibitory population that uses GABA as a transmitter. NSp is the non-specific population of neurons (see Fig. 9.2). (**c**) Rastergrams for the trial shown in (**b**) 10 neurons from each of the four pools of neurons are shown. (**d**) The firing rates on another difficult trial ($\Delta I = 0$) showing prolonged competition between the D1 and D2 attractors until the D1 attractor finally wins after approximately 1,100 ms. (**f**) Firing rate plots for the 4 neuronal populations on a single easy trial ($\Delta I = 160$). (**g**) The synaptic currents in the four neuronal populations on the trial shown in (**f**). (**h**) Rastergrams for the easy trial shown in (**f** and **g**)

firing rate of the winning attractor after the network has settled into the correct decision attractor is higher on easy trials (with large ΔI, and when certainty and confidence are high) than on difficult trials. This is because the exact firing rate in the attractor is a result not only of the internal recurrent collateral effect, but also of the external input to the neurons, which in Fig. 9.3 is 32 Hz to each neuron (summed across all synapses) of D1 and D2, but in Fig. 9.3a is increased by 80 Hz to D1 and decreased by 80 Hz to D2 (i.e. the total external input to the network is the same, but $\Delta I = 0$ for Fig. 9.3a and $.\Delta I = 160$ for Fig. 9.3b). Third, the variability of the firing rate is high, with the standard deviations of the mean firing rate calculated in 50 ms epochs indicated in order to quantify the variability. The large standard deviations on difficult trials for the first second after the decision cues are applied at $t = 2$s reflects the fact that on some trials the network has entered an attractor state after 1,000 ms, but on other trials it has not yet reached the attractor, although it does so later. This trial-by-trial variability is indicated by the firing rates on individual trials and the rastergrams in the lower part of Fig. 9.3.

The effects evident in Fig. 9.3 are quantified and elucidated over a range of values for ΔI elsewhere (Rolls et al. 2010a, b). They show that a continuous-valued representation of decision certainty or decision confidence is encoded in the firing rates of the neurons in a decision-making attractor.

9.4.2 A Model for Decisions About Confidence Estimates

We have seen that a continuous-valued representation of decision certainty or decision confidence is encoded in the firing rates of the neurons in a decision-making attractor. What happens if instead of having to report or assess the continuous-valued representation of confidence in a decision one has taken, one needs to take a decision based on one's confidence estimate that one has just made a correct or incorrect decision? One might, for example, wait for a reward if one thinks one's decision was correct, or alternatively stop waiting on that trial and start another trial or action. We suggest that in this case, one needs a second decision-making network that takes decisions based on one's decision confidence (Insabato et al. 2010).

The architecture has a decision-making network, and a separate confidence decision network that receives inputs from the decision-making network, as shown in Fig. 9.4. The decision-making network has two main pools or populations of neurons, D1 which become active for decision 1, and D2 which become active for decision 2. Pool D1 receives sensory information about stimulus 1 (e.g. odor A), and Pool D2 receives sensory information about stimulus 2 (e.g. odor B). Each of these pools has strong recurrent collateral connections between its own neurons, so that each operates as an attractor population. There are inhibitory neurons with global connectivity to implement the competition between the attractor subpopulations. When stimulus 1 is applied, pool D1 will usually win the competition and end up with high firing indication that decision 1 has been reached. When stimulus 2 is applied, pool D2 will usually win the competition and end up

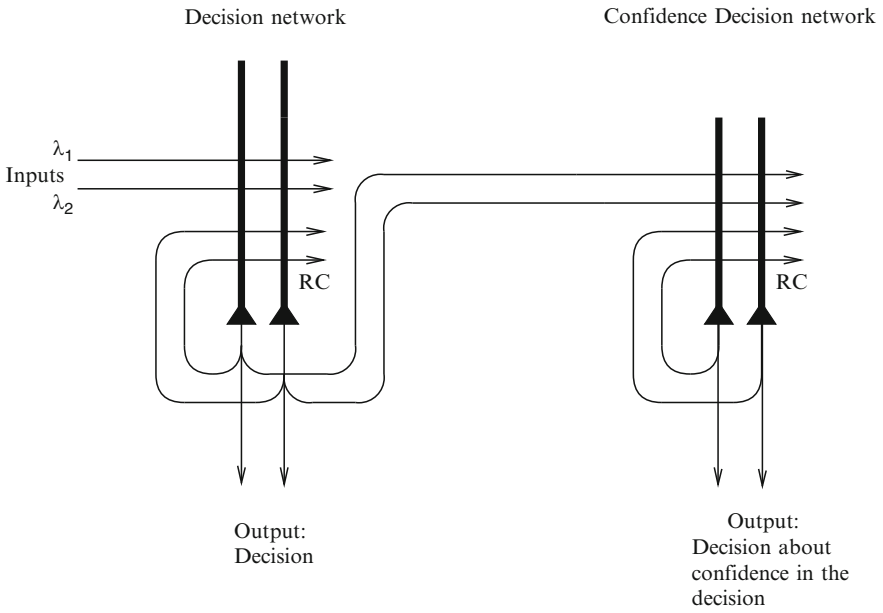

Fig. 9.4 Decisions about confidence estimates. The first network is a decision-making network, and its outputs are sent to a second network that makes decisions based on the firing rates from the first network, which reflect the decision confidence. In the first network, high firing of population D_1 represents decision 1, and high firing of population D_2 represents decision 2. The second network is a confidence decision network and receives inputs from the first network. The confidence 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). In each network of the integrate-and-fire simulations, the excitatory pool is divided into three subpopulations: a non-specific one and two stimulus selective populations. The selective pools are endowed with strong recurrent connections (w_+), while the connections between the two selective pools are weak (w_-). All other connections are set to the default value 1. All neurons in the network receive an input (λ_{ext}) emulating the spontaneous firing of neurons in other cortical areas. Pools D_1 and D_2 receive also a stimulus related input (respectively λ_1 and λ_2 which is the information that will drive the decision) (After Insabato et al. 2010)

with high firing indication that decision 2 has been reached. When a mixture is applied, the decision-making network will probabilistically choose stimulus 1 or 2 influenced by the proportion of stimulus 1 and 2 in the mixture. The decision-making is probabilistic because the neurons in the network have approximately Poisson spike time firings which are a source of noise, which, in a finite size system, cause coherent statistical fluctuations, as described in more detail elsewhere (Deco et al. 2009; Rolls and Deco 2010). The proportion of correct decisions increases as the proportion of stimulus 1 and 2 in the mixture is altered from 50% (corresponding to $\Delta I = 0$) to 100% of one and 0% of the other (corresponding to a large ΔI). The firing rates of the neurons in the two selective populations as a function of the proportion of stimulus 1 and 2 in the mixture are similar to those illustrated in Fig. 9.3. The neurons that win on correct trials have higher firing rates as the dif-

ference in the proportion of A (stimulus 1) and B (stimulus 2) in the mixture, which alters ΔI, becomes larger in magnitude. (ΔI is defined as $(A-B)/((A+B)/2)$. ΔI is thus large and positive if only A is present, is large and negative if only B is present, and is 0 if A and B are present in equal proportions.) The reason that the firing rates of the winning pool become higher as ΔI becomes larger in magnitude is that the external inputs from the stimuli 1 or 2 then support the winning attractor and add to the firing rates being produced by the recurrent collateral connections in the winning attractor. On the other hand, the firing rates of the winning pool become lower on error trials as ΔI increases, because then the external sensory inputs are inconsistent with the decision that has been taken and do not support and increase the firing rate of the winning pool.

The decision network sends outputs from its decision-making selective pools D1 and D2 to the confidence network. The confidence 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). If the output firing of D1 and D2 is high because the decision just taken has sensory inputs consonant with the decision, then the confidence network acting as a second level network takes the decision, probabilistically as before, to have confidence in the decision, and the C population wins the competition. If the output firing of D1 and D2 is low because the decision just taken has sensory inputs that are not consonant with the decision, then the confidence network takes the decision, probabilistically as before, to have a lack of confidence in the decision, and the LC population wins the competition. The confidence network thus acts as a decision-making network to make confident decisions if the firing rates from the first, decision-making, network are high and to make lack of confidence decisions if the firing rates from the first, decision-making network are low.

In this situation, we find in integrate-and-fire simulations (Insabato et al. 2010) that on correct trials with high ΔI (easy decisions), C has a high firing rate, whereas it has a lower rate for $\Delta I = 0$, that is difficult decisions. Conversely, on error trials when the firing rates in the first level, decision-making, networks are lower, and the confidence neurons C have firing rates that decrease as the magnitude of ΔI increases. The LC attractor neurons do not receive inputs from the first level, decision-making, network, and thus through the inhibitory neurons has the opposite type of firing to the confidence pool C. That is, the firing rates of the LC are in general high on error trials (when the firing rates of the first-level neurons are low) and increase as ΔI increases (Insabato et al. 2010).

This new theoretical approach to confidence-related decision-making accounts for neuronal responses in the rat orbitofrontal cortex related to confidence in decisions (Kepecs et al. 2008). The rats had to perform a binary categorization task with a mixture of two pure odourants (A, caproic acid; B, 1-hexanol) by entering one of two ports to indicate that the mixture was more like odour A or odour B. Correct choices were rewarded after a delay of 0.3–2 s. Varying the relative concentration of the odourants allowed the difficulty of the trial to be altered. Neural activity related to decision confidence should occur just after the decision

is taken and before the trial outcome. Kepecs et al., therefore, analyzed recordings of neuronal activity during a delay period after a decision had been taken before a reward was given. (The single neuron recordings were made in the rat orbitofrontal cortex, though exactly what area in primates and humans corresponds to the area in which recordings were made is not yet clear.) The neurons were divided into two groups based on whether they fired faster on correct or on error trials. Kepecs et al. found that the group of neurons with an increased firing rate on error trials had higher firing rates with easier stimuli. The same neurons fired at a substantially lower rate on correct trials, and on these trials the firing rates were lower when the decision was made easier. This produced opposing V-shaped curves. The authors argued that this pattern of activity encoded decision confidence.

In the experiment of Kepecs et al. (2008), C corresponds to a decision to stay and wait for a reward, i.e. what they call the positive outcome population, though it really represents confidence or a prediction that the decision just taken will have a positive outcome. LC corresponds to a decision to abort a trial and not wait for a possible reward, i.e. what they call the negative outcome population, though it really represents lack of confidence that the decision just taken will have a positive outcome.

Kepecs et al. (2008) then performed a second experiment to investigate if rats were able to make use of this information encoded by orbitofrontal cortex neurons. The delay period was prolonged up to 8 s in order to allow the rat to reinitiate the trial. The subject could decide to leave the choice port and repeat the trial or could wait for the reward. It was found that when the likelihood of a reward was low, due to the decision difficulty and the choice just made, the rat returned to the odour port. The probability that the rat would restart a trial as a function of stimulus difficulty and the choice just made were consistent with the responses of the neurons with activity described as encoding decision confidence: in our terminology, in taking decisions based on one's confidence in an earlier decision. The model makes predictions about the firing rates of these neuronal populations when the second, confidence decision, network itself makes an error due to the noise in the system (Insabato et al. 2010).

These results indicate that confidence estimates, previously suggested to "objectively measure awareness" (Koch and Preuschoff 2007; Persaud et al. 2007), can be computed with relatively simple operations, involving the firing rates of the neurons when a decision-making network falls into an attractor. Moreover, adding a second attractor decision-making network even enables decisions to be made about the confidence in the first decision. There would seem to be no awareness in either of the networks. The implication is that we need to be very careful in ascribing awareness to processes that at first seem complex and closely tied to awareness. The implication in turn is that activity in a special and different processing system, argued above to be that capable of thoughts about thoughts (higher order syntactic thoughts, HOSTs), is what is occurring when we are aware, that is when we have phenomenal consciousness.

9.5 Oscillations and Stimulus-Dependent Neuronal Synchrony: Their Role in Information Processing in the Ventral Visual System and in Consciousness

We now turn in Sects. 5–7 to some other computational issues related to the implementation of consciousness in the brain. The first is on whether oscillations are important in consciousness.

It has been suggested that syntax in real neuronal networks is implemented by temporal binding (see Malsburg 1990), which would be evident as, for example, stimulus-dependent synchrony (Singer 1999). According to this hypothesis, the binding between features common to an object could be implemented by neurons coding for that object firing in synchrony, whereas if the features belong to different objects, the neurons would not fire in synchrony. Crick and Koch (1990) postulated that oscillations and synchronization are necessary bases of consciousness. It is difficult to see what useful purpose oscillations per se could perform for neural information processing, apart from perhaps resetting a population of neurons to low activity so that they can restart some attractor process (see e.g. Rolls and Treves 1998), acting as a reference phase to allow neurons to provide some additional information by virtue of the time that they fire with respect to the reference waveform (Huxter et al. 2003) or increasing spike numbers (Smerieri et al. 2010). Neither putative function seems to be closely related to consciousness. However, stimulus-dependent synchrony, by implementing binding, a function that has been related to attention (Treisman 1996) might perhaps be related to consciousness. Let us consider the evidence on whether stimulus-dependent synchrony between neurons provides significant information related to object recognition and top-down attention in the ventral visual system.

This has been investigated by developing information theoretic methods for measuring the information present in stimulus-dependent synchrony (Panzeri et al. 1999; Rolls et al. 2003; Franco et al. 2004) and applying them to the analysis of neuronal activity in the macaque inferior temporal visual cortex during object recognition and attention. This brain region represents both features, such as parts of objects and faces, and whole objects in which the features must be bound in the correct spatial relationship for the neurons to respond (Rolls and Deco 2002; Rolls 2007b, 2008b). It has been shown that simultaneously recorded single neurons do sometimes show stimulus-dependent synchrony, but that the information available is less than 5% of that available from the spike counts (Rolls et al. 2003, 2004; Franco et al. 2004; Aggelopoulos et al. 2005; Rolls 2008b).

The neurophysiological studies performed have included situations in which feature binding is likely to be needed, that is when the monkey had to choose to touch one of two simultaneously presented objects, with the stimuli presented in a complex natural background in a top-down attentional task (Aggelopoulos et al. 2005). We found that between 99 and 94% of the information was present in the firing rates of inferior temporal cortex neurons, and less than 5% in any stimulus-dependent synchrony that was present, as illustrated in Fig. 9.3. The implication of these results is that any stimulus-dependent synchrony that is present is not quantitatively important as measured by information theoretic analyses under natural scene conditions.

The point of the experimental design used was to test whether when the visual system is operating normally, in natural scenes and even searching for a particular object, stimulus-dependent synchrony is quantitatively important for encoding information about objects, and it was found not to be in the inferior temporal visual cortex. It will be of interest to apply the same quantitative information theoretic methods to earlier cortical visual areas, but the clear implication of the findings is that even when features must be bound together in the correct relative spatial positions to form object representations, and these must be segmented from the background, then stimulus-dependent synchrony is not quantitatively important in information encoding (Aggelopoulos et al. 2005; Rolls 2008b). Further, it was shown that there was little redundancy (less than 6%) between the information provided by the spike counts of the simultaneously recorded neurons, making spike counts an efficient population code with a high encoding capacity (Rolls 2008b).

The findings (Aggelopoulos et al. 2005) are consistent with the hypothesis that feature binding is implemented by neurons that respond to features in the correct relative spatial locations (Elliffe et al. 2002; Rolls and Deco 2002; Rolls 2008b) and not by temporal synchrony and attention (Malsburg 1990; Singer 1999). In any case, the computational point is that even if stimulus-dependent synchrony was useful for grouping, it would not without much extra machinery be useful for binding the relative spatial positions of features within an object or for that matter of the positions of objects in a scene which appears to be encoded in a different way by using receptive fields that become asymmetric in crowded scenes (Aggelopoulos and Rolls 2005). The computational problem is that synchronization does not by itself define the spatial relations between the features being bound, so is not as a binding mechanism adequate for shape recognition. For example, temporal binding might enable features 1, 2 and 3, which might define one stimulus to be bound together and kept separate from, for example, another stimulus consisting of features 2, 3 and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2 and 3 in the 123 stimulus, so that it can be discriminated from, for example, 312 (Rolls 2008b). However, the required computation for binding can be performed by the use of neurons that respond to combinations of features with a particular spatial arrangement (Elliffe et al. 2002; Rolls and Deco 2002; Rolls and Stringer 2006; Rolls 2008b).

Another type of evidence that stimulus-dependent neuronal synchrony is not likely to be crucial for information encoding, at least in the ventral visual system, is that the code about which visual stimulus has been shown can be read off from the end of the visual system in short times of 20–50 ms, and cortical neurons need fire for only this long during the identification of objects (Tovee et al. 1993; Rolls and Tovee 1994; Rolls et al. 1994b, 2006; Tovee and Rolls 1995; Rolls 2008b). These are rather short time windows for the expression of multiple separate populations of synchronized neurons.

If large populations of neurons become synchronized, oscillations are likely to be evident in cortical recordings. In fact, oscillations are not an obvious property of neuronal firing in the primate temporal cortical visual areas involved in the representation of faces and objects when the system is operating normally in the awake behaving macaque (Tovee and Rolls 1992). 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 synchronization are much more difficult to demonstrate in the temporal cortical visual areas of awake behaving monkeys (Aggelopoulos et al. 2005) might just mean that during evolution to primates the cortex has become better able to avoid parasitic oscillations, as a result of developing better feedforward and feedback inhibitory circuits (see Rolls and Deco 2002; Rolls and Treves 1998).

However, in addition there is an interesting computational argument against the utility of oscillations. The computational argument is related to the speed of information processing in cortical circuits with recurrent collateral connections. It has been shown that if attractor networks have integrate-and-fire neurons and spontaneous activity, then memory recall into a basin of attraction can occur in approximately 1.5 time constants of the synapses, i.e. in times in the order of 15 ms (Treves 1993; Simmen et al. 1996; Battaglia and Treves 1998; Rolls and Treves 1998; Panzeri et al. 2001). One factor in this rapid dynamics of autoassociative networks with brain-like integrate-and-fire membrane and synaptic properties is that with some spontaneous activity, some of the neurons in the network are close to threshold already before the recall cue is applied, and hence some of the neurons are very quickly pushed by the recall cue into firing, so that information starts to be exchanged very rapidly (within 1–2 ms of brain time) through the modified synapses by the neurons in the network. The progressive exchange of information starting early on within what would otherwise be thought of as an iteration period (of perhaps 20 ms, corresponding to a neuronal firing rate of 50 spikes/s) is the mechanism accounting for rapid recall in an autoassociative neuronal network made biologically realistic in this way. However, this process relies on spontaneous random firings of different neurons, so that some will always be close to threshold when the retrieval cue is applied. If many of the neurons were firing synchronously in an oscillatory pattern, then there might be no neurons close to threshold and ready to be activated by the retrieval cue, so that the network might act much more like a discrete time network with fixed timesteps, which typically takes 8–15 iterations to settle, equivalent to perhaps 100 ms of brain time, and much too slow for cortical processing within any one area (Rolls and Treves 1998; Panzeri et al. 2001; Rolls 2008b). The implication is that oscillations would tend to be detrimental to cortical computation by slowing down any process using attractor dynamics. Attractor dynamics are likely to be implemented not only by the recurrent collateral connections between pyramidal neurons in a given cortical area but also by the reciprocal feedforward and feedback connections between adjacent layers in cortical processing hierarchies (Rolls 2008b). However, if oscillations increase spike numbers in a process like stochastic resonance, this could speed information processing (Smerieri et al. 2010).

Another computational argument is that it is possible to account for many aspects of attention, including the non-linear interactions between top-down and bottom-up inputs, in integrate-and-fire neuronal networks that do not oscillate or show stimulus-dependent synchrony (Deco and Rolls 2005a, b, c; Rolls 2008b).

The implication of these findings is that stimulus-dependent neuronal synchronization and oscillatory activity are unlikely to be quantitatively important in cortical processing, at least in the ventral visual stream. To the extent that we can be conscious of activity that has been processed in the ventral visual stream (made evident, for example, by reports of the appearance of objects), stimulus-dependent synchrony and oscillations are unlikely to be important in the neural mechanisms of consciousness.

9.6 A Neural Threshold for Consciousness: The Neurophysiology of Backward Masking

Damage to the primary (striate) visual cortex can result in blindsight, in which patients report that they do not see stimuli consciously, yet when making forced choices can discriminate some properties of the stimuli such as motion, position, some aspects of form, and even face expression (Azzopardi and Cowey 1997; Weiskrantz 1997, 1998; De Gelder et al. 1999). In normal human subjects, backward masking of visual stimuli, in which another visual stimulus closely follows the short presentation of a test stimulus, reduces the visual perception of the test visual stimulus, and this paradigm has been widely used in psychophysics (Humphreys and Bruce 1991). In this section, I consider how much information is present in neuronal firing in the part of the visual system that represents faces and objects, the inferior temporal visual cortex (Rolls and Deco 2002; Rolls 2008b), when human subjects can discriminate face identity in forced choice testing but cannot consciously perceive the person's face, and how much information is present when they become conscious of perceiving the stimulus. From this evidence, I argue that even *within* a particular processing stream the processing may not be conscious yet can lead to behaviour, and that with higher and longer neuronal firing, events in that system become conscious. From this evidence, I argue that the threshold for consciousness is normally higher than for some behavioural response. I then suggest a computational hypothesis for why this might be adaptive. A fuller account of this issue is available elsewhere (Rolls 2007a).

9.6.1 The Neurophysiology and Psychophysics of Backward Masking

The responses of single neurons in the macaque inferior temporal visual cortex have been investigated during backward visual masking (Rolls and Tovee 1994; Rolls et al. 1994b). Recordings were made from neurons that were selective for faces, using distributed encoding (Rolls and Deco 2002; Rolls 2007b, 2008b), during presentation of a test stimulus, a face, that lasted for 16 ms. The test stimulus was followed on different trials by a mask with stimulus onset asynchrony (S.O.A.)

values of 20, 40, 60, 100 or 1,000 ms. (The S.O.A. is the time between the onset of the test stimulus and the onset of the mask.) The duration of the pattern masking stimulus (letters of the alphabet) was 300 ms, and the neuron did not respond to the masking stimulus.

One important conclusion from these results is that the effect of a backward masking stimulus on cortical visual information processing is to limit the duration of neuronal responses by interrupting neuronal firing. This persistence of cortical neuronal firing when a masking stimulus is not present is probably related to cortical recurrent collateral connections which could implement an autoassociative network with attractor and short-term memory properties (see Rolls and Treves 1998; Rolls and Deco 2002; Rolls 2008b), because such continuing post-stimulus neuronal firing is not observed in the lateral geniculate nucleus (K. Martin, personal communication).

Information theoretic analyses (Rolls et al. 1999) showed that as the S.O.A. is reduced towards 20 ms the information does reduce rapidly, but that nevertheless at an S.O.A. of 20 ms there is still considerable information about which stimulus was shown. Rolls et al. (1994b) performed human psychophysical experiments with the same set of stimuli and with the same apparatus used for the neurophysiological experiments so that the neuronal responses could be closely related to the identification that was possible of which face was shown. Five different faces were used as stimuli. All the faces were well known to each of the eight observers who participated in the experiment. In the forced choice paradigm, the observers specified whether the face was normal or rearranged (i.e. with the features jumbled) and identified whose face they thought had been presented. Even if the observers were unsure of their judgement they were instructed to respond with their best guess. The data were corrected for guessing. Forced choice discrimination of face identity was better than chance at an S.O.A. of 20 ms. However, at this S.O.A., the subjects were not conscious of seeing the face, or of the identity of the face, and felt that their guessing about which face had been shown was not correct. The subjects did know that something had changed on the screen (and this was not just brightness, as this was constant throughout a trial). Sometimes the subjects had some conscious feeling that a part of a face (such as a mouth) had been shown. However, the subjects were not conscious of seeing a whole face or of seeing the face of a particular person. At an S.O.A. of 40 ms, the subjects' forced choice performance of face identification was close to 100%, and at this S.O.A., the subjects became much more consciously aware of the identity of which face had been shown (Rolls et al. 1994b).

Comparing the human performance purely with the changes in firing rate under the same stimulus conditions suggested that when it is just possible to identify which face has been seen, neurons in a given cortical area may be responding for only approximately 30 ms (Rolls and Tovee 1994; Rolls et al. 1994b; Rolls 2007a). The implication is that 30 ms is enough time for a neuron to perform sufficient computation to enable its output to be used for identification. When the S.O.A. was increased to 40 ms, the inferior temporal cortex neurons responded for approximately 50 ms and encoded approximately 0.14 bits of information. At this S.O.A., not only was face identification 97% correct, but the subjects were much more likely to be able

to report consciously seeing a face and/or whose face had been shown. One further way in which the conscious perception of the faces was measured quantitatively was by asking subjects to rate the clarity of the faces. This was a subjective assessment and therefore reflected conscious processing and was made using magnitude estimation. It was found that the subjective clarity of the stimuli was low at 20 ms S.O.A., was higher at 40 ms S.O.A. and was almost complete by 60 ms S.O.A (Rolls et al. 1994b; Rolls 2003, 2005a, 2007a).

It is suggested that the threshold for conscious visual perception may be set to be higher than the level at which small but significant sensory information is present so that the systems in the brain that implement the type of information processing involved in conscious thoughts are not interrupted by small signals that could be noise in sensory pathways. It is suggested that the processing related to consciousness involves a HOST system used to correct first order syntactic thoughts, and that these processes are inherently serial because of the way that the binding problems associated with the syntactic binding of symbols are treated by the brain. The argument is that it would be inefficient and would not be adaptive to interrupt this serial processing if the signal was very small and might be related to noise. Interruption of the serial processing would mean that the processing would need to start again, as when a train of thought is interrupted. The small signals that do not interrupt conscious processing but are present in sensory systems may nevertheless be useful for some implicit (non-conscious) functions, such as orienting the eyes towards the source of the input, and may be reflected in the better than chance recognition performance at short S.O.A.s even without conscious awareness.

9.6.2 The Relation to Blindsight

These quantitative analyses of neuronal activity in an area of the ventral visual system involved in face and object identification which show that significant neuronal processing can occur that is sufficient to support forced choice but implicit (unconscious) discrimination in the absence of conscious awareness of the identity of the face is of interest in relation to studies of blindsight (Azzopardi and Cowey 1997; Weiskrantz 1997, 1998; De Gelder et al. 1999). It has been argued that the results in blindsight are not due just to reduced visual processing, because some aspects of visual processing are less impaired than others (Azzopardi and Cowey 1997; Weiskrantz 1997,1998, 2001). It is though suggested that some of the visual capacities that do remain in blindsight reflect processing via visual pathways that are alternatives to the V1 processing stream (Weiskrantz 1997, 1998, 2001). If some of those pathways are normally involved in implicit processing, this may help to give an account of why some implicit (unconscious) performance is possible in blindsight patients. Further, it has been suggested that ventral visual stream processing is especially involved in consciousness, because it is information about objects and faces that needs to enter a system to select and plan actions (Milner and Goodale 1995; Rolls 2008b), and the planning of actions that involves the

operation and correction of flexible one-time multiple-step plans may be closely related to conscious processing (Rolls 1999b, 2005b, 2008b). In contrast, dorsal stream visual processing may be more closely related to executing an action on an object once the action has been selected, and the details of this action execution can take place implicitly (unconsciously) (Milner and Goodale 1995; Rolls 2008b), perhaps because they do not require multiple step syntactic planning (Rolls 1999b, 2005b, 2008b).

One of the implications of blindsight thus seems to be that some visual pathways are more involved in implicit processing and other pathways in explicit processing. In contrast, the results described here suggest that short and information-poor signals in a sensory system involved in conscious processing do not reach consciousness and do not interrupt ongoing or engage conscious processing. This evidence described here thus provides interesting and direct evidence that there may be a threshold for activity in a sensory stream that must be exceeded in order to lead to consciousness, even when that activity is sufficient for some types of visual processing such as visual identification of faces at well above chance in an implicit mode. The latter implicit mode processing can be revealed by forced choice tests and by direct measurements of neuronal responses. Complementary evidence at the purely psychophysical level using backward masking has been obtained by Marcel (1983a, b) and discussed by Weiskrantz (1998, 2001). Possible reasons for this relatively high threshold for consciousness are considered above.

9.7 The Speed of Visual Processing Within a Cortical Visual Area Shows That Top-Down Interactions with Bottom-Up Processes Are Not Essential for Conscious Visual Perception

The results of the information analysis of backward masking (Rolls et al. 1999) emphasize that very considerable information about which stimulus was shown is available in a short epoch of, for example, 50 ms of neuronal firing. This confirms and is consistent with many further findings on the speed of processing of inferior temporal cortex neurons (Tovee et al. 1993; Tovee and Rolls 1995; Rolls et al. 2006; Rolls 2008b) and facilitates the rapid read-out of information from the inferior temporal visual cortex. One direct implication of the 30 ms firing with the 20 ms S.O.A. is that this is sufficient time both for a cortical area to perform its computation and for the information to be read out from a cortical area, given that psychophysical performance is 50% correct at this S.O.A. Another implication is that the recognition of visual stimuli can be performed using feedforward processing in the multi-stage hierarchically organized ventral visual system comprising at least V1–V2–V4–Inferior Temporal Visual Cortex, in that the typical shortest neuronal response latencies in macaque V1 are approximately 40 ms, and increase by approximately 15–17 ms per stage to produce a value of approximately 90 ms in the inferior temporal visual cortex (Dinse and Kruger 1994; Nowak and Bullier 1997;

Rolls and Deco 2002; Rolls 2008b). Given these timings, it would not be possible in the 20 ms S.O.A. condition for inferior temporal cortex neuronal responses to feed back to influence V1 neuronal responses to the test stimulus before the mask stimulus produced its effects on the V1 neurons. (In an example, in the 20 ms S.O.A. condition with 30 ms of firing, the V1 neurons would stop responding to the stimulus at 40 + 30 = 70 ms, but would not be influenced by backprojected information from the inferior temporal cortex until 90 + (3 stages ×15 ms per stage) = 135 ms. In another example for conscious processing, in the 40 ms S.O.A. condition with 50 ms of firing, the V1 neurons would stop responding to the stimulus at 40 + 50 = 90 ms, but would not be influenced by backprojected information from the inferior temporal cortex until 90 + (3 stages ×15 ms per stage) = 135 ms.) This shows that not only recognition, but also conscious awareness, of visual stimuli is possible without top-down backprojection effects from the inferior temporal visual cortex to early cortical processing areas that could interact with the processing in the early cortical areas.

The same information theoretic analyses (Tovee et al. 1993; Tovee and Rolls 1995; Rolls et al. 1999, 2006; Rolls 2008b) show that from the earliest spikes of the anterior inferior temporal cortex neurons described here after they have started to respond (at approximately 80 ms after stimulus onset), the neuronal response is specific to the stimulus, and it is only in more posterior parts of the inferior temporal visual cortex that neurons may have an earlier short period of firing (of perhaps 20 ms) which is not selective for a particular stimulus. The neurons described by Sugase et al. (1999) thus behaved like more posterior inferior temporal cortex neurons and not like typical anterior inferior temporal cortex neurons. This evidence thus suggests that in the anterior inferior temporal cortex, recurrent processing may help to sharpen up representations to minimize early non-specific firing (cf Lamme and Roelfsema 2000).

9.8 Comparisons with Other Approaches to Consciousness

The theory described here suggests that it feels like something to be an organism or machine that can think about its own (linguistic and semantically based) thoughts. It is suggested that qualia, raw sensory and emotional subjective feelings arise secondary to having evolved such a HOST system, and that sensory and emotional processing feels like something because it would be unparsimonious for it to enter the planning, HOST system and *not* 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. Some issues that arise in relation to this theory are discussed by Rolls (2000, 2004b, 2005b); reasons why the ventral visual system is more closely related to explicit than implicit processing (because reasoning about objects may be important) are considered by Rolls (2003) and by

Rolls and Deco (2002); and reasons why explicit, conscious, processing may have a higher threshold in sensory processing than implicit processing are considered by Rolls (2003, 2005a, b).

I now compare this approach to consciousness with those that place emphasis on working memory (LeDoux 2008). LeDoux (1996), 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 (p. 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 (1996) concedes that consciousness, especially its phenomenal or subjective nature, is not completely explained by the computational processes that underlie working memory (p. 281).

LeDoux (2008) notes that the term "working memory" can refer to a number of different processes. In attentional systems, a short term memory is needed to hold on-line the subject of the attention, for example the position in space at which an object must be identified. There is much evidence that this short term memory is implemented in the prefrontal cortex by an attractor network implemented by associatively modifiable recurrent collateral connections between cortical pyramidal cells, which keep the population active during the attentional task. This short term memory then biases posterior perceptual and memory networks in the temporal and parietal lobes in a biased competition process (Miller and Cohen 2001; Rolls and Deco 2002; Deco and Rolls 2005a, b; Rolls 2008b). The operation of this type of short term memory acting using biased competition to implement top-down attention does not appear to be central to consciousness, for as LeDoux (2008) agrees, prefrontal cortex lesions that have major effects on attention do not impair subjective feelings of consciousness. The same evidence suggests that attention itself is not a fundamental computational process that is necessary for consciousness, as the neural networks that implement short term memory and operate to produce biased competition with non-linear effects do not appear to be closely related to consciousness (Deco and Rolls 2005a; Rolls 2008b), though of course if attention is directed towards particular perceptual events, this will increase the gain of the perceptual processing (Deco and Rolls 2005a, b; Rolls 2008b), making the attended phenomena stronger.

Another process ascribed to working memory is that items can be manipulated in working memory, for example, placed into a different order. This process implies at the computational level some type of syntactic processing, for each item (or symbol) could occur in any position relative to the others, and each item might occur more than once. To keep the items separate yet manipulable into any relation to each other, just having each item represented by the firing of a different set of neurons is insufficient, for this provides no information about the order or more generally the relations between the items being manipulated (Rolls and Deco 2002; Rolls 2008b). In this sense, some form of syntax, that is a way to relate to each other the firing of the different populations of neurons each representing an item, is required. If we go this far (and LeDoux (1996) p. 280 does appear to), then we see that this aspect

of working memory is very close to the concept I propose of syntactic thought in my HOST theory. My particular approach though makes it clear what the function is to be performed (syntactic operations), whereas the term working memory can be used to refer to many different types of processing and is in this sense less well defined computationally. My approach of course argues that it is thoughts about the first order thoughts that may be very closely linked to consciousness. In our simple case, the HOST might be "Do I have the items now in the correct reversed order? Should the *X* come before or after the *Y*?" To perform this syntactic manipulation, I argue that there is a special syntactic processor, perhaps in cortex near Broca's area, that performs the manipulations on the items, and that the dorsolateral prefrontal cortex itself provides the short-term store that holds the items on which the syntactic processor operates (Rolls 2008b). In this scenario, dorsolateral prefrontal cortex damage would affect the number of items that could be manipulated, but not consciousness or the ability to manipulate the items syntactically and to monitor and comment on the result to check that it is correct.

A property often attributed to consciousness is that it is *unitary*. LeDoux (2008) might relate this to the limitations of a working memory system. 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 (McLeod et al. 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 relation between, and the limited capacity of, both the stream of consciousness, and auditory-verbal short term memory, may arise because both require implementation of the capacity for syntax in neural networks. My suggestion is that it is the difficulty the brain has in implementing the syntax required for manipulating items in working memory, and therefore for multiple step planning, and for then correcting these plans, that provides a close link between working memory concepts and my theory of higher order syntactic processing. The theory I describe makes it clear what the underlying computational problem is (how syntactic operations are performed in the system, and how they are corrected), and argues that when there are thoughts about the system, i.e. HOSTs, and the system is reflecting on its first order thoughts (cf. Weiskrantz 1997), then it is a property of the system that it feels conscious. As I argued above, first order linguistic thoughts, which presumably involve working memory (which must be clearly defined for the purposes of this discussion), need not necessarily be conscious.

The theory is also different from some other theories of consciousness (Carruthers 1996; Gennaro 2004; Rosenthal 2004, 2005) in that it provides an account of the evolutionary, adaptive, value of a HOST system in helping to solve a credit assignment problem that arises in a multi-step syntactic plan, links this type of processing to consciousness and therefore emphasizes a role for syntactic processing in consciousness. 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 or simpler, as it involves primarily the syntactic manipulation of symbols (Fodor 1994; Rolls 2004b, 2005b).

The current theory holds that it is HOSTs that are closely associated with consciousness, and this may differ from Rosenthal's 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 multi-step flexible "on the fly" reasoning procedure that errors which cannot be easily corrected by reward or punishment received at the end of the reasoning need "thoughts about thoughts", that is some type of supervisory and monitoring process, to detect where errors in the reasoning have occurred. This suggestion on the adaptive value in evolution of such a higher order linguistic thought process for multi-step planning ahead, and correcting such plans, may also be different from earlier work. Put another way, this point is that *credit assignment* when reward or punishment are received is straightforward in a one layer network (in which the reinforcement can be used directly to correct nodes in error, or responses), but is very difficult in a multi-step linguistic process executed once "on the fly". Very complex mappings in a multilayer network can be learned if hundreds of learning trials are provided. But once these complex mappings are learned, their success or failure in a new situation on a given trial cannot be evaluated and corrected by the network. Indeed, the complex mappings achieved by such networks (e.g. backpropagation nets) mean that after training they operate according to fixed rules and are often quite impenetrable and inflexible (Rolls and Deco 2002). In contrast, to correct a multi-step, single occasion, linguistically based plan or procedure, recall of the steps just made in the reasoning or planning, and perhaps related episodic material, needs to occur, so that the link in the chain which is most likely to be in error can be identified. This may be part of the reason why there is a close relation between declarative memory systems, which can explicitly recall memories and consciousness.

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

Another possible difference from other theories of consciousness 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.

Finally, I provide a short specification of what might have to be implemented in a neural network to implement conscious processing. First, a linguistic system, not

necessarily verbal, but implementing syntax between symbols grounded in the environment would be needed (e.g. a mentalese language system). Then a HOST system also implementing syntax and able to think about the representations in the first order language system, and able to correct the reasoning in the first order linguistic system in a flexible manner, would be needed. So my view is that consciousness can be implemented in neural networks of the artificial and biological type, but that the neural networks would have to implement the type of higher order linguistic processing described in this paper.

Acknowledgements The author acknowledges helpful discussions with Martin Davies, Marian Dawkins and David Rosenthal.

References

Aggelopoulos NC, Rolls ET (2005) Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. European Journal of Neuroscience 22:2903–2916.

Aggelopoulos NC, Franco L, Rolls ET (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 (Marcel AJ, Bisiach E, eds), pp 159–182. Oxford: Oxford University Press.

Armstrong DM, Malcolm N (1984) Consciousness and Causality. Oxford: Blackwell.

Azzopardi P, Cowey A (1997) Is blindsight like normal, near-threshold vision? Proceedings of the National Academy of Sciences USA 94:14190–14194.

Baars BJ (1988) A Cognitive Theory of Consciousness. New York: Cambridge University Press.

Barlow HB (1997) Single neurons, communal goals, and consciousness. In: Cognition, Computation, and Consciousness (Ito M, Miyashita Y, Rolls ET, eds), pp 121–136. Oxford: Oxford University Press.

Battaglia FP, Treves A (1998) Stable and rapid recurrent processing in realistic auto-associative memories. Neural Computation 10:431–450.

Block N (1995) On a confusion about a function of consciousness. Behavioral and Brain Sciences 18:227–247.

Block N (2005) Two neural correlates of consciousness. Trends in Cognitive Sciences 9:46–52.

Booth DA (1985) Food-conditioned eating preferences and aversions with interoceptive elements: learned appetites and satieties. Annals of the New York Academy of Sciences 443:22–37.

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

Chalmers DJ (1996) The Conscious Mind. Oxford: Oxford University Press.

Cheney DL, Seyfarth RM (1990) How Monkeys See the World. Chicago: University of Chicago Press.

Cooney JW, Gazzaniga MS (2003) Neurological disorders and the structure of human consciousness. Trends in Cognitive Sciences 7:161–165.

Crick FHC, Koch C (1990) Towards a neurobiological theory of consciousness. Seminars in the Neurosciences 2:263–275.

Damasio AR (1994) Descartes' Error. New York: Putnam.

Davies MK (2008) Consciousness and explanation. In: Frontiers of Consciousness (Weiskrantz L, Davies MK, eds), pp 1–54. Oxford: Oxford University Press.

Dawkins R (1986) The Blind Watchmaker. Harlow: Longman.

Dawkins R (1989) The Selfish Gene, 2nd Edition. Oxford: Oxford University Press.
De Gelder B, Vroomen J, Pourtois G, Weiskrantz L (1999) Non-conscious recognition of affect in the absence of striate cortex. Neuroreport 10:3759–3763.
Deco G, Rolls ET (2005a) Neurodynamics of biased competition and co-operation for attention: a model with spiking neurons. Journal of Neurophysiology 94:295–313.
Deco G, Rolls ET (2005b) Attention, short-term memory, and action selection: a unifying theory. Progress in Neurobiology 76:236–256.
Deco G, Rolls ET (2005c) Synaptic and spiking dynamics underlying reward reversal in orbitofrontal cortex. Cerebral Cortex 15:15–30.
Deco G, Rolls ET (2006) Decision-making and Weber's Law: a neurophysiological model. European Journal of Neuroscience 24:901–916.
Deco G, Rolls ET, Romo R (2009) Stochastic dynamics as a principle of brain function. Progress in Neurobiology 88:1–16.
Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79:1–37.
Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences 10:204–211.
Dennett DC (1991) Consciousness Explained. London: Penguin.
Dennett DC (2005) Sweet Dreams: Philosophical Obstacles to a Science of Consciousness. Cambridge, MA: MIT.
Dinse HR, Kruger K (1994) The timing of processing along the visual pathway in the cat. Neuroreport 5:893–897.
Elliffe MCM, Rolls ET, Stringer SM (2002) Invariant recognition of feature combinations in the visual system. Biological Cybernetics 86:59–71.
Fodor JA (1994) The Elm and the Expert: mentalese and its semantics. Cambridge, MA: MIT.
Franco L, Rolls ET, Aggelopoulos NC, 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 MS (1988) Brain modularity: towards a philosophy of conscious experience. In: Consciousness in Contemporary Science (Marcel AJ, Bisiach E, eds), pp 218–238. Oxford: Oxford University Press.
Gazzaniga MS (1995) Consciousness and the cerebral hemispheres. In: The Cognitive Neurosciences (Gazzaniga MS, ed), pp 1392–1400. Cambridge, Mass.: MIT.
Gazzaniga MS, LeDoux J (1978) The Integrated Mind. New York: Plenum.
Gennaro RJ, ed (2004) Higher Order Theories of Consciousness. Amsterdam: John Benjamins.
Goldman-Rakic PS (1996) The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. Philosophical Transactions of the Royal Society B 351:1445–1453.
Goodale MA (2004) Perceiving the world and grasping it: dissociations between conscious and unconscious visual processing. In: The Cognitive Neurosciences III (Gazzaniga MS, ed), pp 1159–1172. Cambridge, MA: MIT.
Hamilton W (1964) The genetical evolution of social behaviour. Journal of Theortical Biology 7:1–52.
Hamilton WD (1996) Narrow Roads of Gene Land. New York: W. H. Freeman.
Hampton RR (2001) Rhesus monkeys know when they remember. Proceedings of the National Academy of Sciences of the United States of America 98:5359–5362.
Hampton RR, Zivin A, Murray EA (2004) Rhesus monkeys (Macaca mulatta) discriminate between knowing and not knowing and collect information as needed before acting. Animal Cognition 7:239–246.
Heyes CM (2008) Beast machines? Questions of animal consciousness. In: Frontiers of Consciousness (Weiskrantz L, Davies M, eds), pp 259–274. Oxford: Oxford University Press.
Hornak J, Bramham J, Rolls ET, Morris RG, O'Doherty J, Bullock PR, Polkey CE (2003) Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. Brain 126:1691–1712.

Hornak J, O'Doherty J, Bramham J, Rolls ET, Morris RG, Bullock PR, Polkey CE (2004) Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. Journal of Cognitive Neuroscience 16:463–478.

Humphrey NK (1980) Nature's psychologists. In: Consciousness and the Physical World (Josephson BD, Ramachandran VS, eds), pp 57–80. Oxford: Pergamon.

Humphrey NK (1986) The Inner Eye. London: Faber.

Humphreys GW, Bruce V (1991) Visual Cognition. Hove, East Sussex: Erlbaum.

Huxter J, Burgess N, O'Keefe J (2003) Independent rate and temporal coding in hippocampal pyramidal cells. Nature 425:828–832.

Insabato A, Pannunzi M, Rolls ET, Deco G (2010) Confidence-related decision-making. Journal of Neurophysiology 104:539–547.

Jackendoff R (2002) Foundations of Language. Oxford: Oxford University Press.

Johnson-Laird PN (1988) The Computer and the Mind: An Introduction to Cognitive Science. Cambridge, MA: Harvard University Press.

Kadohisa M, Rolls ET, Verhagen JV (2005) Neuronal representations of stimuli in the mouth: the primate insular taste cortex, orbitofrontal cortex, and amygdala. Chemical Senses 30:401–419.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. Nature 455:227–231.

Koch C, Preuschoff K (2007) Betting the house on consciousness. Nature Neuroscience 10:140–141.

Krebs JR, Kacelnik A (1991) Decision Making. In: Behavioural Ecology, 3rd Edition (Krebs JR, Davies NB, eds), pp 105–136. Oxford: Blackwell.

Lamme VAF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neuroscience 23:571–579.

LeDoux J (2008) Emotional coloration of consciousness: how feelings come about. In: Frontiers of Consciousness (Weiskrantz L, Davies M, eds), pp 69–130. Oxford: Oxford University Press.

LeDoux JE (1996) The Emotional Brain. New York: Simon and Schuster.

Libet B (2002) The timing of mental events: Libet's experimental findings and their implications. Consciousness and Cognition 11:291–299; discussion 304–333.

Malsburg Cvd (1990) A neural architecture for the representation of scenes. In: Brain Organization and Memory: Cells, Systems and Circuits (McGaugh JL, Weinberger NM, Lynch G, eds), pp 356–372. New York: Oxford University Press.

Marcel AJ (1983a) Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes. Cognitive Psychology 15:238–300.

Marcel AJ (1983b) Conscious and unconscious perception: experiments on visual masking and word recognition. Cognitive Psychology 15:197–237.

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

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annual review of neuroscience 24:167–202.

Milner AD (2008) Conscious and unconscious visual processing in the human brain. In: Frontiers of Consciousness (Weiskrantz L, Davies M, eds), pp 169–214. Oxford: Oxford University Press.

Milner AD, Goodale MA (1995) The Visual Brain in Action. Oxford: Oxford University Press.

Nowak LG, Bullier J (1997) The timing of information transfer in the visual system. In: Extrastriate visual cortex in primates (Rockland KS, Kaas JH, Peters A, eds), pp 205–241. New York: Plenum.

Panzeri S, Schultz SR, Treves A, Rolls ET (1999) Correlations and the encoding of information in the nervous system. Proceedings of the Royal Society of London B 266:1001–1012.

Panzeri S, Rolls ET, Battaglia F, Lavis R (2001) Speed of feedforward and recurrent processing in multilayer networks of integrate-and-fire neurons. Network: Computation in Neural Systems 12:423–440.

Persaud N, McLeod P, Cowey A (2007) Post-decision wagering objectively measures awareness. Nature Neuroscience 10:257–261.

Petrides M (1996) Specialized systems for the processing of mnemonic information within the primate frontal cortex. Philosophical Transactions of the Royal Society B 351:1455–1462.
Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175–187.
Ridley M (1993) The Red Queen: Sex and the Evolution of Human Nature. London: Penguin.
Rolls ET (1990) A theory of emotion, and its application to understanding the neural basis of emotion. Cognition and Emotion 4:161–190.
Rolls ET (1995) A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In: The Cognitive Neurosciences (Gazzaniga MS, ed), pp 1091–1106. Cambridge, Mass.: MIT.
Rolls ET (1997a) Brain mechanisms of vision, memory, and consciousness. In: Cognition, Computation, and Consciousness (Ito M, Miyashita Y, Rolls ET, eds), pp 81–120. Oxford: Oxford University Press.
Rolls ET (1997b) Consciousness in neural networks? Neural Networks 10:1227–1240.
Rolls ET (1999a) The functions of the orbitofrontal cortex. Neurocase 5:301–312.
Rolls ET (1999b) The Brain and Emotion. Oxford: Oxford University Press.
Rolls ET (2000) Précis of The Brain and Emotion. Behavioral and Brain Sciences 23:177–233.
Rolls ET (2003) Consciousness absent and present: a neurophysiological exploration. Progress in Brain Research 144:95–106.
Rolls ET (2004a) The functions of the orbitofrontal cortex. Brain and cognition 55:11–29.
Rolls ET (2004b) A higher order syntactic thought (HOST) theory of consciousness. In: Higher-Order Theories of Consciousness: An Anthology (Gennaro RJ, ed), pp 137–172. Amsterdam: John Benjamins.
Rolls ET (2005a) Consciousness absent or present: a neurophysiological exploration of masking. In: The First Half Second: The Microgenesis and Temporal Dynamics of Unconscious and Conscious Visual Processes (Ogmen H, Breitmeyer BG, eds), pp 89–108, chapter 106. Cambridge, MA: MIT.
Rolls ET (2005b) Emotion Explained. Oxford: Oxford University Press.
Rolls ET (2006) Brain mechanisms underlying flavour and appetite. Philosophical Transactions of the Royal Society London B 361:1123–1136.
Rolls ET (2007a) A computational neuroscience approach to consciousness. Neural Networks 20:962–982.
Rolls ET (2007b) The representation of information about faces in the temporal and frontal lobes. Neuropsychologia 45:125–143.
Rolls ET (2007c) The affective neuroscience of consciousness: higher order linguistic thoughts, dual routes to emotion and action, and consciousness. In: Cambridge Handbook of Consciousness (Zelazo P, Moscovitch M, Thompson E, eds), pp 831–859. Cambridge: Cambridge University Press.
Rolls ET (2008a)Emotion, higher order syntactic thoughts, and consciousness. In: Frontiers of Consciousness (Weiskrantz L, Davies MK, eds), pp 131–167. Oxford: Oxford University Press.
Rolls ET (2008b) Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach. Oxford: Oxford University Press.
Rolls ET (2009) The anterior and midcingulate cortices and reward. In: Cingulate Neurobiology and Disease (Vogt BA, ed), pp 191–206. Oxford: Oxford University Press.
Rolls ET (2011) Neuroculture. Oxford: Oxford University Press.
Rolls ET, Tovee MJ (1994) Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proceedings of the Royal Society of London B 257:9–15.
Rolls ET, Treves A (1998) Neural Networks and Brain Function. Oxford: Oxford University Press.
Rolls ET, Deco G (2002) Computational Neuroscience of Vision. Oxford: Oxford University Press.
Rolls ET, Stringer SM (2006) Invariant visual object recognition: a model, with lighting invariance. Journal of Physiology Paris 100:43–62.
Rolls ET, Kesner RP (2006) A computational theory of hippocampal function, and empirical tests of the theory. Progress in Neurobiology 79:1–48.

Rolls ET, Grabenhorst F (2008) The orbitofrontal cortex and beyond: from affect to decision-making. Progress in Neurobiology 86:216–244.
Rolls ET, Deco G (2010) The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function. Oxford: Oxford University Press.
Rolls ET, Tovee MJ, Panzeri S (1999) The neurophysiology of backward visual masking: information analysis. Journal of Cognitive Neuroscience 11:335–346.
Rolls ET, Grabenhorst F, Deco G (2010a) Choice, difficulty, and confidence in the brain. NeuroImage 53:694–706.
Rolls ET, Grabenhorst F, Deco G (2010) Decision-making, errors, and confidence in the brain. Journal of Neurophysiology 104:2359–2374.
Rolls ET, 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 ET, Franco L, Aggelopoulos NC, Reece S (2003) An information theoretic approach to the contributions of the firing rates and correlations between the firing of neurons. Journal of Neurophysiology 89:2810–2822.
Rolls ET, Aggelopoulos NC, Franco L, Treves A (2004) Information encoding in the inferior temporal cortex: contributions of the firing rates and correlations between the firing of neurons. Biological Cybernetics 90:19–32.
Rolls ET, Franco L, Aggelopoulos NC, Perez JM (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research 46:4193–4205.
Rolls ET, Tovee MJ, Purcell DG, Stewart AL, Azzopardi P (1994b) The responses of neurons in the temporal cortex of primates, and face identification and detection. Experimental Brain Research 101:473–484.
Rosenthal DM (1986) Two concepts of consciousness. Philosophical Studies 49:329–359.
Rosenthal DM (1990) A Theory of Consciousness. Bielefeld, Germany: Zentrum für Interdisziplinaire Forschung.
Rosenthal DM (1993) Thinking that one thinks. In: Consciousness (Davies M, Humphreys GW, eds), pp 197–223. Oxford: Blackwell.
Rosenthal DM (2004) Varieties of Higher-Order Theory. In: Higher Order Theories of Consciousness (Gennaro RJ, ed). Amsterdam: John Benjamins.
Rosenthal DM (2005) Consciousness and Mind. Oxford: Oxford University Press.
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Rumelhart DE, McClelland JL, Group TPR, eds). Cambridge, Mass.: MIT.
Shallice T, Burgess P (1996) The domain of supervisory processes and temporal organization of behaviour. Philosophical Transactions of the Royal Society B 351:1405–1411.
Simmen MW, Treves A, Rolls ET (1996) Pattern retrieval in threshold-linear associative nets. Network (Bristol, England) 7:109–122.
Singer W (1999) Neuronal synchrony: A versatile code for the definition of relations? Neuron 24:49–65.
Smerieri A, Rolls ET, Feng J (2010) Decision time, slow inhibition, and theta rhythm. Journal of Neuroscience 30:14173–14181.
Smith-Swintosky VL, Plata-Salaman CR, Scott TR (1991) Gustatory neural encoding in the monkey cortex: stimulus quality. Journal of Neurophysiology 66:1156–1165.
Smith PL, Ratcliff R (2004) Psychology and neurobiology of simple decisions. Trends in Neurosciences 27:161–168.
Squire LR, Zola SM (1996) Structure and function of declarative and nondeclarative memory systems. Proceedings of the National Academy of Sciences USA 93:13515–13522.
Sugase Y, Yamane S, Ueno S, Kawano K (1999) Global and fine information coded by single neurons in the temporal visual cortex. Nature 400:869–873.
Tovee MJ, Rolls ET (1992) The functional nature of neuronal oscillations. Trends in Neurosciences 15:387.

Tovee MJ, Rolls ET (1995) Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Visual Cognition 2:35–58.
Tovee MJ, Rolls ET, Treves A, Bellis RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. Journal of Neurophysiology 70:640–654.
Treisman A (1996) The binding problem. Current Opinion in Neurobiology 6:171–178.
Treves A (1993) Mean-field analysis of neuronal spike dynamics. Network (Bristol, England) 4:259–284.
Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36:955–968.
Weiskrantz L (1997) Consciousness Lost and Found. Oxford: Oxford University Press.
Weiskrantz L (1998) Blindsight. A Case Study and Implications, 2nd Edition. Oxford: Oxford University Press.
Weiskrantz L (2001) Blindsight – putting beta (β) on the back burner. In: Out of Mind: Varieties of Unconscious Processes (De Gelder B, De Haan E, Heywood C, eds), pp 20–31. Oxford: Oxford University Press.
Wong KF, Wang XJ (2006) A recurrent network mechanism of time integration in perceptual decisions. Journal of Neuroscience 26:1314–1328.
Yaxley S, Rolls ET, Sienkiewicz ZJ (1990) Gustatory responses of single neurons in the insula of the macaque monkey. Journal of Neurophysiology 63:689–700.