A Higher Order Syntactic Thought (HOST) Theory of Consciousness

Edmund T. Rolls
University of Oxford, Department of Experimental Psychology,
South Parks Road, Oxford OX1 3UD, England

Please address correspondence to Professor E.T. Rolls
Tel: +44-1865-271348
Fax: +44-1865-310447
Email: Edmund.Rolls@psy.ox.ac.uk
Web: www.cns.ox.ac.uk

.pdf copy of:

Rolls, E.T. (2004) A higher order syntactic thought (HOST) theory of consciousness. Ch 7, pp. 137-172 in Higher Order Theories of Consciousness. Ed. R. J.Gennaro. John Benjamins: Amsterdam.

1 Background

The background to the HOST theory of consciousness described here is a theory of emotion based on the neuroscience of emotion (Rolls 1999a; 2000e; 1990), as described in this section.

1a. A theory of emotion

Rolls' theory of emotion holds that emotions can usefully be defined as states elicited by rewards and punishers which have particular functions (Rolls 1999a). A reward is anything for which an animal (which includes humans) will work. A punisher is anything that an animal will escape from or avoid. An example of an emotion might thus be happiness produced by being given a reward, such as a pleasant touch, praise, or winning a large sum of money. Another example of an emotion might be fear produced by the sound of a rapidly approaching bus, or the sight of an angry expression on someone's face. We will work to avoid such stimuli, which are punishing. Another example would be frustration, anger, or sadness produced by the omission of an expected reward such as a prize, or the termination of a reward such as the death of a loved one. Another example would be relief, produced by the omission or termination of a punishing stimulus such as the removal of a painful stimulus, or sailing out of danger. These examples indicate how emotions can be produced by the delivery, omission, or termination of rewarding or punishing stimuli, and go some way to indicate how different emotions could be produced and classified in terms of the rewards and punishments received, omitted, or terminated. This approach has been greatly extended to account for many types of emotion (Rolls 1999a; 2000e; 1990).

It is worth raising the issue that philosophers usually categorize fear in the example as an emotion, but not pain. The distinction they make may be that primary (unlearned) reinforcers do not produce emotions, whereas secondary reinforcers (stimuli associated by stimulus-reinforcement learning with primary reinforcers) do. They describe the pain as a sensation. But neutral stimuli (such as a table) can produce sensations when touched. It accordingly seems to be much more useful to categorise stimuli according to whether they are reinforcing (in which case they produce emotions), or are not reinforcing (in which case they do not produce emotions). Clearly there is a difference between primary reinforcers and learned reinforcers; but this is most precisely caught by noting that this is the difference, and that it is whether a stimulus is reinforcing that determines whether it is related to emotion.

1b. The Functions of Emotion

The functions of emotion also provide insight into the nature of emotion. These functions, described more fully elsewhere (Rolls 1999a; 2000e; 1990) can be summarized as follows:

1. The *elicitation of autonomic responses* (e.g., a change in heart rate) *and endocrine responses* (e.g., the release of adrenaline). These prepare the body for action.

2. *Flexibility of behavioral responses to reinforcing stimuli*. Emotional (and motivational) states allow a simple interface between sensory inputs and action systems. The essence of this idea is that goals for behavior are specified by reward and punishment evaluation. When an environmental stimulus has been decoded as a primary reward or punishment, or (after previous stimulus-reinforcer association learning) a secondary rewarding or punishing stimulus, then it becomes a goal for action. The animal can then perform any action (instrumental response) to obtain the reward, or to avoid the punisher. Thus there is flexibility of action, and this is in contrast with stimulus-response, or habit, learning in which a particular response to a particular stimulus is learned. The emotional route to action is flexible not only because any action can be performed to obtain the reward or avoid the punishment, but also because the animal can learn in as little as one trial that a reward or punishment is associated with a particular stimulus, in what is termed "stimulus-reinforcer association learning".

To summarize and formalize, two processes are involved in the actions being described. The first is stimulus-reinforcer association learning, and the second is instrumental learning of an operant response made to approach and obtain the reward or to avoid or escape from the punisher. Emotion is an integral part of this, for it is the state elicited in the first stage, by stimuli which are decoded as rewards or punishers, and this state has the property that it is motivating. The motivation is to obtain the reward or avoid the punisher, and animals must be built to obtain certain rewards and avoid certain punishers. Indeed, primary or unlearned rewards and punishers are specified by genes which effectively specify the

goals for action. This, Rolls proposes (Rolls 1999a) is the solution which natural selection has found for how genes can influence behavior to promote their fitness (as measured by reproductive success), and for how the brain could interface sensory systems to action systems.

Selecting between available rewards with their associated costs, and avoiding punishers with their associated costs, is a process which can take place both implicitly (unconsciously), and explicitly using a language system to enable long-term plans to be made (Rolls 1999a). These many different brain systems, some involving implicit evaluation of rewards, and others explicit, verbal, conscious, evaluation of rewards and planned long-term goals, must all enter into the selector of behavior (see Fig. 1). This selector is poorly understood, but it might include a process of competition between all the competing calls on output, and might involve the basal ganglia in the brain (see Fig. 1 and Rolls 1999a).

[Insert Fig. 1 near here]

3. Emotion is *motivating*, as just described. For example, fear learned by stimulus-reinforcement association provides the motivation for actions performed to avoid noxious stimuli.

4. *Communication*. Monkeys for example may communicate their emotional state to others, by making an open-mouth threat to indicate the extent to which they are willing to compete for resources, and this may influence the behavior of other animals. This aspect of emotion was emphasized by Darwin (Darwin 1872) and has been studied more recently by Ekman (1982; 1993). As shown elsewhere (Rolls 2000a; 2000c), there are neural systems in the amygdala and overlying temporal cortical visual areas which are specialized for the face-related aspects of this processing.

5. *Social bonding*. Examples of this are the emotions associated with the attachment of the parents to their young, and the attachment of the young to their parents.

6. The current mood state can affect the *cognitive evaluation of events or memories* (see Oatley and Jenkins 1996). This may facilitate continuity in the interpretation of the reinforcing value of events in the environment. A hypothesis that backprojections from parts of the brain involved in emotion such as the orbitofrontal cortex and amygdala implement this is described in *The Brain and Emotion*.

7. Emotion may facilitate the *storage of memories*. One way this occurs is that episodic memory (i.e., one's memory of particular episodes) is facilitated by emotional states. This may be advantageous in that storing many details of the prevailing situation when a strong reinforcer is delivered may be useful in generating appropriate behavior in situations with some similarities in the future. This function may be implemented by the relatively nonspecific projecting systems to the cerebral cortex and hippocampus, including the cholinergic pathways in the basal forebrain and medial septum, and the ascending noradrenergic pathways (Rolls 1999a; Rolls and Treves 1998). A second way in which emotion may affect the storage of memories is that the current emotional state may be stored with episodic memories, providing a mechanism for the current emotional state to affect which memories are recalled. A third way that emotion may affect the storage of memories is by guiding the cerebral cortex in the representations of the world which are set up. For example, in the visual system it may be useful for perceptual representations or analyzers to be built which are different from each other if they are associated with different reinforcers, and for these to be less likely to be built if they have no association with reinforcement. Ways in which backprojections from parts of the brain important in emotion (such as the amygdala) to parts of the cerebral cortex could perform this function are discussed by (Rolls and Treves 1998).

8. Another function of emotion is that by enduring for minutes or longer after a reinforcing stimulus has occurred, it may help to produce *persistent and continuing motivation and direction of behavior*, to help achieve a goal or goals.

9. Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala backprojections to the cortex could perform this for emotion in a way analogous to that in which the hippocampus could implement the retrieval in the neocortex of recent (episodic) memories (Rolls and Stringer 2001; Rolls and Treves 1998).

1c. To what extent is consciousness involved in the different types of processing initiated by emotional states?

It might be possible to build a computer which would perform the functions of emotions described above and in more detail by Rolls (1999a; 2000e), and yet we might not want to ascribe

emotional *feelings* to the computer. We might even build the computer with some of the main processing stages present in the brain, and implemented using neural networks which simulate the operation of the real neural networks in the brain (Rolls and Deco 2002; Rolls and Treves 1998), yet we might not still wish to ascribe emotional feelings to this computer. In a sense, the functions of reward and punishment in emotional behaviour are described by the above types of process and their underlying brain mechanisms in structures such as the amygdala and orbitofrontal cortex as described by Rolls (1999a; 2000c; 2000d), but what about the subjective aspects of emotion, what about the pleasure? A similar point arises when we consider the parts of the taste, olfactory, and visual systems in which the reward value of the taste, smell and sight of food are represented. One such brain region is the orbitofrontal cortex (Rolls 1999a; 2002; 2000d; 1997b). Although the neuronal representation in the orbitofrontal cortex is clearly related to the reward value of food, is this where the pleasantness (the subjective hedonic aspect) of the taste, smell and sight of food is represented? Again, we could (in principle at least) build a computer with neural networks to simulate each of the processing stages for the taste, smell and sight of food which are described by Rolls (1999a; and Rolls and Deco 2002; and more formally in terms of neural networks Rolls and Treves 1998), and yet would probably not wish to ascribe feelings of pleasantness to the system we have simulated on the computer. The point I am making here is that much of the processing related to the control of emotional and motivational behavior could take place without any need to invoke the possibility to be doing the type of processing that leads to reward decoding, and the initiation of behavioral responses. Part of the evidence for this hypothesis is that patients with orbitofrontal cortex damage may not be able to implement reward-related associative learning and reversal, yet may be able to comment verbally and consciously on the behavioral choices that they should be making (Hornak 2003a; Hornak 2003b; Rolls et al. 1994) .This dissociation between different systems involved in the implementation of behavior is described further in sections 1d and 3 on dual routes to action. (I do believe that provided that the right type of processing is implemented in a computer, it would be conscious. In this sense I am a functionalist. In this Chapter I develop a hypothesis on what would have to be implemented in a computer for it to be conscious.)

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. 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 analyse (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 this Chapter is to address the issue of consciousness, especially of qualia, in relation to emotional feelings and actions, 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, at least, should not be taken to have practical implications.

1d. Dual routes to action, and consciousness.

According to the present formulation, there are two types of route to action performed in relation to reward or punishment in humans (Rolls 1999a; Rolls 2003). 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, by the processes of incentive motivation and sensory-specific satiety (Rolls 1999a). 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 which are concerned with the initiation of actions (see Fig. 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" conditional 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 in order to implement for example the conditional statements involved in each step of the plan. 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. 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. I emphasise that this type of processing would involve syntax, in that each step of the plan might have its own symbols that would need to be kept apart from those utilized in other steps of the plan, and further that each step of the plan might itself require syntax, of the type required to implement for example “if ... then” conditionals. The need for syntax can be illustrated also by considering a neural network in which different populations of neurons are active, each representing different symbols. If each symbol is represented by a set of neuronal firings, how does the system implement the relations between the symbols, e.g. that symbol A acts on symbol B and not vice versa. Thus need for syntax in neural networks has been recognised, and solutions to this binding problem such as stimulus-dependent synchronisation of the firing of different neuronal populations that need to

be related to each other have been proposed (Singer 1999), but do not seem to be adequate for the job required (Rolls and Deco 2002, section 13.7). 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 (see Goldman-Rakic 1996; Petrides 1996), and may be part of the reason why prefrontal cortex damage impairs planning (Rolls and Deco 2002; see Shallice and Burgess 1996).

Of these two routes (see Fig. 1), it is the second which I suggest is related to consciousness (see Rolls 1999a). 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 I argue of enabling one to correct long multi-step syntactic plans and thus solving a credit assignment problem, as described below. This second 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 verbalised 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. These points are developed next.

2. A Theory of Consciousness

A starting point is that many actions can be performed relatively automatically, without apparent conscious intervention. An example sometimes given is driving a car. Such actions could involve control of behaviour by brain systems which are old in evolutionary terms such as the basal ganglia. It is of interest that the basal ganglia (and cerebellum) do not have backprojection systems to most of the parts of the cerebral cortex from which they receive inputs (Rolls 1994; Rolls and Johnstone 1992; Rolls et al. 1998). In contrast, parts of the brain such as the hippocampus and amygdala, involved in functions such as episodic memory and emotion respectively, about which we can make (verbal) declarations (hence declarative memory, Squire 1992) do have major backprojection systems to the high parts of the cerebral cortex from which they receive forward projections (2000b; Rolls 1996; Rolls and Deco 2002; Rolls and Treves 1998; Treves and Rolls 1994). It may be that evolutionarily newer parts of the brain, such as the language areas and parts of the prefrontal cortex, are involved in an alternative type of control of behaviour, in which actions can be planned with the use of a (language) system which allows relatively arbitrary (syntactic) manipulation of semantic entities (symbols).

The general view that there are many routes to behavioural output is supported by the evidence that there are many input systems to the basal ganglia (from almost all areas of the cerebral cortex), and that neuronal activity in each part of the striatum reflects the activity in the overlying cortical area (Rolls 1994; Rolls and Johnstone 1992; Rolls and Treves 1998). The evidence is consistent with the possibility that different cortical areas, each specialised for a different type of computation, have their outputs directed to the basal ganglia, which then select the strongest input, and map this into action (via outputs directed for example to the premotor cortex) (Rolls and Johnstone 1992; Rolls and Treves 1998). Within

this scheme, the language areas would offer one of many routes to action, but a route particularly suited to planning actions, because of the syntactic manipulation of semantic entities which may make long-term planning possible. A schematic diagram of this suggestion is provided in Fig. 1.

Some of the evidence that supports the hypothesis of multiple routes to action, only some of which utilize language, is the evidence that split-brain patients may not be aware of actions being performed by the "non-dominant" hemisphere (Gazzaniga 1988; 1995; Gazzaniga and LeDoux 1978). Another important line of evidence for multiple, including non-verbal, routes to action, is that patients with focal brain damage, for example to the orbitofrontal cortex, may perform actions, yet comment verbally that they should not be performing those actions (Hornak 2003b; 1999b; Rolls et al. 1994). The actions which appear to be performed implicitly, with surprise expressed later by the explicit system, include making behavioral responses to a no-longer rewarded visual stimulus in a visual discrimination reversal (Hornak 2003b; Rolls 1994). 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 which actually triggered the action (Gazzaniga 1988; 1995; Gazzaniga and LeDoux 1978; Rolls 1994). It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region such as those involved in some types of rewarded behaviour, the language system may subsequently elaborate a coherent account of why that action was performed (i.e., confabulate). This would be consistent with a general view of brain evolution in which as areas of the cortex evolve, they are laid on top of existing circuitry connecting inputs to outputs, and in which each level in this hierarchy of separate input-output pathways may control behaviour according to the specialised function it can perform (see schematic in Fig. 1). (It is of interest that mathematicians may get a hunch that something is correct, yet not be able to verbalise why. They may then resort to formal, more serial and language-like, theorems to prove the case, and these seem to require conscious processing. This is a further indication of a close association between linguistic processing, and consciousness. The linguistic processing need not, as in reading, involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements, and to perform syntactic operations on the semantic representations of states in the world, would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g., I think that he thinks that ...), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others. (Second order thoughts are thoughts about thoughts. Higher order thoughts refer to second order, third order etc. thoughts about thoughts...) This capability for higher order thoughts would also enable reflection on past events, which would also be useful in planning, and in particular in correcting these multistep plans. In this sense, higher order thoughts I propose help to solve a credit assignment problem (see below). 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. This functionality is termed by some philosophers mentalese. The functionality required is not as strong as that required for natural language, which implies a universal grammar.

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 (cf. Dennett 1991; 1990; 1993; Rosenthal 1986). 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 focussed consciousness of one's mental states. It is noted that not all of the higher order thoughts need themselves be conscious (many mental states are not). However, according to the analysis, having a higher-order thought about a lower order thought is necessary for the lower order thought to be conscious. A slightly weaker position than Rosenthal's on this is that a conscious state corresponds to a first order thought that has the *capacity* to cause a second order thought or judgement about it (Carruthers 1996). [Another position that is close in some respects to that of Carruthers and the present position is that of (Chalmers 1996), that awareness is something that has *direct availability for behavioral 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 which can have second or higher order thoughts about its own operation, including its planning and linguistic operation, must itself be a language processor, in that it must be able to bind correctly to the symbols and syntax in the first order system. According to this explanation, the feeling of anything is the state which is present when linguistic processing that involves second or higher order thoughts is being performed.

It might be objected that this captures some of the process aspects of consciousness, what it is good for in an information processing system, but does not capture the phenomenal aspect of consciousness. I agree that there is an element of "mystery" that is invoked at this step of the argument, when I say that it feels like something for a machine with higher order thoughts to be thinking about its own first or lower order thoughts. But the return point is the following: *if a human with second order thoughts is thinking about its own first order thoughts, surely it is very difficult for us to conceive that this would NOT feel like something?* (Perhaps the higher order thoughts in thinking about the first order thoughts would need to have in doing this some sense of continuity or self, so that the first order thoughts would be related to the same system that had thought of something else a few minutes ago. But even this continuity aspect may not be a requirement for consciousness. Humans with anterograde amnesia cannot remember what they felt a few minutes ago; yet their current state does feel like something.)

It is suggested that part of the evolutionary adaptive significance of this type of higher order thought is that is enables correction of errors made in first order linguistic or in non-linguistic processing. Indeed, the ability to reflect on previous events is extremely important for learning from them, including setting up new long-term semantic structures. It was shown elsewhere (Rolls and Treves 1998)that the hippocampus may be a system for such "declarative" recall of recent memories. Its close relation to "conscious" processing in humans (Squire 1992, has classified it as a declarative memory system) may be simply that it enables the recall of recent memories, which can then be reflected upon in conscious, higher order, processing (Rolls 1996). Another part of the adaptive value of a higher order thought system may be that by thinking about its own thoughts in a given situation, it may be able to better understand the thoughts of another individual in a similar situation, and therefore predict that individual's behaviour better (cf. Barlow 1997; 1986; Humphrey 1980).

As a point of clarification, I note that according to this theory, a language processing system is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have higher order thoughts, and a first order language processor (that might be perfectly competent at language) would not be conscious, in that it could not think about its own or others' thoughts. One can perfectly well conceive of a system which obeyed the rules of language (which is the aim of much connectionist modelling), and implemented a first-order linguistic system, that would not be conscious. [Possible examples of language processing that might be performed non-consciously include computer programs implementing aspects of language, or ritualized human conversations, e.g., about the weather. These might require syntax and correctly grounded semantics, and yet be performed non-consciously. A more complex example, illustrating that syntax could be used, might be "If A does X, then B will probably do Y, and then C would be able to do Z." A first order language system could process this statement. Moreover, the first order language system could apply the rule usefully in the world, provided that the symbols in the language system (A, B, X, Y etc) are grounded (have meaning) in the world.] In line with the argument on the adaptive value of higher order thoughts and thus consciousness given above, that they are useful for correcting lower order thoughts, I now suggest that correction using higher order thoughts of lower order thoughts would have adaptive value primarily if the

lower order thoughts are sufficiently complex to benefit from correction in this way. The nature of the complexity is specific: that it should involve syntactic manipulation of symbols, probably with several steps in the chain, and that the chain of steps should be a one-off (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 higher order thoughts, that I suggest that there is a close link between this reflective consciousness and language. The hypothesis is that by thinking about lower order thoughts, the higher order thoughts can discover what may be weak links in the chain of reasoning at the lower order level, and having detected the weak link, might alter the plan, to see if this gives better success. In our example above, if it transpired that C could not do Z, how might the plan have failed? Instead of having to go through endless random changes to the plan to see if by trial and error some combination does happen to produce results, what I am suggesting is that by thinking about the previous plan, one might for example using knowledge of the situation and the probabilities that operate in it, guess that the step where the plan failed was that B did not in fact do Y. So by thinking about the plan (the first or lower order thought), one might correct the original plan, in such a way that the weak link in that chain, that "B will probably do Y", is circumvented. To draw a parallel with neural networks: there is a "*credit assignment*" problem in such 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. The suggestion is that this is the function of higher order thoughts and is why systems with higher order thoughts evolved. The suggestion I then make is that if a system were doing this type of processing (thinking about its own thoughts), it would then be very plausible that it should feel like something to be doing this. I even suggest to the reader that it is not plausible to suggest that it would not feel like anything to a system if it were doing this.

Two other points in the argument should be emphasised for clarity. One is that the system that is having syntactic thoughts about its own syntactic thoughts would have to have its symbols grounded in the real world for it to feel like something to be having higher order thoughts. The intention of this clarification is to exclude systems such as a computer running a program when there is in addition some sort of control or even overseeing program checking the operation of the first program. We would want to say that in such a situation it would feel like something to be running the higher level control program only if the first order program was symbolically performing operations on the world and receiving input about the results of those operations, and if the higher order system understood what the first order system was trying to do in the world. The issue of symbol grounding is considered further by Rolls (Rolls 1999a, section 10.4; 2000e). 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. (This suggestion may be close to the view that thoughts may be grounded by the way they function in "belief-desire psychology" which considers the functions of intentional states and attitudinal states such as beliefs, desires etc. as discussed by Fodor (1994; Fodor 1987; 1990). The notion of what constitutes a thought is itself a major issue. Animals can be said to have “expectations”; but if they are based on functionality that implements Stimulus-Response habits or stimulus-reinforcement associations, they would not I believe constitute thoughts. For the purposes of this Chapter, “thought” can be read as “human thought”, and would normally involve symbols and syntactic operations. The issue of the extent to which animals have thoughts which operate in this way remains to be fully examined and assessed, and is in principle a matter that can be resolved empirically. I am thus open-minded about the operation in animals of the type of processing described in this Chapter as higher order syntactic thought.) Indeed, it is pointed out in *The Brain and Emotion* (Rolls 1999a, pp. 252-253) that it is no accident that the shape of the multidimensional phenomenal (sensory etc) space does map so clearly onto the space defined by neuronal activity in sensory systems, for if this were not the case, reasoning about the state of affairs in the world would not map onto the world, and would not be useful. Good examples of this close correspondence are found in the taste system, in which subjective space (Schiffman and Erikson 1971) maps simply onto the multidimensional space represented by neuronal firing in primate cortical taste areas. In particular, if a

three-dimensional space reflecting the distances between the representations of different tastes provided by macaque neurons in the cortical taste areas is constructed, then the distances between the subjective ratings by humans of different tastes is very similar (Plata-Salaman et al. 1996; Smith-Swintosky et al. 1991; Yaxley et al. 1990). Similarly, the changes in human subjective ratings of the pleasantness of the taste, smell and sight of food parallel very closely the responses of neurons in the macaque orbitofrontal cortex (see *The Brain and Emotion*, Chapter 2). The representations in the first order linguistic processor that the HOSTs process include beliefs (for example "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 summarized by Rolls, (Rolls 2000e), 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 *The Brain and Emotion*, Fig. 9.3 and p. 253). When the sensory information is about the identity of the taste, the inputs to the first order linguistic system must come from the primary taste cortex, in that the identity of taste, independent of its pleasantness (in that the representation is independent of hunger) must come from the primary taste cortex. In contrast, when the information that reaches the first order linguistic system is about the pleasantness of taste, it must come from the secondary taste cortex, in that there the representation of taste depends on hunger.

The second clarification is that the plan would have to be a unique string of steps, in much the same way as a sentence can be a unique and one-off (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.

Sensory 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 being used (or are available to be used) in this type of (linguistically based) higher order thought processing, and this is what I propose.

The explanation for 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 specialised linguistic symbol-manipulation system that is part of a higher order thought system that is capable of reflecting on and correcting its lower order thoughts involved for example in the flexible planning of actions. It would require a very special machine to enable this higher-order linguistically-based thought processing, which is conscious by its nature, to occur without the sensory, emotional and motivational states (which must be taken into account by the higher order thought system) becoming felt qualia. The 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. To expand on this, qualia of low-level sensory details may not themselves involve higher order thoughts (thoughts about thoughts), and may involve thoughts about sensory processes, but nevertheless become conscious because they enter the processing system that implements HOSTs. The HOST processing system may sometimes have to reason (perform multistep planning and evaluation) about low-level sensory features of objects, e.g. to establish whether an antique has the correct coloring and surface features for it to be genuine.

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 1990,), and rule-following (implemented e.g., using habit or stimulus-response learning in the basal ganglia, Rolls 1994; Rolls and Johnstone 1992). Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals close to humans, with demonstrable syntactic manipulation of symbols (termed mentalese), and the ability to think about these linguistic processes) would be different from other animals, and would have evolved qualia. (The hypothesis described here implies that consciousness in animals could be related to the extent to which mentalese is possible in each given type of animal, and does not require natural language.)

In order for processing in a part of our brain to be able to reach consciousness, appropriate pathways must be present. Certain constraints arise here. For example, in the sensory pathways, the nature of the representation may change as it passes through a hierarchy of processing levels, and in order to be conscious of the information in the form in which it is represented in early processing stages, the early processing stages must have access to the part of the brain necessary for consciousness. An example is provided by processing in the taste system. In the primate primary taste cortex, neurons respond to taste independently of hunger, yet in the secondary taste cortex, food-related taste neurons (e.g., responding to sweet taste) only respond to food if hunger is present, and gradually stop responding to that taste during feeding to satiety (see Rolls 1997b). Now the quality of the tastant (sweet, salt etc) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety. The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex), and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it). The same argument arises for reward in general, and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is reflected in the responses of neurons to visual stimuli (see Rolls 1999a; 1995a; 1995b; 1990). It is also of interest that reward signals (e.g., the taste of food when we are hungry) are associated with subjective feelings of pleasure (see 1999a; 1997a; 1995a; 1997b; Rolls 1990). I suggest that this correspondence arises because pleasure is the subjective state that represents in the conscious system a signal that is positively reinforcing (rewarding), and that inconsistent behaviour would result if the representations did not correspond to a signal for positive reinforcement in both the conscious and the non-conscious processing systems.

Do these arguments mean that the conscious sensation of e.g., taste quality (i.e., identity and intensity) is represented or occurs in the primary taste cortex, and of the pleasantness of taste in the secondary taste cortex, and that activity in these areas is sufficient for conscious sensations (qualia) to occur? I do not suggest this at all. Instead the arguments I have put forward above suggest that we are only conscious of representations when we have high order thoughts about them. The implication then is that pathways must connect from each of the brain areas in which information is represented about which we can be conscious, to the system which has the higher order thoughts, which as I have argued above, requires language. Thus, in the example given, there must be connections to the language areas from the primary taste cortex, which need not be direct, but which must bypass the secondary taste cortex, in which the information is represented differently (see Fig. 2 and Rolls 1995a,). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have higher order thoughts about the pleasantness of the representation in the secondary taste cortex (see Fig. 2). 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. 1, which would in turn need connections to the language areas). I note that given the distributed nature of neuronal representations (see Rolls and Deco 2002), there need be no loss of information given the large numbers of neurons in any representation for information from early vs late cortical areas, provided that reasonable numbers of connections are present.

[Insert Fig. 2 near here]

One of the arguments that Rosenthal uses to support a role for higher order thoughts in consciousness is that learning a repertoire of HOTs makes the two rather similar sensory inputs available to consciousness, and consciously discriminable. An example is with wine. The blackcurrant or peppermint notes in a wine may

not be conscious at first; but after you have been trained with language, using different words to describe the different notes, then you become conscious of the different notes. So having a (higher order) thought may make some quality enter consciousness. However, words are inherently orthogonal representations, and these representations could by the top-down backprojections to earlier cortical areas could make the representations of two rather similar qualities become more different in the early cortical areas by influencing the categories being formed in competitive networks during learning, in ways described by Rolls and Treves (Rolls and Treves 1998, section 4.5) and by Rolls and Deco (Rolls and Deco 2002, section 7.4.5). If this is the mechanism, then, at least after the learning, the different notes may be brought into consciousness because of a better sensory representation, rather than by a special operation of higher order thoughts on an unchanged sensory input.

A causal role for consciousness

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 higher order thoughts play a role in correcting or influencing lower order thoughts. The sense in which the consciousness is causal is then it is suggested, that the higher order thought is causally involved in correcting the lower order thought; and that it is a property of the higher order thought system that it feels like something when it is operating. As we have seen, some behavioural responses can be elicited when there is not this type of reflective control of lower order processing, nor indeed any contribution of language (see further Rolls 2003 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. 1).

It is of interest to comment on how the evolution of a system for flexible planning might affect emotions. Consider grief which may occur when a reward is terminated and no immediate action is possible (see Rolls 1990,). 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.)

Free will

This account of consciousness also leads to a suggestion about the processing that underlies the feeling of free will. Free will would in this scheme involve the use of language to check many moves ahead on a number of possible series of actions and their outcomes, and then with this information to make a choice from the likely outcomes of different possible series of actions. (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 (the system) 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 which 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, and one of the reasons for the feeling of the *unity of consciousness*. The feeling of the unity of consciousness may also be related to the suggested involvement of syntactic processing in consciousness, which appears to be inherently serial and with a limited binding capacity. Indeed, these properties of the implementation of syntax in the brain may provide some of the underlying computational reasons why consciousness feels unitary.

In the operation of such a free will system, the uncertainties introduced by the limited information possible about the likely outcomes of series of actions, and the inability to use optimal algorithms when combining conditional probabilities, would be much more important factors than whether the brain operates deterministically or not. (The operation of brain machinery must be relatively deterministic, for it has evolved to provide reliable outputs for given inputs. What I am suggesting here though is that an interesting question about free will is not whether it reflects the operation of deterministic machinery or not, but instead is what information processing and computations are taking place when we feel that we have free will, and the confabulations that may arise when we operate using implicit computations, i.e. computations which are not available to consciousness, such as for example the reward reversal implemented by the orbitofrontal cortex, see Rolls et al. 1994.)

Self-identity and the unity of consciousness

Before leaving these thoughts, it may be worth commenting on the feeling of continuing self-identity that is characteristic of humans, and the unity of consciousness. Why might these arise? One suggestion is that if one is an organism that can think about its own long-term multi-step plans, then for those plans to be consistently and thus adaptively executed, the goals of the plans would need to remain stable, as would memories of how far one had proceeded along the execution path of each plan. If one felt each time one came to execute, perhaps on another day, the next step of a plan, that the goals were different; or if one did not remember which steps had already been taken in a multi-step plan, the plan would never be usefully executed. So, given that it does feel like something to be doing this type of planning using higher order thoughts, it would have to feel as if one were the same agent, acting towards the same goals, from day to day. Thus it is suggested that the feeling of continuing self-identity falls out of a situation in which there is an actor with consistent long-term goals, and long-term recall. If it feels like anything to be the actor, according to the suggestions of the higher order thought theory, then it should feel like the same thing from occasion to occasion to be the actor, and no special further construct is needed to account for self-identity. Humans without such a feeling of being the same person from day to day might be expected to have for example inconsistent goals from day to day, or a poor recall memory. It may be noted that the ability to recall previous steps in a plan, and bring them into the conscious, higher-order thought system, is an important prerequisite for long-term planning which involves checking each step in a multi-step process.

These are my initial thoughts on why we have consciousness, and are conscious of sensory, emotional and motivational qualia, as well as qualia associated with first-order linguistic thoughts. However, as stated above, one does not feel that there are straightforward criteria in this philosophical field of enquiry for knowing whether the suggested theory is correct; so it is likely that theories of consciousness will continue to undergo rapid development; and current theories should not be taken to have practical implications.

3. Dual routes to Action, and Decisions

The question 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. 1). 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 (e.g.,Rolls 1994). 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. Another factor which may influence the balance between control by the implicit and explicit systems is the presence of pharmacological agents such as alcohol, which may alter the balance towards control by the implicit system, may allow the implicit system to influence more the explanations made by the explicit system, and may within the explicit system alter the relative value it places on caution and restraint versus commitment to a risky action or plan.

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. An example of this might be that if a pregnant woman feels that she would like to escape a cruel mate, but is aware that she may not survive in the jungle, then it would be adaptive if the explicit system could suppress some aspects of her implicit behaviour towards her mate, so that she does not give signals that she is displeased with her situation. [In the literature on self-deception, it has been suggested that unconscious desires may not be made explicit in consciousness (or actually repressed), so as not to compromise the explicit system in what it produces (1979; see e.g., Alexander 1975; and the review by Nesse and Lloyd 1992; Trivers 1976; 1985)]. Another example might be that the explicit system might because of its long-term plans influence the implicit system to increase its response to for example a positive reinforcer. One way in which the explicit system might influence the implicit system is by setting up the conditions in which for example when a given stimulus (e.g., person) is present, positive reinforcers are given, to facilitate stimulus-reinforcement association learning by the implicit system of the person receiving the positive reinforcers. Conversely, the implicit system may influence the explicit system, for example by highlighting certain stimuli in the environment that are currently associated with reward, to guide the attention of the explicit system to such stimuli.

However, 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 (Nesse and Lloyd 1992). As Nesse and Lloyd (Nesse and Lloyd 1992) describe, analysts have come to a somewhat similar position, for they hold that intrapsychic conflicts usually seem to have two sides, with impulses on one side and inhibitions on the other. Analysts describe the source of the impulses as the *id*, and the modules that inhibit the expression of impulses, because of external and internal constraints, the *ego* and *superego* respectively (Leak and Christopher 1982; see Nesse and Lloyd 1992, p. 613). The superego can be thought of as the conscience, while the ego is the locus of executive functions that balance satisfaction of impulses with anticipated internal and external costs. A difference of the present position is that it is based on identification of dual routes to action implemented by different systems in the brain, each with its own selective advantage.

Some investigations in non-human primates on deception have been interpreted as showing that animals can plan to deceive others {see e.g., ; Trivers 1985) , that is to utilize "Machiavellian intelligence". For example, a baboon may "deliberately" mislead another animal in order to obtain a resource such as food (e.g., by screaming to summon assistance in order to have a competing animal chased from a food patch) or sex (e.g., a female baboon who very gradually moved into a position from which the dominant male could not see her grooming a subadult baboon) (see Dawkins 1993). The attraction of the Machiavellian argument is that the behaviour for which it accounts seems to imply that there is a concept of another animal's mind, and that one animal is trying occasionally to mislead another, which implies some planning. However, such observations tend by their nature to be field-based, and may have an anecdotal character, in that the previous experience of the animals in this type of behaviour, and the reinforcements obtained, are not known (Dawkins 1993). It is possible for example that some behavioural responses that appear to be Machiavellian may have been the result of previous instrumental learning in which reinforcement was obtained for particular types of response, or of observational learning, with again learning from the outcome observed. However, in any case, most examples of Machiavellian intelligence in non-human primates do not involve multiple stages of "if....then" planning requiring syntax to keep the symbols apart (but may involve learning of the type "if the dominant male sees me grooming a subadult male, I will be punished") (see Dawkins 1993). Nevertheless, the possible advantage of such Machiavellian *planning* could be one of the adaptive guiding factors in evolution which provided advantage to a multi-step, syntactic system which enables long-term planning, the best example of such a system being human language. However, another, not necessarily exclusive, advantage for the evolution of a linguistic multi-step planning system could well be not Machiavellian planning, but planning for social co-operation and advantage. Perhaps in general an "if...then" multi-step syntactic planning ability is useful primarily in evolution in social situations of the type: "if X does this, then Y does that; then I would

/ should do that, and the outcome would be ... ". It is not yet at all clear whether such planning is required in order to explain the social behaviour of social animals such as hunting dogs; or socialising monkeys (Dawkins 1993). However, in humans there is evidence that members of "primitive" hunting tribes spend hours recounting tales of recent events (perhaps who did what, when; who then did what, etc), perhaps to help learn from experience about good strategies, necessary for example when physically weak men take on large animals (see Pinker and Bloom 1992). Thus, social co-operation may be as powerful a driving force in the evolution of syntactical planning systems as Machiavellian intelligence. What is common to both is that they involve social situations. However, such a syntactic planning system would have advantages not only in social systems, for such planning may be useful in obtaining resources purely in a physical (non-social) world. An example might be planning how to cross terrain given current environmental constraints in order to reach a particular place.

The thrust of this argument 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 consideration 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. Although it has been argued by Rolls (Rolls 1999a, Chapter 3) 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 which control behaviour. For example, during female orgasm the hormone oxytocin may be released, and this may influence the implicit system to help develop positive reinforcement associations and thus attachment.

4 Discussion

Some ways in which the current theory may be different from other related theories follow. The current theory holds that it is higher order *syntactic* thoughts (HOSTs) that are closely associated with consciousness, and this may differ from Rosenthal's higher order thoughts (HOTs) theory (1990; 1993; Rosenthal 1986), 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 type of language required in the theory described here is sometimes termed “mentalese” by philosophers. 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. 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 high 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 earlier theories is that raw sensory feels are suggested to arise as a consequence of having a system that can think about its own thoughts. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution.

A property often attributed to consciousness is that it is *unitary*. The current theory would account for this by the limited syntactic capability of neuronal networks in the brain, which render it difficult to implement more than a few syntactic bindings of symbols simultaneously (McLeod et al. 1998; see Rolls and Treves 1998). This limitation makes it difficult to run several "streams of consciousness" simultaneously. In addition, given that a linguistic system can control behavioural output, several parallel streams might produce maladaptive behaviour (apparent as e.g., indecision), and might be selected against. The close relation between, and the limited capacity of, both the stream of consciousness, and auditory-verbal short term memory, may be that both implement the capacity for syntax in neural networks. Whether syntax in real neuronal networks is implemented by temporal binding (see von der Malsburg 1990) is still very much an unresolved issue (Rolls and Deco 2002; Rolls and Treves 1998). However, the hypothesis that syntactic binding is necessary for consciousness is one of the postulates of the theory I am describing (for the system I describe must be capable of correcting its own syntactic thoughts); and the fact that the binding must be implemented in neuronal networks may well place limitations on consciousness, which lead to some of its properties, such as its unitary nature. The postulate of Crick and Koch (Crick and Koch 1990) that oscillations and synchronization are necessary bases of consciousness could thus be related to the present theory if it turns out that oscillations or neuronal synchronization are the way the brain implements syntactic binding. However, the fact that oscillations and neuronal synchronization are especially evident in anaesthetized cats does not impress as strong evidence that oscillations and synchronization are critical features of consciousness, for most people would hold that anaesthetized cats are *not* conscious. The fact that oscillations and synchronization are much more difficult to demonstrate in the temporal cortical visual areas of awake behaving monkeys 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).

The current theory holds that consciousness arises by virtue of a system that can think linguistically about its own linguistic thoughts. The advantages for a system of being able to do this have been described, and this has been suggested as the reason why consciousness evolved. The evidence that consciousness arises by virtue of having a system that can perform higher order linguistic processing is however, and I think may remain, circumstantial. (Why must it feel like something when we are performing a certain type of information processing? The evidence described here suggests that it does feel like something when we are performing a certain type of information processing, but does not produce a strong reason for why it has to feel like something. It just does, when we are using this linguistic processing system capable of higher order thoughts.) The evidence, summarized above, includes the points that we think of ourselves as conscious when for example we recall earlier events, compare them with current events, and plan many steps ahead. Evidence also comes from neurological cases, from for example split brain patients (who may confabulate conscious stories about what is happening in their other, non-language, hemisphere; and from cases such as frontal lobe patients who can tell one consciously what they should be doing, but nevertheless may be doing the opposite. (The force of this type of case is that much of our behaviour may normally be produced by routes about which we cannot verbalize, and are not conscious about.) This raises the issue of the causal role of consciousness. Does consciousness cause our behaviour?[1] The view that I currently hold is that the information processing which

[1] This raises the issue of the causal relation between mental events and neurophysiological events, part of the mind-body problem. My view is that the relation between mental events and neurophysiological events is similar (apart from the problem of consciousness) to the relation between the program running in a computer and the hardware on the computer. In a sense, the program causes

is related to consciousness (activity in a linguistic system capable of higher order thoughts, and used for planning and correcting the operation of lower order linguistic systems) can play a causal role in producing our behaviour (see Fig. 1). It is, I postulate, a *property* of processing in this system (capable of higher order thoughts) that it feels like something to be performing that type of processing. It is in this sense that I suggest that consciousness can act causally to influence our behaviour: consciousness is the property that occurs when a linguistic system is thinking about its lower order thoughts. The hypothesis that it does feel like something when this processing is taking place is at least to some extent testable: humans performing this type of higher order linguistic processing, for example recalling episodic memories and comparing them with current circumstances, who denied being conscious, would *prima facie* constitute evidence against the theory. Most humans would find it very implausible though to posit that they could be thinking about their own thoughts, and reflecting on their own thoughts, without being conscious. This type of processing does appear to be for most humans to be necessarily conscious.

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

5 Conclusion

It is suggested 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 feels, arise secondary to having evolved such a higher order thought system, and that sensory and emotional processing feels like something because it would be unparsimonious for it to enter the planning, higher order thought, 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 (Rolls 2000e); reasons why the ventral visual system is more closely related to explicit than implicit processing are considered by Rolls and Deco (Rolls and Deco 2002) and by Rolls (Rolls 2003): and reasons why explicit, conscious, processing may have a higher threshold in sensory processing than implicit processing are considered by Rolls (Rolls 2003).

It may be useful to comment that this theory, while sharing much with Rosenthal's HOT theory of consciousness (1990; 1993; Rosenthal 1986), may be different in a number of respects. I take an information processing / computational approach, and try to define the computations that appear to be taking place when we are conscious. I take a brain design approach, and try to link implicit vs conscious computations with processes in different brain regions, as a way of constraining the theory, and providing useful links to medical concerns and issues. I propose that it is syntactic thoughts implementing multistep planning operations that raise a credit assignment problem, and that the *adaptive value* of higher order thoughts is to help solve this credit assignment problem. My HOST theory is based therefore on the adaptive utility of higher order thoughts for correcting lower order syntactic operations such as those involved in planning, and it is because the higher order system must be able to correct a lower order syntactic processing system that the higher order system needs to be a Higher Order Syntactic Thought (HOST) system. When I use the term "thoughts", I can thus be read as meaning "human thoughts", which imply the ability to perform syntactic processing, that is to

---

the logic gates to move to the next state. This move causes the program to move to its next state. Effectively, we are looking at different levels of what is overall the operation of a *system*, and causal explanations can usefully be understood as operating both within levels (causing one step of the program to move to the next), as well as between levels (e.g., software to hardware and vice versa). This is the solution I propose to this aspect of the mind-body (or mind-brain) problem.

implement the correct and flexible binding of symbols in relational operations. (This does not imply a need for human language, and I am open-minded about the extent to which this type of information processing may be possible in animals.) When Rosenthal uses the term “thought” he may be using it in a wider sense, and I hope that this is an issue dealt with by Rosenthal (Rosenthal 2004).

Acknowledgements. I am very grateful to David Rosenthal (City University of New York), Marian Dawkins (Oxford University), and Martin Davies (Australian National University) for many fascinating and inspiring discussions on the topics considered here.

Figure Legends

Fig. 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 (verbalizable) decisions involving multi-step syntactic planning to be implemented. (After Rolls 1999a, Fig 9.4) (emdualroutes.eps)

Fig. 2. Schematic illustration indicating that early cortical stages in information processing may need access to language areas that bypass subsequent levels in the hierarchy, so that consciousness of what is represented in early cortical stages, and which may not be represented in later cortical stages, can occur. Higher-order syntactic thoughts could be implemented in the language cortex itself, which for the theory described here needs to implement syntactic processing on symbols, but not natural language. (Natural language implies a universal grammar.) Backprojections, a notable feature of cortical connectivity, with many probable functions including recall, attention, and influencing the categories formed in earlier cortical areas during learning (see Rolls and Deco 2002), probably reciprocate all the connections shown.
(After Rolls 1999a, Fig. 9.3).

Fig. 1

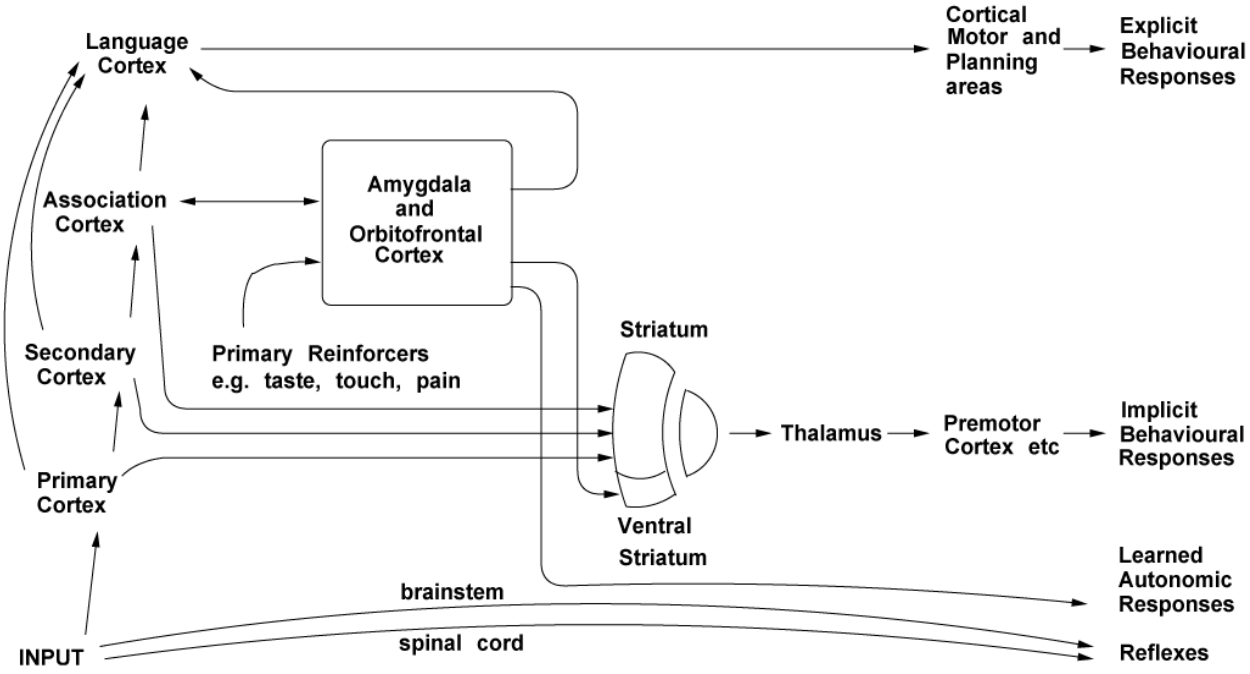
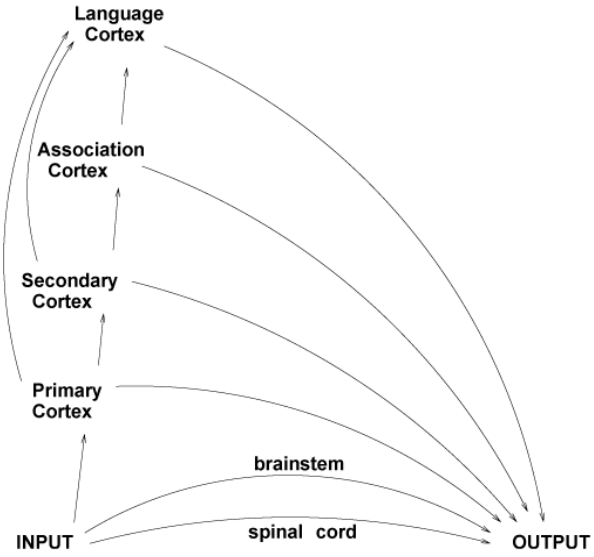

Language Cortex
Cortical Motor and Planning areas
Explicit Behavioural Responses
Association Cortex
Amygdala and Orbitofrontal Cortex
Striatum
Secondary Cortex
Primary Reinforcers e.g. taste, touch, pain
Thalamus
Premotor Cortex etc
Implicit Behavioural Responses
Primary Cortex
Ventral Striatum
Learned Autonomic Responses
brainstem
spinal cord
Reflexes
INPUT

Fig. 2.


Language
Cortex
Association
Cortex
Secondary
Cortex
Primary
Cortex
brainstem
INPUT
spinal cord
OUTPUT

Alexander RD. The search for a general theory of behavior. Behav Sci 20: 77-100, 1975.

Alexander RD. Darwinism and Human Affairs. Seattle: University of Washington Press, 1979.

Allport A. What concept of consciousness? In: Conscousness in Contemporary Science, edited by Marcel AJ and Bisiach E. Oxford: Oxford University Press, 1988, p. 159-182.

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

Barlow HB. Single neurons, communal goals, and consciousness. In: Cognition, Computation, and Consciousness, edited by Ito M, Miyashita Y and Rolls ET. Oxford: Oxford University Press, 1997, p. 121-136.

Block N. On a confusion about a function of consciousness. Behav Brain Sci 18: 227-247, 1995.

Booth DA. Food-conditioned eating preferences and aversions with interoceptive elements: learned appetites and satieties. Ann NY Acad Sci 443: 22-37, 1985.

Carruthers P. Language, Thought and Consciosness. Cambrudge: Cambridge University Press, 1996.

Chalmers DJ. The Conscious Mind. Oxford: Oxford University Press, 1996.

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

Crick FHC and Koch C. Towards a neurobiological theory of consciousness. Sem Neurosci 2: 263-275, 1990.

Darwin C. The Expression of the Emotions in Man and Animals. Chicago: University of Chicago Press, 1872.

Dawkins MS. Through Our Eyes Only? The Search for Animal Consciousness. Oxford: Freeman, 1993.

Dennett DC. Consciousness Explained. London: Penguin, 1991.

Ekman P. Emotion in the Human Face. Cambridge: Cambridge University Press, 1982.

Ekman P. Facial expression and emotion. Am Psych 48: 384-392, 1993.

Fodor JA. Psychosemantics: the problem of meaning in the philosophy of mind. Cambridge, MA: MIT Press, 1987.

Fodor JA. A Theory of Content and other Essays. Cambridge, MA: MIT Press, 1990.

Fodor JA. The Elm and the Expert: mentalese and its semantics. Cambridge, MA: MIT Press, 1994.

Gazzaniga MS and LeDoux J. The Integrated Mind. New York: Plenum, 1978.

Gazzaniga MS. Brain modularity: towards a phlosophy of conscious experience. In: Consciousness in Contempoary Science, edited by Marcel AJ and Bisiach E. Oxford: Oxford University Press, 1988, p. 218-238.

Gazzaniga MS. Consciousness and the cerebral hemispheres. In: The Cognitive Neurosciences, edited by Gazzaniga MS. Cambridge, Mass.: MIT Press, 1995, p. 1392-1400.

Goldman-Rakic PS. The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. Phil Trans R Soc Lond B 351: 1445-1453, 1996.

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

Hornak J, O'Doherty, J., Bramham, J., Rolls, E.T., Morris, R.G., Bullock, P.R. and Polkey, C.E. Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. J Cogn Neurosci in press, 2003b.

Humphrey NK. Nature's psychologists. In: Consciousness and the Physical World, edited by Josephson BD and Ramachandran VS. Oxford: Pergamon, 1980, p. 57-80.

Humphrey NK. The Inner Eye. London: Faber, 1986.

Krebs JR and Kacelnik A. Decision Making. In: Behavioural Ecology (3rd ed.), edited by Krebs JR and Davies NB. Oxford: Blackwell, 1991, p. 105-136.

Leak GK and Christopher SB. Freudian psychoanalysis and sociobiology: a synthesis. Am Psych 37: 313-322, 1982.

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

Nesse RM and Lloyd AT. The evolution of psychodynamic mechanisms. In: The Adapted Mind, edited by Barkow JH, Cosmides L and Tooby J. New York: Oxford University Press, 1992, p. 601-624.

Oatley K and Jenkins JM. Understanding Emotions. Oxford: Backwell, 1996.

Petrides M. Specialized systems for the processing of mnemonic information within the primate frontal cortex. Philosophical Transactions of the Royal Society B 351: 1455-1462, 1996.

Pinker S and Bloom P. Natural language and natural selection. In: The Adapted Mind, edited by Barkow JH, Cosmides L and Tooby J. New York: Oxford University Press, 1992, p. 451-493.

Plata-Salaman CR, Smith-Swintosky VL, and and Scott TR. Gustatory neural coding in the monkey cortex: mixtures. J Neurophysiol: 2369-2379, 1996.

Rolls ET. A theory of emotion, and its application to understanding the neural basis of emotion. Cog

Emot 4: 161-190, 1990.

Rolls ET. Neurophysiology and cognitive functions of the striatum. Rev Neurol (Paris) 150: 648-660, 1994.

Rolls ET. Central taste anatomy and neurophysiology. In: Handbook of Olfaction and Gustation, edited by R.L.Doty. New York.: Dekker:, 1995a, p. Ch. 24, pp. 549-573.

Rolls ET. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In: The Cognitive Neurosciences, edited by Gazzaniga MS. Cambridge, Mass.: MIT Press, 1995b, p. 1091-1106.

Rolls ET. A theory of hippocampal function in memory. Hippocampus 6: 601-620, 1996.

Rolls ET. Brain mechanisms of vision, memory, and consciousness. In: Cognition, Computation, and Consciousness, edited by Ito M, Miyashita Y and Rolls ET. Oxford: Oxford University Press, 1997a, p. 81-120.

Rolls ET. Taste and olfactory processing in the brain and its relation to the control of eating. Critical Reviews in Neurobiology 11: 263-287, 1997b.

Rolls ET. The Brain and Emotion. Oxford: Oxford University Press, 1999a.

Rolls ET. The functions of the orbitofrontal cortex. Neurocase 5: 301-312, 1999b.

Rolls ET. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron 27: 205-218, 2000a.

Rolls ET. Hippocampo-cortical and cortico-cortical backprojections. Hippocampus 10: 380-388, 2000b.

Rolls ET. Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In: The Amygdala: A Functional Analysis (Second Edition ed.), edited by Aggleton JP. Oxford: Oxford University Press, 2000c.

Rolls ET. The orbitofrontal cortex and reward. Cereb Cortex 10: 284-294, 2000d.

Rolls ET. Précis of The Brain and Emotion. Behav Brain Sci 23: 177-233, 2000e.

Rolls ET. The functions of the orbitofrontal cortex. In: Principles of Frontal Lobe Function, edited by Stuss DT, Knight, R.T. New York: Oxford University Press, 2002, p. Chap.23, 354-375.

Rolls ET. Consciousness absent and present: a neurophysiological exploration. Progress in Brain Research, in press., 2003.

Rolls ET and Johnstone S. Neurophysiological analysis of striatal function. In: Neuropsychological Disorders Associated with Subcortical Lesions, edited by Vallar G and Wallesch CW. Oxford: Oxford University Press, 1992, p. 61-97.

Rolls ET, Hornak J, Wade D, and McGrath J. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J Neurol Neurosurg and Psychiat 57: 1518-1524, 1994.

Rolls ET and Treves A. Neural Networks and Brain Function. Oxford, UK: Oxford University Press, 1998.

Rolls ET, Treves A, Robertson RG, Georges-Francois P, and Panzeri S. Information about spatial view in an ensemble of primate hippocampal cells. J Neurophysiol 79: 1797-1813, 1998.

Rolls ET and Stringer SM. A model of the interaction between mood and memory. Network: Computation in Neural Systems 12: 111-129, 2001.

Rolls ET and Deco G. Computational Neuroscience of Vision. Oxford.: Oxford University Press:, 2002.

Rosenthal DM. Two concepts of consciousness. Phil Stud 49: 329-359, 1986.

Rosenthal DM. A theory of consciousness. In: ZIF. Bielefeld, Germany: Zentrum für Interdisziplinaire Forschung, 1990.

Rosenthal DM. Thinking that one thinks. In: Consciousness, edited by Davies M and Humphreys GW. Oxford: Blackwell, 1993, p. 197-223.

Rosenthal DM. Varieties of Higher-Order Theory. In: Higher Order Theories of Consciousness, edited by Gennaro RJ. Amsterdam: John Benjamins, 2004.

Schiffman SS and Erikson RP. A psychophysical model for gustatory quality. Physiol Behav 7: 617-633, 1971.

Shallice T and Burgess P. The domain of supervisory processes and temporal organization of behaviour. Philosophical Transactions of the Royal Society B 351: 1405-1411, 1996.

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

Smith-Swintosky VL, Plata-Salaman CR, and Scott TR. Gustatory neural encoding in the monkey cortex: stimulus quality. J Neurophysiol 66: 1156-1165, 1991.

Squire LR. Memory and the hippocampus: A synthesis from findings with rats, monkeys and humans. Psych Rev 99: 195-231, 1992.

Treves A and Rolls ET. A computational analysis of the role of the hippocampus in memory. Hippocampus 4: 374-391, 1994.

Trivers RL. Foreword. In: The Selfish Gene, edited by Dawkins R. Oxford: Oxford Univerisity Press, 1976.

Trivers RL. Social Evolution. California: Benjamin Cummings, 1985.

von der Malsburg C. A neural architecture for the representation of scenes. In: Brain Organisation and Memory: Cells, Systems and Circuits, edited by McGaugh JL, Weinberger NM and Lynch G. New York, NY: Oxford University Press, 1990, p. 356-372.

Yaxley S, T. RE, and Sienkiewicz ZJ. Gustatory responses of single neurons in the insula of the macaque monkey. J Neurophysiol 63: 689-700, 1990.