Chapter 4

Emotion, higher-order syntactic thoughts, and consciousness

Edmund T. Rolls

4.1 Introduction

LeDoux (1996), in line with Johnson-Laird (1988) and Baars (1988) (see also Dehaene and Naccache 2001; Dehaene *et al*. 2006), emphasizes the role of working memory in consciousness, where he views working memory as a limited-capacity serial processor that creates and manipulates symbolic representations (p 280). He thus holds that much emotional processing is unconscious, and that when it becomes conscious it is because emotional information is entered into a working memory system. However, LeDoux (1996) concedes that consciousness, especially its phenomenal or subjective nature, is not completely explained by the computational processes that underlie working memory (p. 281).

LeDoux (2007) notes that the term working memory can refer to a number of different processes. In top-down attentional processing, a short-term memory is needed to hold online the subject of the attention, for example the position in space at which an object must be identified, or the object that must be found (Rolls and Deco 2002; Rolls 2008). There is much evidence that this short-term memory is implemented in the prefrontal cortex by an attractor network implemented by associatively modifiable recurrent collateral connections between cortical pyramidal cells, which keep the population active during the attentional task (Rolls 2008). This short-term memory then biases posterior perceptual and memory networks in the temporal and parietal lobes in a biased competition process (Miller and Cohen 2001; Rolls and Deco 2002; Deco and Rolls 2005a, 2005b; Rolls 2008). The operation of this type of short-term memory acting using biased competition to implement top-down attention does not appear to be central to consciousness, for as LeDoux (2007) agrees, prefrontal cortex lesions that have major effects on attention and short-term memory do not impair subjective feelings of consciousness (Rolls 2008). Thus in the absence of any top-down modulation from a short-term memory to implement top-down attention by biased competition, consciousness in a landscape without top-down biasing can still occur. In this scenario, there is no top-down 'attentional spotlight' anywhere. The same evidence suggests that top-down attention itself is not a fundamental process that is necessary for consciousness, though of course if attention is directed towards particular perceptual events, this will increase the gain of the perceptual processing (Deco and Rolls 2005a, 2005b; Rolls 2008), making the attended phenomena stronger.

In this chapter, I compare this approach with another approach to emotion and consciousness (Rolls 2005a). Emotion is considered first, and this then sets a framework for approaching the relation between affect and consciousness. I describe multiple routes to action, some of which involve implicit (unconscious) emotional processing, and one of which involves multiple-step planning and leads to a higher-order syntactic theory of consciousness. Then this theory of emotion and consciousness is compared with that of LeDoux (1996, 2007).

4.2 **Emotions as states**

Emotions can usefully be defined (operationally) as states elicited by rewards and punishers which have particular functions (Rolls 1999a, 2005a). The functions are defined below, and include working to obtain or avoid the rewards and punishers. 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. A diagram summarizing some of the emotions associated with the delivery of reward or punishment or a stimulus associated with them, or with the omission of a reward or punishment, is shown in Fig 4.1.

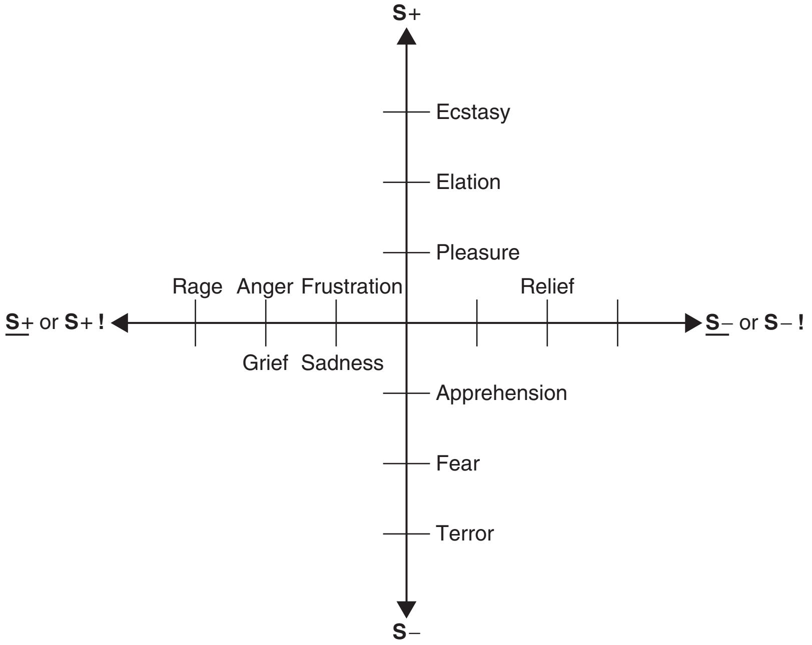

Fig. 4.1. Some of the emotions associated with different reinforcement contingencies are indicated. Intensity increases away from the centre of the diagram, on a continuous scale. The classification scheme created by the different reinforcement contingencies consists of (1) the presentation of a positive reinforcer (S+), (2) the presentation of a negative reinforcer (S–), (3) the omission of a positive reinforcer (*S+*) or the termination of a positive reinforcer (S+!), and (4) the omission of a negative reinforcer (S–) or the termination of a negative reinforcer (S–!).

Before accepting this approach, we should consider whether there are any exceptions to the proposed rule. Are any emotions caused by stimuli, events, or remembered events that are not rewarding or punishing? Do any rewarding or punishing stimuli not cause emotions? We will consider these questions in more detail below. The point is that if there are no major exceptions, or if any exceptions can be clearly encapsulated, then we may have a good working definition at least of what causes emotions. Moreover, it is worth pointing out that many approaches to or theories of emotion (Strongman 1996) have in common that part of the process involves 'appraisal' (Frijda 1986; Lazarus 1991; Oatley and Jenkins 1996). In all these theories the concept of appraisal presumably involves assessing whether something is rewarding or punishing. The description in terms of reward or punishment adopted here seems more tightly and operationally specified. I next consider a slightly more formal definition than rewards or punishments, in which the concept of reinforcers is introduced, and show how there has been a considerable history in the development of ideas along this line.

The proposal that emotions can be usefully seen as states produced by instrumental reinforcing stimuli follows earlier work by Millenson (1967), Weiskrantz (1968), Gray (1975, 1987), and Rolls (1986a, 1986b, 1990, 1999a, 2000a, 2005a). (Instrumental reinforcers are stimuli which, if their occurrence, termination, or omission is made contingent upon the making of a response, alter the probability of the future emission of that response.) Some stimuli are unlearned reinforcers (e.g. the taste of food if the animal is hungry, or pain); while others may become reinforcing by learning, because of their association with such primary reinforcers, thereby becoming 'secondary reinforcers'. This type of learning may thus be called 'stimulus–reinforcement association', and occurs via a process like classical conditioning. If a reinforcer increases the probability of emission of a response on which it is contingent, it is said to be a 'positive reinforcer' or 'reward'; if it decreases the probability of such a response it is a 'negative reinforcer' or 'punisher'. For example, fear is an emotional state which might be produced by a sound (the conditioned stimulus) that has previously been associated with an electric shock (the primary reinforcer).

The converse reinforcement contingencies produce the opposite effects on behaviour. The omission or termination of a positive reinforcer ('extinction' and 'time out' respectively, sometimes described as 'punishing') decreases the probability of responses. Responses followed by the omission or termination of a negative reinforcer increase in probability, this pair of negative reinforcement operations being termed 'active avoidance' and 'escape' respectively (Rolls 2005a).

This foundation has been developed (see Rolls 1986a, 1986b, 1990, 1999a, 2000a, 2005a) to show how a very wide range of emotions can be accounted for, as a result of the operation of a number of factors, including the following:

1. The *reinforcement contingency* (e.g. whether reward or punishment is given, or withheld) (see Fig. 4.1).
2. The *intensity* of the reinforcer (see Fig. 4.1).
3. Any environmental stimulus might have a *number of different reinforcement associations*. (For example, a stimulus might be associated both with the presentation of a reward and of a punisher, allowing states such as conflict and guilt to arise.)
4. Emotions elicited by stimuli associated with *different primary reinforcers* will be different.

5 Emotions elicited by *different secondary reinforcing stimuli* will be different from each other (even if the primary reinforcer is similar).

6 The emotion elicited can depend on whether an *active or passive behavioural response* is possible. (For example, if an active behavioural response can occur to the omission of a positive reinforcer, then anger might be produced, but if only passive behaviour is possible, then sadness, depression or grief might occur.)

By combining these six factors, it is possible to account for a very wide range of emotions (for elaboration see Rolls 2005a). It is also worth noting that emotions can be produced just as much by the recall of reinforcing events as by external reinforcing stimuli; that cognitive processing (whether conscious or not) is important in many emotions, for very complex cognitive processing may be required to determine whether or not environmental events are reinforcing. Indeed, emotions normally consist of cognitive processing which analyses the stimulus, and then determines its reinforcing valence; and then an elicited mood change if the valence is positive or negative. In that an emotion is produced by a stimulus, philosophers say that emotions have an object in the world, and that emotional states are intentional, in that they are about something. We note that a mood or affective state may occur in the absence of an external stimulus, as in some types of depression, but that normally the mood or affective state is produced by an external stimulus, with the whole process of stimulus representation, evaluation in terms of reward or punishment, and the resulting mood or affect being referred to as emotion.

It is worth raising the issue that some philosophers categorize fear in the example as an emotion, but not pain. The distinction they make may be that primary (unlearned or innate) reinforcers (for example pain) do not produce emotions, whereas secondary reinforcers (stimuli associated by stimulus–reinforcement learning with primary reinforcers) do. (An example is fear, which is a state produced by a secondary reinforcing stimulus such as the sight of an image associated by learning with a primary reinforcer such as pain.) They describe the pain as a sensation. But neutral stimuli (such as a table) can produce sensations when touched. Thus whether a stimulus produces a sensation or not does not seem to be a useful distinction that has anything to do with affective or emotional states. It accordingly seems to be much more useful to categorize stimuli according to whether they are reinforcing (in which case they produce emotions or affective states, produced by both primary and secondary reinforcers), or are not reinforcing (in which case they do not produce emotions of affective states such as pleasantness or unpleasantness). 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 affective states and emotion. These points are considered in more detail by Rolls (2005a), who provides many examples of primary versus secondary reinforcers, all of which elicit affective states.

4.3 The functions of emotion

The functions of emotion also provide insight into the nature of emotion. These functions, described more fully elsewhere (Rolls 1990, 1999a, 2005a), 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/epinephrine). These prepare the body for action.
2. *Flexibility of behavioural 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 behaviour 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 is the solution which natural selection has found for how genes can influence behaviour to promote their fitness (as measured by reproductive success), and for how the brain could interface sensory systems to action systems, and is an important part of Rolls' theory of emotion (1990, 1999a, 2005a).

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 2005a, 2008). 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 behaviour (see Fig. 4.2). This selector is poorly understood, but it might include a process of competition between all the

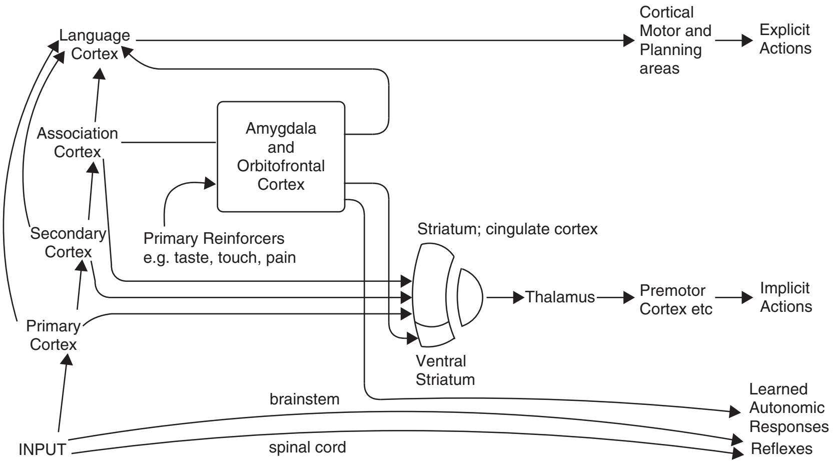

Fig. 4.2. Dual routes to the initiation of action in response to rewarding and punishing stimuli. The inputs from different sensory systems to brain structures such as the orbitofrontal cortex and amygdala allow these brain structures to evaluate the reward- or punishment-related value of incoming stimuli, or of remembered stimuli. The different sensory inputs enable evaluations within the orbitofrontal cortex and amygdala based mainly on the primary (unlearned) reinforcement value for taste, touch, and olfactory stimuli, and on the secondary (learned) reinforcement value for visual and auditory stimuli. In the case of vision, the 'association cortex' which outputs representations of objects to the amygdala and orbitofrontal cortex is the inferior temporal visual cortex. One route for the outputs from these evaluative brain structures is via projections directly to structures such as the basal ganglia (including the striatum and ventral striatum) to enable implicit, direct behavioural responses based on the reward or punishment-related evaluation of the stimuli to be made. The second route is via the language systems of the brain, which allow explicit decisions involving multi-step syntactic planning to be implemented.

competing calls on output, and might involve the anterior cingulate cortex and basal ganglia in the brain (Rolls 2005a, 2008) (see Fig. 4.2).

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 behaviour of other animals. This aspect of emotion was emphasized by Darwin (1872), and has been studied more recently by Ekman (1982, 1993). He reviews evidence that humans can categorize facial expressions into the categories happy, sad, fearful, angry, surprised, and disgusted, and that this categorization may operate similarly in different cultures. As shown elsewhere, there are neural systems in the orbitofrontal cortex, amygdala and overlying temporal cortical visual areas which are specialized for the face-related aspects of this processing (Rolls 2005a, 2007b; Rolls *et al*. 2006).

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 back-projections from parts of the brain involved in emotion such as the orbitofrontal cortex and amygdala implement this is described in *Emotion Explained* (Rolls 2005a).

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 (Rolls 2005a, 2008). 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 (Rolls 2008).

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 behaviour*, to help achieve a goal or goals.

9 Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala back-projections 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 2008).

4.4 Reward, punishment, and emotion in brain design: an evolutionary approach

The theory of the functions of emotion is further developed in *Emotion Explained* (Rolls 2005a). Some of the points made help to elaborate greatly on 3.2 above. Rolls (2005a) considers the fundamental question of why we and other animals are built to use rewards and punishments to guide or determine our behaviour. Why are we built to have emotions, as well as motivational states? Is there any reasonable alternative around which evolution could have built complex animals?

Rolls argues that a role of natural selection is to guide animals to build sensory systems that will respond to dimensions of stimuli in the natural environment along which actions can lead to better ability to pass genes on to the next generation, that is to increased fitness. The animals must be built by such natural selection to make actions that will enable them to obtain more rewards, that is to work to obtain stimuli that will increase their fitness. Correspondingly, animals must be built to make responses that will enable them to escape from, or learn to avoid, stimuli that will reduce their fitness. There are likely to be many dimensions of environmental stimuli along which responses can alter fitness. Each of these dimensions may be a separate reward–punishment dimension. An example of one of these dimensions might be food reward. It increases fitness to be able to sense nutrient need, to have sensors that respond to the taste of food, and to perform behavioural responses to obtain such reward stimuli when in that need or motivational state. Similarly, another dimension is water reward, in which the taste of water becomes rewarding when there is body fluid depletion (see Chapter 6 of *Emotion Explained*).

With many reward–punishment dimensions for which actions may be performed (see Table 2.1 of *Emotion Explained* for a non-exhaustive list!), a selection mechanism for actions performed is needed. In this sense, rewards and punishers provide a *common currency* for inputs to response selection mechanisms. Evolution must set the magnitudes of each of the different reward systems so that each will be chosen for action in such a way as to maximize overall fitness. Food reward must be chosen as the aim for action if a nutrient is depleted; but water reward as a target for action must be selected if current water depletion poses a greater threat to fitness than the current food depletion. This indicates that each reward must be carefully calibrated by evolution to have the right value in the common currency for the competitive selection process. Other types of behaviour, such as sexual behaviour, must be selected sometimes, but probably less frequently, in order to maximize fitness (as measured by gene transmission into the next generation). Many processes contribute to increasing the chances that a wide set of different environmental rewards will be chosen over a period of time, including not only need-related satiety mechanisms which decrease the rewards within a dimension, but also sensory-specific satiety mechanisms, which facilitate switching to another reward stimulus (sometimes within and sometimes outside the same main dimension), and attraction to novel stimuli. Finding novel stimuli rewarding is one way that organisms are encouraged to explore the multidimensional space in which their genes are operating.

The implication of this comparison is that operation by animals using reward–punishment systems tuned to dimensions of the environment that increase fitness provides a mode of operation that can work in organisms that evolve by natural selection. It is clearly a natural outcome of Darwinian evolution to operate using reward–punishment systems tuned to fitness-related dimensions of the environment, if arbitrary responses are to be made by the animals, rather than just preprogrammed movements such as tropisms and taxes. This view of brain design in terms of reward–punishment systems built by genes that gain their adaptive value by being tuned to a goal for action offers, I believe, a deep insight into how natural selection has shaped many brain systems, and is a fascinating outcome of Darwinian thought.

Part of the value of the approach to emotions described here, that they are states elicited by reinforcers that implement the functions described above, is that it provides a firm foundation for which systems to analyse in the brain, i.e. those brain systems involved in responding to reinforcers to implement these functions (*pace* a statement by LeDoux (2007) who commented on just a part of this definition). This approach has been made use of extensively in *The Brain and Emotion* (Rolls 1999a), in *Emotion Explained* (Rolls 2005a), and elsewhere (Rolls 2000b, 2004b, 2006b, 2007c; O'Doherty *et al*. 2001b; Kringelbach and Rolls 2003;, 2004; de Araujo *et al*. 2005) where fuller details of the neuroscience are provided than can be included here.

4.5 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 (2005a), 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 2008), 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 (2005a), but what about the subjective aspects of emotion, what about the feeling of 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 2004b, 2005a, 2006b). 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 (2005a) (and more formally in terms of neural networks by Rolls 2008 and Rolls and Deco 2002), and yet would probably not wish to ascribe feelings of pleasantness to the system we have simulated on the computer.

What is it about neural processing that makes it feel like something when some types of information processing are taking place? It is clearly not a general property of processing in neural networks, for there is much processing, for example that 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.

4.6 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 back-projection systems to most of the parts of the cerebral cortex from which they receive inputs (Rolls and Treves 1998; Rolls 2005a). 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 and Zola 1996) do have major back-projection systems to the high parts of the cerebral cortex from which they receive forward projections (Treves and Rolls 1994; Rolls 2008). 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, 2005a). The evidence is consistent with the possibility that different cortical areas, each specialized for a different type of computation, have their outputs directed to the basal ganglia, which then select the strongest input, and map this into action (via outputs directed for example to the premotor cortex) (Rolls 2005a). 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. 4.2.

Consistent with the hypothesis of multiple routes to action, only some of which utilize language, is the evidence that split-brain patients may not be aware of actions being performed by the ‘non-dominant’ hemisphere (Gazzaniga and LeDoux 1978; Gazzaniga 1988, 1995; Cooney and Gazzaniga 2003).

Also consistent with multiple including non-verbal routes to action, patients with focal brain damage, for example to the prefrontal cortex, may perform actions, yet comment verbally that they should not be performing those actions (Rolls *et al.* 1994a, 1999b, 2005a; Hornak *et al.* 2003, 2004;). The actions which appear to be performed implicitly, with surprise expressed later by the explicit system, include making behavioural responses to a no-longer-rewarded visual stimulus in a visual discrimination reversal (Rolls *et al.* 1994a; Hornak *et al.* 2004). In both these types of patient, confabulation may occur, in that a verbal account of why the action was performed may be given, and this may not be related at all to the environmental event which actually triggered the action (Gazzaniga and LeDoux 1978; Gazzaniga 1988, 1995; Rolls *et al.* 1994a, 2005a; LeDoux 2007).

Also consistent with multiple (including non-verbal) routes to action is the evidence that in backward masking at short time delays between the stimulus and the mask, neurons in the inferior temporal visual cortex respond selectively to different faces, and humans guess which face was presented 50% better than chance, yet report having not seen the face consciously (Rolls 2003).

It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region, such as those involved in some types of rewarded behaviour, the language system may subsequently elaborate a coherent account of why that action was performed (i.e. confabulate). This would be consistent with a general view of brain evolution in which as areas of the cortex evolve, they are laid on top of existing circuitry connecting inputs to outputs, and in which each level in this hierarchy of separate input–output pathways may control behaviour according to the specialized function it can perform (see schematic in Fig. 4.2). (It is of interest that mathematicians may get a hunch that something is correct, yet not be able to verbalize why. They may then resort to formal, more serial and language-like, theorems to prove the case, and these seem to require conscious processing. This is a further indication of a close association between linguistic processing, and consciousness. The linguistic processing need not, as in reading, involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements, and to perform syntactic operations on the semantic representations of states in the world, would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g. I think that he thinks that …), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others.[1] This capability for higher-order thoughts would also enable reflection on past events, which would also be useful in planning. In contrast, non-linguistic behaviour would be driven by learned reinforcement associations, learned rules, etc., but not by flexible planning for many steps ahead involving a model of the world including others' behaviour. (For an earlier view which is close to this part of the argument see Humphrey 1980.) The examples of behaviour by 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.

It is next suggested that this arbitrary symbol-manipulation using important aspects of language processing and used for planning but not in initiating all types of behaviour is close to what consciousness is about. In particular, consciousness may *be* the state which arises in a system that can think about (or reflect on) its own (or other peoples') thoughts, that is in a system capable of second or higher-order thoughts (Rosenthal 1986, 1990, 1993, 2004, 2005; Dennett 1991; Rolls 1995, 1997a, 1997b, 1999a, 2004a, 2005a, 2007a; Carruthers 1996; Gennaro 2004). On this account, a mental state is non-introspectively (i.e. non-reflectively) conscious if one has a roughly simultaneous thought that one is in that mental state. Following from this, introspective consciousness (or reflexive consciousness, or self-consciousness) is the attentive, deliberately focused consciousness of one's mental states. It is noted that not all of the 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 (and mine) on this is that a conscious

[1] Second-order thoughts are thoughts about thoughts. Higher-order thoughts refer to second-order, third-order, etc. thoughts about thoughts.

state corresponds to a first-order thought that has the *capacity* to cause a second-order thought or judgement about it (Carruthers 1996). (Another position which is close in some respects to that of Carruthers and the present position is that of Chalmers 1996, that awareness is something that has *direct availability for behavioural control*, which amounts effectively for him in humans to saying that consciousness is what we can report (verbally) about.) This analysis is consistent with the points made above that the brain systems that are required for consciousness and language are similar. In particular, a system 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 (discussed further below) is the following: *if a human with second-order thoughts is thinking about its own first-order thoughts, surely it is very difficult for us to conceive that this would NOT feel like something?* (Perhaps the 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 above that the hippocampus may be a system for such 'declarative' recall of recent memories. Its close relation to 'conscious' processing in humans (Squire and Zola 1996 have classified it as a declarative memory system) may be simply that it enables the recall of recent memories, which can then be reflected upon in conscious, higher-order, processing (Rolls and Kesner 2006; Rolls 2008). 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. Humphrey 1980, 1986; Barlow 1997).

As a point of clarification, I note that according to this theory, a language processing system (let alone a working memory, LeDoux 2007) is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have higher-order thoughts, and a firstorder 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.

I draw a parallel with neural networks: there is a '*credit assignment*' problem in such multi-step syntactic plans, in that if the whole plan fails, how does the system assign credit or blame to particular steps of the plan. (In multilayer neural networks, the credit assignment problem is that if errors are being specified at the output layer, the problem arises about how to propagate back the error to earlier, hidden, layers of the network to assign credit or blame to individual synaptic connection; see Rolls and Deco 2002, Rumelhart *et al.* 1986 and Rolls 2008.) The suggestion is that this is the function of higher-order thoughts and is why systems with higher-order thoughts evolved. The suggestion I then make is that if a system were doing this type of processing (thinking about its own thoughts), it would then be very plausible that it should feel like something to be doing this. I even suggest to the reader that it is not plausible to suggest that it would not feel like anything to a system if it were doing this.

Two other points in the argument should be emphasized for clarity. One is that the system that is having syntactic thoughts about its own syntactic thoughts (higher-order syntactic thoughts or HOSTs) would have to have its symbols grounded in the real world for it to feel like something to be having 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 (2005a). The symbols (or symbolic representations) are symbols in the sense that they can take part in syntactic processing. The symbolic representations are grounded in the world in that they refer to events in the world. The symbolic representations must have a great deal of information about what is referred to in the world, including the quality and intensity of sensory events, emotional states, etc.

The need for this is that the reasoning in the symbolic system must be about stimuli, events, and states, and remembered stimuli, events and states, and for the reasoning to be correct, all the information that can affect the reasoning must be represented in the symbolic system, including for example just how light or strong the touch was, etc. Indeed, it is pointed out in *Emotion Explained* that it is no accident that the shape of the multidimensional phenomenal (sensory, etc.) space does map so clearly onto the space defined by neuronal activity in sensory systems, for if this were not the case, reasoning about the state of affairs in the world would not map onto the world, and would not be useful. Good examples of this close correspondence are found in the taste system, in which subjective space maps simply onto the multidimensional space represented by neuronal firing in primate cortical taste areas. In particular, if a three-dimensional space reflecting the distances between the representations of different tastes provided by macaque neurons in the cortical taste areas is constructed, then the distances between the subjective ratings by humans of different tastes is very similar (Yaxley *et al.* 1990; Smith-Swintosky *et al.* 1991; Kadohisa *et al.* 2005). Similarly, the changes in human subjective ratings of the pleasantness of the taste, smell, and sight of food parallel very closely the responses of neurons in the macaque orbitofrontal cortex (see *Emotion Explained*).

The representations in the first-order linguistic processor that the HOSTs process include beliefs (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 argued by Rolls (1999a, 2005a), representations of sensory processes and emotional states must be processed by the first-order linguistic system, and HOSTs may be about these representations of sensory processes and emotional states capable of taking part in the syntactic operations of the first-order linguistic processor. Such sensory and emotional information may reach the first-order linguistic system from many parts of the brain, including those such as the orbitofrontal cortex and amygdala implicated in emotional states (see Fig. 4.2 and *Emotion Explained*, Fig. 10.3). When the sensory information is about the identity of the taste, the inputs to the first order linguistic system must come from the primary taste cortex, in that the identity of taste, 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.

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 specialized linguistic symbol-manipulation system that is part of a higher-order thought system that is capable of reflecting on and correcting its lower-order thoughts involved for example in the flexible planning of actions. It would require a very special machine to 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.

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 1999a, 2004b, 2005a), and rule-following (implemented e.g. using habit or stimulus–response learning in the basal ganglia (Rolls 2005a).

Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals close to humans, with demonstrable syntactic manipulation of symbols, and the ability to think about these linguistic processes) would be different from other animals, and would have evolved qualia.

In order for processing in a part of our brain to be able to reach consciousness, appropriate pathways must be present. Certain constraints arise here. For example, in the sensory pathways, the nature of the representation may change as it passes through a hierarchy of processing levels, and in order to be conscious of the information in the form in which it is represented in early processing stages, the early processing stages must have access to the part of the brain necessary for consciousness (see Fig. 4.2). An example is provided by processing in the taste system. In the primate primary taste cortex, neurons respond to taste independently of hunger, yet in the secondary taste cortex, food-related taste neurons (e.g. responding to sweet taste) only respond to food if hunger is present, and gradually stop responding to that taste during feeding to satiety (Rolls 2005a, 2006a). Now the quality of the tastant (sweet, salt, etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety (Rolls 2005a). The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex), and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it).

The same argument arises for reward in general, and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is reflected in the responses of neurons to visual stimuli (Rolls 2005a, 2006a). It is also of interest that reward signals (e.g. the taste of food when we are hungry) are associated with subjective feelings of pleasure (Rolls 2005a, 2006a). 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 conscious of representations only 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 (Rolls 2005a). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have higher-order thoughts about the pleasantness of the representation in the secondary taste cortex. There would also need to be pathways from the hippocampus, implicated in the recall of declarative memories, back to the language areas of the cerebral cortex (at least via the cortical areas which receive backprojections from the amygdala, orbitofrontal cortex, and hippocampus, see Fig. 4.2, which would in turn need connections to the language areas).

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, 2005b) 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. 4.2).

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, 2005a). 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.)

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

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.)

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

I now consider some clarifications of the present proposal, and how it deals with some issues that arise when considering theories of the phenomenal aspects of consciousness. First, the present proposal has as its foundation the type of computation that is being performed, and suggests that it is a property of a HOST system used for correcting multistep plans with its representations grounded in the world that it would feel like something for a system to be doing this type of processing. To do this type of processing, the system would have to be able to recall previous multistep plans, and would require syntax to keep the symbols in each step of the plan separate. In a sense, the system would have to be able to recall and take into consideration its earlier multistep plans, and in this sense *report* to itself, on those earlier plans. Some approaches to consciousness take the ability to report on or make a *commentary* on events as being an important marker for consciousness (Weiskrantz 1997), and the computational approach I propose suggests why there should be a close relation between consciousness and the ability to report or provide a commentary, for the ability to report is involved in using higher-order syntactic thoughts to correct a multistep plan. Second, the implication of the present approach is that the type of linguistic processing or reporting need not be verbal, using natural language, for what is required to correct the plan is the ability to manipulate symbols syntactically, and this could be implemented in a much simpler type of mentalese or syntactic system (Fodor 1994; Jackendoff 2002; Rolls 2004a) than verbal language or natural language which implies a universal grammar. Third, this approach to consciousness suggests that the information must be being processed in a system capable of implementing

HOSTs for the information to be conscious, and in this sense is more specific than global workspace hypotheses (Baars 1988; Dehaene and Naccache 2001; Dehaene *et al.* 2006). Indeed, the present approach suggests that a workspace could be sufficiently global to enable even the complex processing involved in driving a car to be performed, and yet the processing might be performed unconsciously, unless HOST (supervisory, monitory, correcting) processing was involved. Fourth, the present approach suggests that it just is a property of HOST computational processing with the representations grounded in the world that it feels like something. There is to some extent an element of mystery about why it feels like something, why it is phenomenal, but the explanatory gap does not seem so large when one holds that the system is recalling, reporting on, reflecting on, and reorganizing information about itself in the world in order to prepare new or revised plans. In terms of the physicalist debate (see for a review see Davies, this volume), an important aspect of my proposal is that it is a *necessary* property of this type of (HOST) processing that it feels like something (the philosophical description is that this is an absolute metaphysical necessity), and given this view, then it is up to one to decide whether this view is consistent with one's particular view of physicalism or not. Similarly, the possibility of a zombie is inconsistent with the present hypothesis, which proposes that it is by virtue of performing processing in a specialized system that can perform higher-order syntactic processing with the representations grounded in the world that phenomenal consciousness is necessarily present.

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.

4.7 **Dual routes to action**

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

The first route is via the brain systems that have been present in non-human primates such as monkeys, and to some extent in other mammals, for millions of years. These systems include the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. These systems control behaviour in relation to previous associations of stimuli with reinforcement. The computation which controls the action thus involves assessment of the reinforcement-related value of a stimulus. This assessment may be based on a number of different factors. One is the previous reinforcement history, which involves stimulus–reinforcement association learning using the amygdala, and its rapid updating especially in primates using the orbitofrontal cortex. This stimulus–reinforcement association learning may involve quite specific information about a stimulus, for example of the energy associated with each type of food, by the process of conditioned appetite and satiety (Booth 1985). A second is the current motivational state, for example whether hunger is present, whether other needs are satisfied, etc. A third factor which affects the computed reward value of the stimulus is whether that reward has been received recently. If it has been received recently but in small quantity, this may increase the reward value of the stimulus. This is known as incentive motivation or the 'salted peanut' phenomenon. The adaptive value of such a process is that this positive feedback of reward value in the early stages of working for a particular reward tends to lock the organism into behaviour being performed for that reward. This means that animals that are, for example, almost equally hungry and thirsty will show hysteresis in their choice of action, rather than continually switching from eating to drinking and back with each mouthful of water or food. This introduction of hysteresis into the reward evaluation system makes action selection a much more efficient process in a natural environment, for constantly switching between different types of behaviour would be very costly if all the different rewards were not available in the same place at the same time. (For example, walking half a mile between a site where water was available and a site where food was available after every mouthful would be very inefficient.) The amygdala is one structure that may be involved in this increase in the reward value of stimuli early on in a series of presentations, in that lesions of the amygdala (in rats) abolish the expression of this reward-incrementing process which is normally evident in the increasing rate of working for a food reward early on in a meal (Rolls 2005a). A fourth factor is the computed absolute value of the reward or punishment expected or being obtained from a stimulus, e.g., the sweetness of the stimulus (set by evolution so that sweet stimuli will tend to be rewarding, because they are generally associated with energy sources), or the pleasantness of touch (set by evolution to be pleasant according to the extent to which it brings animals of the opposite sex together, and depending on the investment in time that the partner is willing to put into making the touch pleasurable, a sign which indicates the commitment and value for the partner of the relationship). After the reward value of the stimulus has been assessed in these ways, behaviour is then initiated based on approach towards or withdrawal from the stimulus. A critical aspect of the behaviour produced by this type of system is that it is aimed directly towards obtaining a sensed or expected reward, by virtue of connections to brain systems such as the basal ganglia and cingulate cortex (Rolls 2007c) which are concerned with the initiation of actions (see Fig. 4.2). The expectation may of course involve behaviour to obtain stimuli associated with reward, which might even be present in a chain.

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

The second route in humans involves a computation with many 'if … then' statements, to implement a plan to obtain a reward. In this case, the reward may actually be *deferred* as part of the plan, which might involve working first to obtain one reward, and only then to work for a second more highly valued reward, if this was thought to be overall an optimal strategy in terms of resource usage (e.g. time). In this case, syntax is required, because the many symbols (e.g. names of people) that are part of the plan must be correctly linked or bound. Such linking might be of the form: 'if A does this, then B is likely to do this, and this will cause C to do this …'. The requirement of syntax for this type of planning implies that an output to language systems in the brain is required for this type of planning (see Fig. 4.2). This the explicit language system in humans may allow working for deferred rewards by enabling use of a one-off, individual, plan appropriate for each situation. Another building block for such planning operations in the brain may be the type of short-term memory in which the prefrontal cortex is involved. This short-term memory may be, for example in non-human primates, of where in space a response has just been made. A development of this type of short-term response memory system in humans to enable multiple short-term memories to be held in place correctly, preferably with the temporal order of the different items in the short-term memory coded correctly, may be another building block for the multiple step 'if … then' type of computation in order to form a multiple-step plan. Such short-term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of non-human primates and humans (Goldman-Rakic 1996; Petrides 1996; Rolls 2008), and may be part of the reason why prefrontal cortex damage impairs planning (Shallice and Burgess 1996).

Of these two routes (see Fig. 4.2), it is the second which I have suggested above is related to consciousness. The hypothesis is that consciousness is the state which arises by virtue of having the ability to think about one's own thoughts, which has the adaptive value of enabling one to correct long, multi-step syntactic plans. This latter system is thus the one in which explicit, declarative, processing occurs. Processing in this system is frequently associated with reason and rationality, in that many of the consequences of possible actions can be taken into account. The actual computation of how rewarding a particular stimulus or situation is or will be probably still depends on activity in the orbitofrontal and amygdala, as the reward value of stimuli is computed and represented in these regions, and in that it is found that verbalized expressions of the reward (or punishment) value of stimuli are dampened by damage to these systems. (For example, damage to the orbitofrontal cortex renders painful input still identifiable as pain, but without the strong affective, 'unpleasant', reaction to it.) This language system which enables long-term planning may be contrasted with the first system in which behaviour is directed at obtaining the stimulus (including the remembered stimulus) which is currently most rewarding, as computed by brain structures that include the orbitofrontal cortex and amygdala. There are outputs from this system, perhaps those directed at the basal ganglia, which do not pass through the language system, and behaviour produced in this way is described as implicit, and verbal declarations cannot be made directly about the reasons for the choice made. When verbal declarations are made about decisions made in this first system, those verbal declarations may be confabulations, reasonable explanations, or fabrications, of reasons why the choice was made. These reasonable explanations would be generated to be consistent with the sense of continuity and self that is a characteristic of reasoning in the language system.

The question then arises of how decisions are made in animals such as humans that have both the implicit, direct reward-based, and the explicit, rational, planning systems (see Fig. 4.2) (Rolls 2008). One particular situation in which the first, implicit, system may be especially important is when rapid reactions to stimuli with reward or punishment value must be made, for then the direct connections from structures such as the orbitofrontal cortex to the basal ganglia may allow rapid actions (Rolls 2005a). 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 into control behaviour quite frequently, as otherwise the adaptive value of having the explicit system would be less than optimal.

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

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

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

This discussion of dual routes to action has been with respect to the behaviour produced. There is of course in addition a third output of brain regions such as the orbitofrontal cortex and amygdala involved in emotion, that is directed to producing autonomic and endocrine responses (see Fig. 4.2). Although it has been argued by Rolls (2005a) 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.

4.8 **Comparisons with other approaches to emotion and consciousness**

The theory described here suggests that it feels like something to be an organism or machine that can think about its own (linguistic, and semantically based) thoughts. It is suggested that qualia, raw sensory and emotional subjective feelings, arise secondary to having evolved such a 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 (2000a, 2004a, 2005a); reasons why the ventral visual system is more closely related to explicit than implicit processing (because reasoning about objects may be important) are considered by Rolls (2003) and by Rolls and Deco (2002); and reasons why explicit, conscious, processing may have a higher threshold in sensory processing than implicit processing are considered by Rolls (2003, 2005a, 2005b).

I now compare this approach to emotion and consciousness with that of LeDoux which places some emphasis on working memory. A comparison with other approaches to emotion and consciousness is provided elsewhere (Rolls 2003, 2004a, 2005a, 2005b, 2007a).

A process ascribed to working memory is that items can be manipulated in working memory, for example placed into a different order. This process implies at the computational level some type of syntactic processing, for each item (or symbol) could occur in any position relative to the others, and each item might occur more than once. To keep the items separate yet manipulable into any relation to each other, just having each item represented by the firing of a different set of neurons is insufficient, for this provides no information about the order or more generally the relations between the items being manipulated (Rolls and Deco 2002; Rolls 2008). In this sense, some form of syntax, that is a way to relate to each other the firing of the different populations of neurons each representing an item, is required. If we go this far (and LeDoux 1996, p. 280 does appear to), then we see that this aspect of working memory is very close to the concept I propose of syntactic thought in my HOST theory. My particular approach though makes it clear what the function is to be performed (syntactic operations), whereas the term working memory can be used to refer to many different types of processing (Repovs and Baddeley 2006), and is in this sense less well defined computationally. My approach of course argues that it is thoughts about the first-order thoughts that may be very closely linked to consciousness. In our simple case, the higher-order thought might be 'Do I have the items now in the correct reversed order? Should the *X* come before or after the *Y*?' To perform this syntactic manipulation, I argue that there is a special syntactic processor, perhaps in cortex near Broca's area, that performs the manipulations on the items, and that the dorsolateral prefrontal cortex itself provides the short-term store that holds the items on which the syntactic processor operates (Rolls 2008). In this scenario, dorsolateral prefrontal cortex damage would affect the number of items that could be manipulated, but not consciousness or the ability to manipulate the items syntactically and to monitor and comment on the result to check that it is correct.

A property often attributed to consciousness is that it is *unitary*. LeDoux (2007) might relate this to the limitations of a working memory system. The current theory would account for this by the limited syntactic capability of neuronal networks in the brain, which render it difficult to implement more than a few syntactic bindings of symbols simultaneously (McLeod *et al.* 1998; Rolls 2008). 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. My suggestion is that it is the difficulty the brain has in implementing the syntax required for manipulating items in working memory, and therefore for multi-step planning, and for then correcting these plans, that provides a close link between working memory concepts and my theory of higher-order syntactic processing. The theory I describe makes it clear what the underlying computational problem is (how syntactic operations are performed in the system, and how they are corrected), and argues that when there are thoughts about the system, i.e. HOSTs, and the system is reflecting on its first-order thoughts (cf. Weiskrantz 1997), then it is a property of the system that it feels conscious. As I argued above, first-order linguistic thoughts, which presumably involve working memory (which must be clearly defined for the purposes of this discussion), need not necessarily be conscious.

The theory of emotion described here is also different from LeDoux's approach to affect. LeDoux's (1996, 2007) approach to emotion is largely (to quote him) one of automaticity, with emphasis on brain mechanisms involved in the rapid, subcortical, mechanisms involved in fear. Much of the research described has been on the functions of the amygdala in fear conditioning. The importance of this system in humans may be less than in rats, in that human patients with bilateral amygdala damage do not present with overt emotional problems, although some impairments in fear conditioning and in the expression of fear in the face can be identified (Phelps 2004) (see Rolls 2005a). In contrast, the orbitofrontal cortex has developed greatly in primates, including humans, and major changes in emotion in humans follow damage to the orbitofrontal cortex, evident for example in reward reversal learning; in face expression identification; in behaviour, which can become disinhibited, uncooperative, and impulsive; and in altered subjective emotions after the brain damage (Rolls *et al.* 1994a; Hornak *et al.* 1996, 2003, 2004; Berlin *et al.* 2004; *et al.*; Berlin *et al.* 2005; Rolls 2005a). In addition, it is worth noting that the human amygdala is involved in reward-related processing, in that for example it is as much activated by rewarding sweet taste as by unpleasant salt taste (O'Doherty *et al.* 2001a). Also, although the direct, low road, inputs from subcortical structures to the amygdala have been emphasized, it should be noted that most emotions are not to stimuli that require no cortical processing (such as pure tones), but instead are to complex stimuli (such as the facial expression of a particular individual) which require cortical processing. Thus the cortical to amygdala connections are likely to be involved in most emotional processing that engages the amygdala (Rolls 2005a, 2000b). Consistent with this, amygdala face-selective neurons in primates have longer latencies (e.g. 130–180 ms) than those of neurons in the inferior temporal visual cortex (typically 90–110 ms) (Rolls 1984, 2005a; Leonard *et al.* 1985). Simple physical aspects of faces such as horizontal spatial frequencies reflecting the mouth might reach the amygdala by non-cortical routes, but inputs that reflect the identity of a face, as well as the expression, which are important in social behaviour, are likely to require cortical processing because of the complexity of the computations involved in invariant face and object identification (Rolls and Deco 2002; Rolls 2005a; Rolls and Stringer 2006). Although it is of interest that the amygdala can be activated by face stimuli that are not perceived in backward masking experiments (Phillips *et al.* 2004, LeDoux this volume), so too can the inferior temporal visual cortex (Rolls and Tovee 1994; Rolls *et al.* 1994b, 1999; Rolls 2003, 2005b), so the amygdala just follows the cortex in this respect. The implication here is that activation of the amygdala by stimuli that are not consciously perceived need not be due to subcortical routes to the amygdala. Temporal visual cortex activity also occurs when the stimulus is minimized by backward masking, and could provide a route for activity even when not consciously seen to reach the amygdala. It is suggested that the higher threshold for conscious awareness than for unconscious responses to stimuli, as shown by the larger neuronal responses in the inferior temporal cortex for conscious awareness to be reported, may be related to the fact that conscious processing is inherently serial because of the syntactic binding required, and may because it would be inefficient to interrupt this, have a relatively high threshold (Rolls 2003, 2005b).

Finally, I provide a short specification of what might have to be implemented in a neural network to implement conscious processing. First, a linguistic system, not necessarily verbal, but implementing syntax between symbols grounded in the environment would be needed (e.g. a mentalese language system). Then a 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.

References

Allport, A. (1988). What concept of consciousness? In Marcel, A.J. and Bisiach, E. (eds) *Consciousness in Contemporary Science*, pp. 159–182. Oxford: Oxford University Press.

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

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

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

Berlin, H., Rolls, E.T., and Kischka, U. (2004). Impulsivity, time perception, emotion, and reinforcement sensitivity in patients with orbitofrontal cortex lesions. *Brain* 127, 1108–1126.

Berlin, H., Rolls, E.T., and Iversen, S.D. (2005). Borderline personality disorder, impulsivity and the orbitofrontal cortex. *American Journal of Psychiatry* 162, 2360–2373.

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

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

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

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

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

Cooney, J.W. and Gazzaniga, M.S. (2003). Neurological disorders and the structure of human consciousness. *Trends in Cognitive Science* 7, 161–165.

Darwin, C. (1872) *The Expression of the Emotions in Man and Animals*, 3rd edn. Chicago: University of Chicago Press.

de Araujo, I.E.T., Rolls, E.T., Velazco, M.I., Margot, C., and Cayeux, I. (2005). Cognitive modulation of olfactory processing. *Neuron* 46, 671–679.

Deco, G. and Rolls, E.T. (2005)a). Neurodynamics of biased competition and co-operation for attention: a model with spiking neurons. *Journal of Neurophysiology* 94, 295–313.

Deco, G. and Rolls, E.T. (2005b). Attention, short-term memory, and action selection: a unifying theory. *Progress in Neurobiology* 76, 236–256.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37.

Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Science* 10, 204–211.

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

Ekman, P. (1982). *Emotion in the Human Face*, 2nd edn. Cambridge: Cambridge University Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist* 48, 384–392.

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

Frijda, N.H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

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

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

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

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

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

Gray, J.A. (1975). *Elements of a Two-Process Theory of Learning*. London: Academic Press.

Gray, J.A. (1987). *The Psychology of Fear and Stress*, 2nd edn. Cambridge: Cambridge University Press.

Hornak, J., Rolls, E.T., and Wade, D. (1996). Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia* 34, 247–261.

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

Hornak, J., O'Doherty, J., Bramham, J., Rolls, E.T., Morris, R.G., Bullock, P.R., and Polkey, C.E. (2004). Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience* 16, 463–478.

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

Humphrey, N.K. (1986). *The Inner Eye*. London: Faber.

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

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

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

Krebs, J.R. and Kacelnik A (1991). Decision making. In Krebs, J.R. and Davies, N.B. (eds) *Behavioural Ecology*, 3rd edn, pp. 105–136. Oxford: Blackwell.

Kringelbach, M.L. and Rolls, E.T. (2003). Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage* 20, 1371–1383.

Kringelbach, M.L. and Rolls, E.T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology* 72, 341–372.

Lazarus, R.S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

LeDoux, J.E. (1996). *The Emotional Brain*. New York: Simon and Schuster.

Leonard, C.M., Rolls, E.T., Wilson, F.A.W., and Baylis, G.C. (1985). Neurons in the amygdala of the monkey with responses selective for faces. *Behavioural Brain Research* 15, 159–176.

Libet, B. (2002). The timing of mental events: Libet's experimental findings and their implications. *Consciousness and Cognition* 11, 291–299; discussion 304–233.

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

Millenson, J.R. (1967). *Principles of Behavioral Analysis*. New York: Macmillan.

Miller, E.K., Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24, 167–202.

O'Doherty, J., Rolls, E.T., Francis, S., Bowtell, R., and McGlone, F. (2001a). The representation of pleasant and aversive taste in the human brain. *Journal of Neurophysiology* 85, 1315–1321.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001b). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* 4, 95–102.

Oatley, K. and Jenkins, J.M. (1996). *Understanding Emotions*. Oxford: Backwell.

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

Phelps, E.A. (2004). Human emotion and memory: interactions of the amygdala and hippocampal complex. *Current Opinions in Neurobiology* 14, 198–202.

Phillips, M.L., Williams, L.M., Heining, M., Herba, C.M., Russell, T., Andrew, C., Bullmore, E.T., Brammer, M.J., Williams, S.C., Morgan, M., Young, A.W., and Gray, J.A. (2004). Differential neural responses to overt and covert presentations of facial expressions of fear and disgust. *NeuroImage* 21, 1484–1496.

Repovs, G. and Baddeley, A. (2006). The multi-component model of working memory: explorations in experimental cognitive psychology. *Neuroscience* 139, 5–21.

Rolls, E.T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology* 3, 209–222.

Rolls, E.T. (1986a). A theory of emotion, and its application to understanding the neural basis of emotion. In Oomura, Y. (ed.) *Emotions. Neural and Chemical Control*, pp. 325–344. Basel: Karger.

Rolls, E.T. (1986b). Neural systems involved in emotion in primates. In Plutchik, R. and Kellerman, H. (eds) *Emotion: Theory, Research, and Experience. Vol.* 3. *Biological Foundations of Emotion*, pp. 125–143. New York: Academic Press.

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

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

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

Rolls, E.T. (1997a). Consciousness in neural networks? *Neural Networks* 10, 1227–1240.

Rolls, E.T. (1997b). Brain mechanisms of vision, memory, and consciousness. In Ito, M., Miyashita, Y., and Rolls, E.T. (eds) *Cognition, Computation, and Consciousness*, pp. 81–120. Oxford: Oxford University Press.

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

Rolls, E.T. (1999b). The functions of the orbitofrontal cortex. *Neurocase* 5, 301–312.

Rolls, E.T. (2000a). Prècis of *The Brain and Emotion*. *Behavioral and Brain Sciences* 23, 177–233.

Rolls, E.T. (2000b). Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In Aggleton, J.P. (ed.) *The Amygdala: A Functional Analysis*, 2nd edn, pp. 447–478. Oxford: Oxford University Press.

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

Rolls, E.T. (2004a). A higher order syntactic thought (HOST) theory of consciousness. In Gennaro, R.J. (ed.) *Higher-Order Theories of Consciousness: An Anthology*, pp. 137–172. Amsterdam: John Benjamins.

Rolls, E.T. (2004b). The functions of the orbitofrontal cortex. *Brain and Cognition* 55, 11–29.

Rolls, E.T. (2005a). *Emotion Explained*. Oxford: Oxford University Press.

Rolls, E.T. (2005b). Consciousness absent or present: a neurophysiological exploration of masking. In Ogmen, H. and Breitmeyer, B.G. (eds) *The First Half Second: The Microgenesis and Temporal Dynamics of Unconscious and Conscious Visual Processes*, pp. 89–108, Chapter 106. Cambridge, MA: MIT Press.

Rolls, E.T. (2006a). Brain mechanisms underlying flavour and appetite. *Philosophical Transactions of the Royal Society London Series B Biological Sciences* 361, 1123–1136.

Rolls, E.T. (2006b). The neurophysiology and functions of the orbitofrontal cortex. In Zald, D.H. and Rauch, S.L. (eds) *The Orbitofrontal Cortex*, pp. 95–124. Oxford: Oxford University Press.

Rolls, E.T. (2007a). The affective neuroscience of consciousness: higher order linguistic thoughts, dual routes to emotion and action, and consciousness. In Zelazo, P., Moscovitch, M. and Thompson, E. (eds) *Cambridge Handbook of Consciousness*, pp. 831–859. Cambridge: Cambridge University Press.

Rolls, E.T. (2007b). The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia* 45, 125–143.

Rolls, E.T. (2007c). The anterior and midcingulate cortices and reward. In Vogt, B.A. (ed.) *Cingulate Neurobiology and Disease*. Oxford: Oxford University Press.

Rolls, E.T. (2008). *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach*. Oxford: Oxford University Press.

Rolls, E.T. and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rolls, E.T. and Kesner, R.P. (2006). A computational theory of hippocampal function, and empirical tests of the theory. *Progress in Neurobiology* 79, 1–48.

Rolls, E.T. and Stringer, S.M. (2001). A model of the interaction between mood and memory. *Network: Computation in Neural Systems* 12, 111–129.

Rolls, E.T. and Stringer, S.M. (2006). Invariant visual object recognition: a model, with lighting invariance. *Journal of Physiology (Paris)* 100, 43–62.

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

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

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

Rolls, E.T., Tovee, M.J., Purcell, D.G., Stewart, A.L., and Azzopardi, P. (1994b). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research* 101, 473–484.

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

Rolls, E.T., Critchley, H.D., Browning, A.S., and Inoue, K. (2006). Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research* 170, 74–87.

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

Rosenthal, D.M. (1990). *A Theory of Consciousness*. Bielefeld: Zentrum für Interdisziplinaire Forschung.

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

Rosenthal, D.M. (2004). Varieties of higher-order theory. In Gennaro, R.J. (ed.) *Higher Order Theories of Consciousness*. Amsterdam: John Benjamins.

Rosenthal, D.M. (2005). *Consciousness and Mind*. Oxford: Oxford University Press.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L., and Group, T.P.R. (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

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

Smith-Swintosky, V.L., Plata-Salaman, C.R., and Scott, T.R. (1991). Gustatory neural encoding in the monkey cortex: stimulus quality. *Journal of Neurophysiology* 66, 1156–1165.

Squire, L.R. and Zola, S.M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the USA* 93, 13515–13522.

Strongman, K.T. (1996). *The Psychology of Emotion*, 4th edn. London: Wiley.

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

Weiskrantz, L. (1968). Emotion. In Weiskrantz, L. (ed.) *Analysis of Behavioural Change*, pp. 50–90. New York: Harper and Row.

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

Yaxley, S., Rolls, E.T., and Sienkiewicz, Z.J. (1990). Gustatory responses of single neurons in the insula of the macaque monkey. *Journal of Neurophysiology* 63, 689–700.