The neuroscience of purpose, meaning, and morals

Edmund T. Rolls
Oxford Centre for Computational Neuroscience
Oxford, England

Email: Edmund.Rolls@oxcns.org
Web: www.oxcns.org

Chapter 5 pp. 68-86 in Neuroexistentialism: Meaning, Morals and Purpose in the Age of Neuroscience. Eds. Caruso,G.D. and Flanagan,O. (Eds) Oxford University Press: New York. 2018.

Abstract

One process to which 'purpose' can refer is that genes are self-replicating. Another process to which 'purpose' can apply is that genes set some of the goals for actions. These goals are fundamental to understanding emotion. Another process to which 'purpose' can apply is that syntactic multistep reasoning provides a route for goals to be set that are to the advantage of the individual, the phenotype, and not of the genes.

With this approach, the 'meaning of life' can be interpreted as what results through evolution using self-replicating genetic mechanisms that can build individuals with wonderful brains that using this reasoning process can even reflect on themselves, on their emotional states, and on the meaning of life (Rolls, 2016, Cerebral Cortex: Principles of Operation. Oxford University Press).

Meaning within the brain can be achieved by neural representations not only if these representations have mutual information with objects and events in the world, but also by virtue of the goals just described of the 'selfish' genes, and of the individual reasoner. This it is suggested provides a means for even symbolic representations to be grounded in the world.

Morals can be considered as principles that are underpinned by (the sometimes different) biological goals specified by the genes and by the reasoning (rational) systems. Given that what is 'natural' does not correspond to what is 'right', it is suggested that these conflicts within and between individuals can be addressed by a social contract.

In this Chapter, I build on detailed evidence and theories about the neural bases of emotion (Rolls, 2013, 2014b, a, 2016a) and their implications (Rolls, 2012b) that are described elsewhere, and further develop the ideas that arise about purpose, meaning, and ethics.

I emphasise that a great deal of evidence is available in the sources cited, and that evidence provides the foundation for the ideas considered further here.

1. The neuroscience of purpose

There are a number of ways in which the notion of 'purpose' can be approached in neuroscience.

1.1 Gene replication and purpose

One biological sense of purpose is that life is kept going by the self-replicating mechanisms of reproduction. The reproduction can be asexual, with evolution driven mainly by gene mutation, or sexual which has the added advantage that different genes can be brought together in new combinations, which facilitates local hill-climbing in the high dimensional space that genetics can search (Rolls, 2014a, 2016a). Some events of course have to facilitate the start of the whole process, but once self-replicating genes have become possible, the whole process of evolution by the mechanisms of variation and Darwinian natural selection provides a basis for understanding the design of organisms. This is one sense in which the notion of 'purpose' can be considered in neurobiology.

1.2 Seeking gene-identified goals, emotion, and purpose

1.2.1 Emotions as states elicited by instrumental reinforcers (rewards and punishers)

Emotions can usefully be defined (operationally) as states elicited by rewards and punishers which have particular functions (Rolls, 1999; Rolls, 2005a, 2014a). 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 the happiness produced by being given a particular 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, 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 punishers received, omitted, or terminated. A diagram summarizing some of the emotions associated with the delivery of a reward or punisher or a stimulus associated with them, or with the omission of a reward or punisher, is shown in Fig.1.

I consider elsewhere a slightly more formal definition than rewards or punishers, in which the concept of reinforcers is introduced, and it is shown that emotions can be usefully seen as states produced by instrumental reinforcing stimuli (Rolls, 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 associative learning, because of their association with such primary reinforcers, thereby becoming "secondary reinforcers".

This foundation has been developed (Rolls, 2005a, 2014a) 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. 1).
2. The *intensity* of the reinforcer (see Fig. 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 behavioral response* is possible. (For example, if an active behavioral response can occur to the omission of a positive reinforcer, then anger might be produced, but if only passive behavior 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 (Rolls, 2005a, 2014a). 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. I 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 (Rolls, 2014a).

1.2.2 Gene-defined goals provide a purpose for action

Given that emotions can be considered as states elicited by goals for action, we may ask what the evolutionary adaptive value is of emotions. It turns out that the design of brains to seek rewards and avoid punishers is highly adaptive, and provides another neurobiological approach to purpose, in that it is adaptive for animals to be designed by evolution to seek goals, as described next.

The most important function of emotion is that it is related to seeking goals, and obtaining or not-obtaining the goals, as follows, and as described more fully elsewhere (Rolls, 2012b, 2013, 2014a). This is the first function of emotion.

Emotions, and goals that provide a purpose for action.

Emotional (and motivational) states allow a simple interface between sensory inputs and action systems, which allow for flexibility of behavioral responses to reinforcing stimuli. The essence of this idea is that goals for behavior are specified by reward and punishment evaluation. When an environmental stimulus has been decoded as a primary reward or punishment, or (after previous stimulus-reinforcer association learning) a secondary rewarding or punishing stimulus, then it becomes a goal for action. The human can then perform any action to obtain the reward, or to avoid the punisher. (Instrumental learning typically allows any action to be learned, though some actions may be more easily learned than others (Lieberman, 2000; Pearce, 2008).) 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 human 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 emotional behaviour. The first is stimulus-reinforcer association learning; emotional states are produced as a result (Rolls, 2014a). This process is implemented in structures such as the orbitofrontal cortex and amygdala (Fig. 2) (Rolls and Grabenhorst, 2008; Grabenhorst and Rolls, 2011; Rolls, 2014a). The second is instrumental learning of an action made to approach and obtain the reward or to avoid or escape from the punisher. This is action-outcome learning, and involves brain regions such as the cingulate cortex when the actions are being guided by the goals, and the striatum and rest of the basal ganglia when the behaviour becomes automatic, and habit-based, that is, uses stimulus-response connections (Fig. 2) (Rolls, 2005a, 2009; Rushworth et al., 2011; Rolls, 2014a). 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 (the goals for the action),

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 that natural selection has found for how genes can influence behavior to promote their fitness (as measured by reproductive success), and for how the brain could interface sensory systems to action systems, and is an important part of Rolls' theory of emotion (2005a, 2014a).

The implication is that operation by animals (including humans) using reward and 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 and punishment systems tuned to fitness-related dimensions of the environment, if arbitrary actions are to be made by the animals, rather than just preprogrammed movements such as tropisms, taxes, reflexes, and fixed action patterns. This view of brain design in terms of reward and 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 (Rolls, 2005a; Rolls, 2011a; Rolls, 2014a).

The point being made then is that another sense in which behavior can be described as purposive is that when genes specify rewards and punishers, they provide the goals for action, that is, the purposes for actions. We are built to be goal seeking machines, in the interests of our selfish genes, which find the specification of rewards and punishers an efficient way to guide our "purposive" behavior for the genes' reproductive success.

Selecting between available rewards with their associated costs, and avoiding punishers with their associated costs, is a process that can take place both implicitly (unconsciously), and explicitly using a language system to enable long-term plans to be made. These many different brain systems, some involving implicit evaluation of rewards, and others explicit, verbal, conscious, evaluation of rewards and planned long-term goals, must all enter into the selector of behavior (see Fig. 2).

Other functions of emotion are described elsewhere (Rolls, 2012b, 2013, 2014a):

1.2 Goal seeking, reasoning, the individual, and purpose:

A separate, rational, reasoning, conscious system for identifying emotional goals.

I have put forward a position that in addition to the gene-based goal system for emotion described above, there is a separate rational, that is reasoning, system that can plan ahead and work for what are sometimes different, long-term, goals (Rolls, 1997a, 2003, 2004, 2005b, a, 2007a; Rolls, 2007b; Rolls, 2008b; Rolls, 2011b, 2012b, 2013, 2014a). This type of processing involves multistep trains of thought, as might be required to formulate a plan with many steps. Each step has its own symbols (e.g. a word to represent a person), and so syntactic linking (binding) is needed between the symbols within each step, and some syntactic (relational) links must be made between symbols in different steps. I have argued that when we correct such multi-step plans or trains of thought, we need to think about these first order thoughts, and the system that does this is thus a higher order thought system (in that it is thinking about first order thoughts).

There is a fundamentally important distinction here: working for a gene-specified reward, as in many emotions, is performed for the interests of the "selfish" genes. Working for rationally planned rewards may be performed in the interest of the particular individual (e.g. the person, the phenotype), and not in the interests of the genotype (Rolls, 2011b).

It is suggested that this arbitrary symbol manipulation using important aspects of language processing and used for planning but not in initiating all types of behavior 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 people's) thoughts, that is in a system capable of second or higher order thoughts (Rosenthal, 1986, 1990; Dennett, 1991; Rosenthal, 1993; Rolls, 1995; Carruthers, 1996; Rolls, 1997a; Rolls, 1997b, 1999; Gennaro, 2004; Rolls, 2004; Rosenthal, 2004; Rolls, 2005a; Rosenthal, 2005; Rolls, 2007a, 2014a).

It is of great 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 behavior and thus facilitating the possible identification of other positive reinforcers in the environment. In humans, grief may be particularly and especially potent because it becomes represented in a system which can plan ahead, and understand the enduring implications of the loss (Rolls, 2016c).

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. 2) (Rolls, 2008a). 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 structures such as the orbitofrontal cortex may be especially important (Rolls, 2005a, 2014a). 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 habit, or from action-outcome goal-directed, behaviour, to the explicit conscious control system which can evaluate with its long-term planning algorithms what action should be performed next. Indeed, it would be adaptive for the explicit system to regularly be assessing performance by the more automatic system, and to switch itself in to control behavior quite frequently, as otherwise the adaptive value of having the explicit system would be less than optimal.

It may be expected that there is often a conflict between these systems, in that the first, implicit, system is able to guide behavior 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 that may be in the interests of the individual not the genes. For example, an individual might decide not to have children, but instead to devote himself or herself to being a creative individual, or to enjoying opera, etc. 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, 2008a, 2011b).

The point being made here is that another sense in which behavior can be described as purposive is that actions in humans (and perhaps in related animals) may be performed where the goal is the interest of the individual, the phenotype, using a reasoning system that can calculate using multistep reasoning the advantages and costs to the individual of performing an action to obtain a goal. A goal for the individual might be to live a long, intellectually productive, and healthy life, and such a person might decide for example to forgo a gene-identified goal such as the delicious taste of sweet and flavour (mouth-feel) of fat (such as in an ice cream), in order to remain healthy. The point has been developed elsewhere that the implicit and explicit systems that define goals for action, with their somewhat different interests, are likely to remain across a population more or less in balance, for a strong a gene-defined emotional system will facilitate reproduction, whereas a rational system may not do this, but may provide rewards for the individual.

It should be noted that the evolution of the rational system may have occurred because it conferred significant advantages in terms of reproductive success, but that nevertheless once such a system evolved, the properties of this system enabled it to compute goals that might be more in the interests of the individual than of the genes.

2. The neuroscience of meaning:
Content and meaning in representations: How are representations grounded in the world?

One sense of “meaning” is “purpose”, and this has been considered in Section 1. Another sense of meaning is how it is that representations in the brain have “meaning” and content. I now describe what I understand by representations being grounded in the world, which addresses how representations, and even the symbols used in language, have meaning. These concepts were

developed as part of the much larger issue of the nature and functions of consciousness (Rolls, 2012b, 2014a, 2016a).

It is possible to analyse how the firing of populations of neurons encodes information about stimuli in the world (Rolls and Treves, 2011; Rolls, 2016a). For example, from the firing rates of small numbers of neurons in the primate inferior temporal visual cortex, it is possible to know which of 20 faces has been shown to the monkey (Rolls et al., 1997; Rolls and Treves, 2011). Similarly, a population of neurons in the anterior part of the macaque temporal lobe visual cortex has been discovered that has a view-invariant representation of objects (Booth and Rolls, 1998; Rolls, 2012a). From the firing of a small ensemble of neurons in the olfactory part of the orbitofrontal cortex, it is possible to know which of eight odours was presented (Rolls et al., 1996; Rolls et al., 2010). From the firing of small ensembles of neurons in the hippocampus, it is possible to know where in allocentric space a monkey is looking (Rolls et al., 1998). In each of these cases, the number of stimuli that is encoded increases exponentially with the number of neurons in the ensemble, so this is a very powerful representation (Rolls et al., 1997; Franco et al., 2004; Rolls et al., 2004; Aggelopoulos et al., 2005; Rolls and Treves, 2011; Rolls, 2016a).

What is being measured in each example is the mutual information between the firing of an ensemble of neurons and which stimuli are present in the world. In this sense, one can read off the code that is being used at the end of each of these sensory systems (Rolls and Treves, 2011; Rolls, 2016a).

However, what sense does the representation make to the animal? What does the firing of each ensemble of neurons 'mean'? What is the content of the representation? In the visual system, for example, it is suggested that the representation is built by a series of appropriately connected competitive networks, operating with a modified Hebb-learning rule (Rolls, 2012a; Rolls, 2016a). Now competitive networks categorize their inputs without the use of a teacher (Rolls, 2016a, b). So which particular neurons fire as a result of the self-organization to represent a particular object or stimulus is arbitrary. What meaning, therefore, does the particular ensemble that fires to an object have? How is the representation grounded in the real world? The fact that there is mutual information between the firing of the ensemble of cells in the brain and a stimulus or event in the world (Rolls and Treves, 2011; Rolls, 2016a) partly, but does not fully, answer this question.

One answer to this question is that there may be meaning in the case of objects and faces that it is an object or face, and not just a particular view. This is the case in that the representation may be activated by any view of the object or face. This is a step, suggested to be made possible by a short-term memory in the learning rule that enables different views of objects to be associated together (Rolls, 2012a; Rolls, 2016a). But it still does not provide the representation with any meaning in terms of the real world. What actions might one make, or what emotions might one feel, if that arbitrary set of temporal cortex visual cells was activated?

This leads to one of the answers I propose. I suggest that one type of meaning of representations in the brain is provided by their reward (or punishment) value: activation of these representations is the goal for actions. In the case of primary reinforcers such as the taste of food or pain, the activation of these representations would have meaning in the sense that the animal would work to obtain the activation of the taste of food neurons when hungry, and to escape from stimuli that cause the neurons representing pain to be activated. Evolution has built the brain so that genes specify these primary reinforcing stimuli, and so that their representations in the brain should be the targets for actions (Rolls, 2014a, 2016a). In the case of other ensembles of neurons in, for example, the visual cortex that respond to objects with the colour and shape of a banana, and which 'represent' the sight of a banana in that their activation is always and uniquely produced by the sight of a banana, such representations come to have meaning only by association with a primary reinforcer, involving the process of stimulus-reinforcer association learning.

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

this second sense, meaning will be conferred on the visual sensory representation because of its associations in the sensory-motor world. Related views have been developed by the philosopher Ruth Millikan (1984). Thus it is suggested that there are two ways by which sensory representations can be said to be grounded, that is to have meaning, in the real world.

It is suggested that the symbols used in language become grounded in the real world by the same two processes, as follows (Rolls, 2016a).

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

In the second process, the word 'table' may have meaning because it is associated with sensory stimuli produced by tables such as their touch, shape, and sight, as well as other functional properties, such as, for example, being load-bearing, and obstructing movement if they are in the way (Rolls, 2016a).

These points are relevant for my higher-order syntactic thought (HOST) theory of consciousness, by addressing the sense in which the thoughts can be grounded in the world. The HOST theory holds that the thoughts 'mean' something to the individual, in the sense that they may be about the survival of the individual (the phenotype) in the world, which the rational, thought, system aims to maximize (Rolls, 2012b, 2016a).

3. A neuroscience approach to morals

3.1. Biological underpinnings

In this section I consider the neurobiological underpinnings of ethics, with a much fuller account provided by Rolls (2012b).

Rolls (2012b, 2014a)} has argued that much of the foundation of our emotional behaviour arises from specification by genes of primary reinforcers that provide goals for our actions. We have emotional reactions in certain circumstances, such as when we see that we are about to suffer pain, when we fall in love, or if someone does not return a favour in a reciprocal interaction. What is the relation between our emotions, and what we think is right, that is our ethical principles? If we think something is right, such as returning something that has been on loan, is this a fundamental and absolute ethical principle, or might it have arisen from deep-seated biologically based systems shaped to be adaptive by natural selection operating in evolution to select genes that tend to promote the survival of those genes?

Many principles that we regard as ethical principles *might* arise in this way. For example, guilt might arise when there is a conflict between an available reward and a rule or law of society. Jealousy is an emotion that might be aroused in a male if the faithfulness of his partner seems to be threatened by her liaison with another male. In this case the reinforcement contingency that is operating is produced by a punisher, and it may be that males are specified genetically to find this punishing because it indicates a potential threat to their paternity and parental investment. Similarly, a female may become jealous if her partner has a liaison with another female, because the resources available to the 'wife' useful to bring up her children are threatened. Again, the punisher here may be gene-specified. Such emotional responses might influence what we build into some of the ethical principles that surround marriage and partnerships for raising children.

Many other similar examples can be surmised from the area of evolutionary psychology (Ridley, 1993, 1996; Buss, 2015). For example, there may be a set of reinforcers that are genetically specified to help promote social cooperation and even reciprocal altruism, and that might thus influence what we

regard as ethical, or at least what we are willing to accept as ethical principles. Such genes might specify that emotion should be elicited, and behavioural changes should occur, if a cooperating partner defects or 'cheats' (Cosmides and Tooby, 1999). Moreover, the genes may build brains with genetically specified rules that are useful heuristics for social cooperation, such as acting with a strategy of 'generous tit-for tat', which can be more adaptive than strict 'tit-for-tat', in that being generous occasionally is a good strategy to help promote further cooperation that has failed when both partners defect in a strict 'tit-for-tat' scenario (Ridley, 1996). Genes that specify good heuristics to promote social cooperation may thus underlie such complex emotional states as feeling forgiving. There are many other examples, including kin-altruism, which has a genetic basis.

The situation is clarified by the ideas I have advanced about a rational syntactically based reasoning system and how this interacts with an evolutionarily older emotional system with gene-specified rewards. The rational system enables us for example to defer immediate gene-specified rewards, and make longer-term plans for actions that in the long term may have more useful outcomes. This rational system enables us to make reasoned choices, and to reason about what is right. Indeed, it is because of the linguistic system that the naturalistic fallacy becomes an issue. In particular, we should not believe that what is right is what is natural (*the naturalistic fallacy*), because we have a rational system that can go beyond simpler gene-specified rewards and punishers that may influence our actions through brain systems that operate at least partly implicitly, i.e. unconsciously.

The suggestion I make is that in all these cases, and in many others, there are biological underpinnings that determine what we find rewarding or punishing, designed into genes by evolution to lead to appropriate behaviour that helps to increase the fitness of the genes. When these implicit systems for rewards and punishers start to be expressed explicitly (in language) in humans, the explicit rules, rights, and laws that are formalized are those that set out in language what the biological underpinnings 'want' to occur.

Clearly in formulating the explicit rights and laws, some compromise is necessary in order to keep the society stable. When the rights and laws are formulated in small societies, it is likely that individuals in that society will have many of the same genes, and rules such as 'help your neighbour' (but 'make war with "foreigners" ') will probably be to the advantage of one's genes. However, when the society increases in size beyond a small village (in the order of 1000), then the explicitly formalized rules, rights, and laws may no longer produce behaviour that turns out to be to the advantage of an individual's genes. In addition, it may no longer be possible to keep track of individuals in order to maintain the stability of 'tit-for-tat' cooperative social strategies (Dunbar, 1996; Ridley, 1996). In such cases, other factors doubtless come into play to additionally influence what groups hold to be right. For example, a group of subjects in a society might demand the `right' to free speech because it is to their economic advantage.

Thus overall it is suggested that many aspects of what a society holds as right and moral, and of what becomes enshrined in explicit 'rights' and laws, are related to biological underpinnings, which have usually evolved because of the advantage to the individual's genes, but that as societies develop, other factors also start to influence what is believed to be 'right' by groups of individuals, related to socioeconomic factors. In both cases, the laws and rules of the society develop so that these 'rights' are protected, but often involve compromise in such a way that a large proportion of the society will agree to, or can be made subject to, what is held as right.

Society may set down certain propositions of what is 'right'. One reason for this is that it may be too difficult on every occasion, and for everyone, to work out explicitly what all the payoffs of each rule of conduct are. A second reason is that what is promulgated as 'right' could actually be to someone else's advantage, and it would not be wise to expose this fully. One way to convince members of society not to do what is apparently in their immediate interest is to promise a reward later. Such deferred rewards are often offered by religions (Rolls, 2012b). The ability to work for a deferred reward using a one-off plan in this way becomes possible, it is suggested, with the evolution of the explicit, propositional, system.

3.2 Ethical principles arising from advantages for the phenotype versus for the genotype

Many of the examples described above of the biological underpinnings to ethical behaviour, what we have come to regard as right and just, reflect advantages to the genotype. These underpinnings related to advantages to the genotype are present in many non-human animals, though

developed to greater or lesser extents. Examples include kin altruism (common in non-human animals), reciprocal altruism including tit-for-tat exchanges and forgiveness, and stakeholder altruism (Rolls, 2012b).

However, I wish to introduce here a new concept related to the underlying biology, that some aspects of what we regard as ethical and right are related to advantage to the phenotype. Phenotypic advantage may not necessarily be to the advantage of the genotype, and indeed we can contrast the 'selfish phene' with the 'selfish gene' (Rolls, 2011b, 2012b). The concept of phenotypic advantage in relation to ethics is that by reasoning, the multiple step syntactic thought brain system may lead people to agree to rules of society that are to the advantage of their bodies, even when there may be no genetic advantage (Rolls, 2012b). Consider 'Thou shalt not kill' and 'Thou shalt not steal'. Both rules could be to the advantage of the individual person, who may wish to stay alive and not be robbed of all the things that she or he enjoys, even though that may not necessarily be to the advantage of genes. Indeed, in the animal kingdom, conflict and killings even within species that are driven by genotypic advantage are common, as in intra-sexual competition as part of sexual selection (Rolls, 2012b), and a lion killing the cubs of his new lioness by a previous father. In this context, the rules 'Thou shalt not kill' and 'Thou shalt not steal' may not be to the genetic advantage of an individual, but may well be to the advantage of the individual person who may wish to stay alive to enjoy life and property, even beyond the age of reproduction.

Helping and care for the elderly is another example where individuals may contract into a society where help and care for the elderly is valued, and is even a 'right', because it may be to their own advantage later on in life, and they are willing to pay some cost (e.g. not stealing from others) to be part of the society in which they can enjoy 'rights' such as not having their possessions stolen. This social contract, agreeing to the rules of a society, has similarities with a contract with an insurance company, in which there is a cost, but also a potential benefit to an individual, and where that advantage may not necessarily be genetic, and may never have been selected for genetically.

The concept here is as follows. At least humans (and possibly some other animals) have evolved to have a brain that can reason, and that has many advantages in enabling them to pursue long-term goals rather than immediate gene-defined goals. However, once one has a reasoning system, it can reason that some of the things that one has been built to enjoy (perhaps for genetic reasons originally, such as wealth, power, good food) might be enjoyed even when there is no genetic advantage, for example by prolonging life into old age, which becomes a 'right' where it is possible.

This results in ethical values and 'rights' which have their (biological) basis in the good of the phenotype, and not necessarily the good of the genotype (though they are not mutually exclusive, and may often work together). This phenotypically underpinned system of rights that becomes enshrined in a social contract is somewhat like the 'right' to a pension and to medical care when elderly (and beyond the age at which genetic potential is likely to be enhanced by longer life), which allows the individual person or phenotype to benefit, by having obeyed the rules of the society earlier, contributing to its insurance provisions. Such a social contract has similarities with a contract with an insurance company: it costs a bit, but may protect your body in the long term. A pension is a bit like such a social contract. One agrees rationally to a cost, as it may benefit you (and not necessarily your genes) in the long term.

3.3 The Social Contract

The view that one is led to is that some of our moral beliefs may be explicit, verbal, formulations of what may reflect factors built genetically by kin selection into behaviour, namely a tendency to favour kin, because they are likely to share some of an individual's genes. In a small society this explicit formulation may be 'appropriate' (from the point of view of the genes), in that many members of that society will be related to that individual. When the society becomes larger, the relatedness may decrease, yet the explicit formulation of the rules or laws of society may not change. In such a situation, it is presumably appropriate for society to make it clear to its members that its rules for what is acceptable and 'right' behaviour are set in place so that individuals can live in safety, and with some expectation of help from society in general.

It is argued that the second biological underpinning of ethics is the evolution in (at least) humans of a reasoning system, which leads humans to values based on phenotypic advantage, which may not always correspond to genetic advantage.

The operation of this reasoning system may encourage acceptance of a social contract, in part because it may be judged to be useful for the individual's kin and genetic fitness, and also for the individual's phenotype, which may for example accept a cost of not harming others and benefitting from that, in order not to be harmed.

Indeed, the view to which this approach based on neuroscience and the evolution of the brain leads is that there may be no absolute rights, or god-given rights (Rolls, 2012b), but that instead there may be rules and laws of a society, and indeed perhaps 'universally' across societies, that may be agreed by a social contract, and these rules and laws may imply 'rights'. Justice may in this approach be used to refer to the implementation of these laws and rules (Hobbes, 1651; Locke, 1689; Rawls, 1971). For example, Thomas Hobbes, beginning from a mechanistic understanding of human beings and the passions, postulates what life would be like without government, a condition which he calls the state of nature. In that state, each person would have a right, or license, to everything in the world. This, Hobbes argues, would lead to a "war of all against all"' (bellum omnium contra omnes), and thus lives that are "solitary, poor, nasty, brutish, and short". This is close to my argument about what would represent the interests of the genotype. In this context, Hobbes then argues that to escape this state of war, men in the state of nature accede to a social contract and establish a civil society, a commonwealth. John Locke (1689) continued the development of Hobbes' argument, though more from a starting point that humans are rational, reasoning people.

Other factors that can influence what is held to be right might reflect socioeconomic advantage to groups or alliances of individuals. It would be then in a sense up to individuals to decide whether they wished to accept the rules, with the costs and benefits provided by the rules of that society, in a form of Social Contract. Individuals who did not agree to the social contract might wish to transfer to another society with a different place on the continuum of costs and potential benefits to the individuals, or to influence the laws and policies of their own society, but acting within its laws. Individuals who attempt to cheat the system or break the laws that operate within a society would be expected to pay a cost in terms of punishment meted out by the society in accordance with its rules and laws, what it considers to be 'right' and 'just', with the underpinnings in genotypic and phenotypic advantage described here and in more detail by Rolls (2012b).

4. Summary and conclusions

One process to which 'purpose' can refer is that genes are self-replicating. Another process to which 'purpose' can apply is that genes set some of the goals for actions. These goals are fundamental to understanding emotion. Another process to which 'purpose' can apply is that syntactic multistep reason provides a route for goals to be set that are to the advantage of the individual, the phenotype, and not of the genes.

Meaning can be achieved by neural representations not only if these representations have mutual information with objects and events in the world, but also by virtue of the goals just described of the 'selfish' genes, and of the individual reasoner. This it is suggested provides a means for even symbolic representations to be grounded in the world.

Morals can be considered as principles that are underpinned by (the sometimes different) biological goals specified by the genes and by the reasoning (rational) systems. Given that what is 'natural' does not correspond to what is 'right', it is suggested that these conflicts within and between individuals can be addressed by a social contract.

Figure Legends

Fig. 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-!). It should be understood that each different reinforcer will produce different emotional states: this diagram just summarizes the types of emotion that may be elicited by different contingencies, but the actual emotions will be different for each reinforcer (Rolls, 2013, 2014a). (emreinf.eps)

Fig. 2. Multiple routes to the initiation of actions and other behavioural responses in response to rewarding and punishing stimuli (Rolls, 2016a). 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. One type of route to behaviour may be implicit, and includes the anterior cingulate cortex for action-outcome, goal-related, learning; and the striatum and rest of the basal ganglia for stimulus-response habits. Another type of route is via the language systems of the brain, which allow explicit (verbalizable) decisions involving multistep syntactic planning to be implemented. Outputs for autonomic responses can also be produced using outputs from the orbitofrontal cortex and anterior cingulate cortex (some of which are routed via the anterior insular cortex) and amygdala. (9_4c.eps)

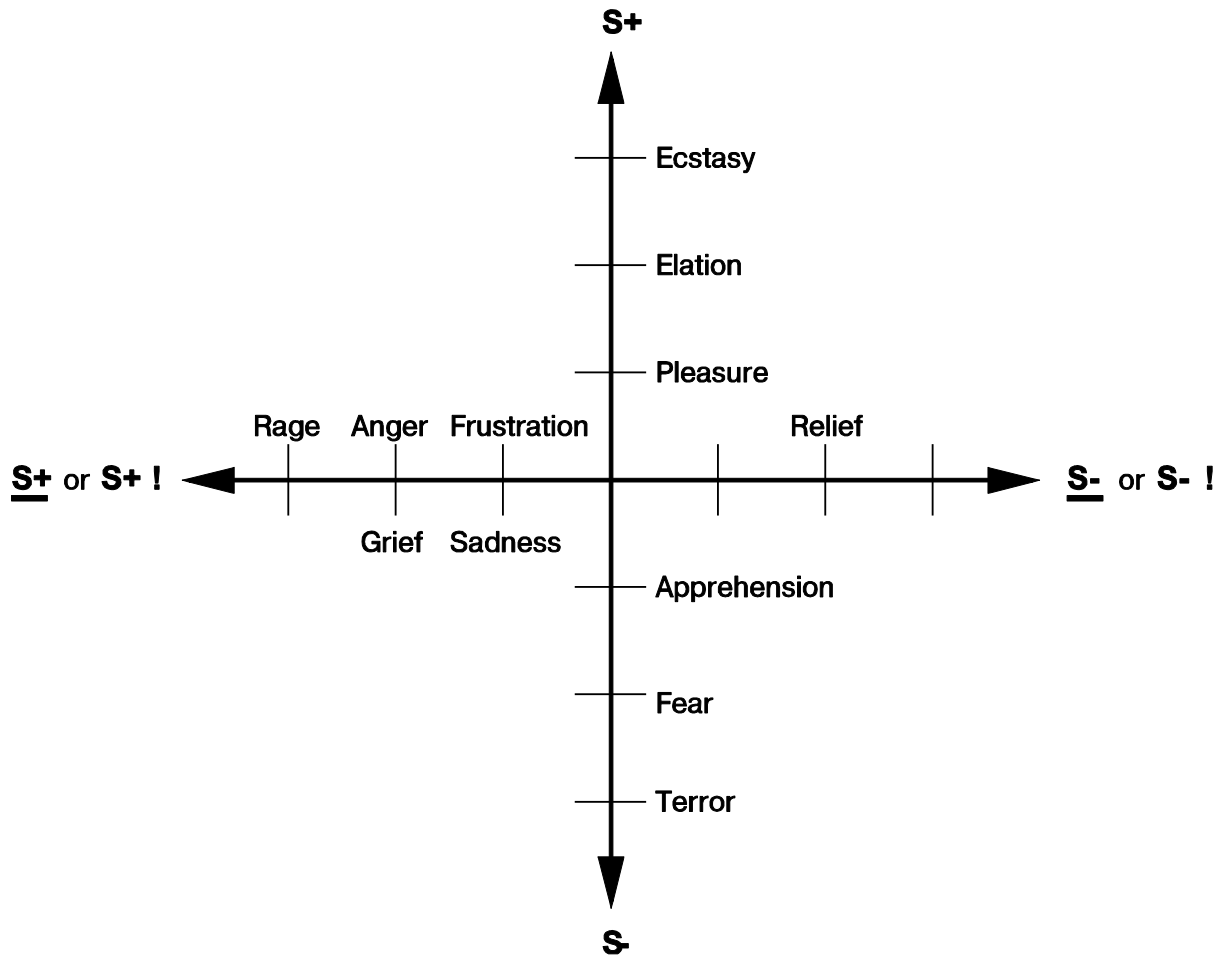
S+
Ecstasy
Elation
Pleasure
Rage
Anger
Frustration
Relief
S+ or S+ !
S- or S- !
Grief
Sadness
Apprehension
Fear
Terror
S-

Fig. 1

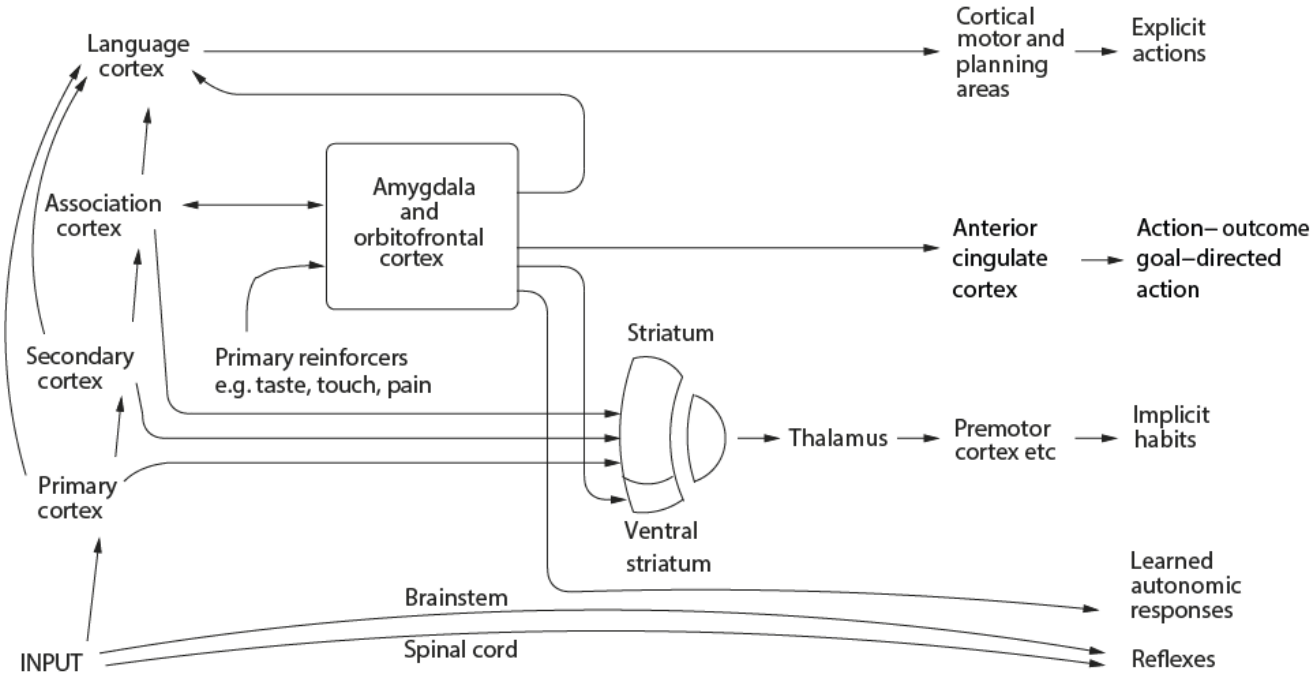
Language cortex
Cortical motor and planning areas
Explicit actions
Association cortex
Amygdala and orbitofrontal cortex
Anterior cingulate cortex
Action–outcome goal–directed action
Striatum
Secondary cortex
Primary reinforcers e.g. taste, touch, pain
Thalamus
Premotor cortex etc
Implicit habits
Primary cortex
Ventral striatum
Learned autonomic responses
Brainstem
Spinal cord
Reflexes
INPUT

Fig. 2

References

Aggelopoulos NC, Franco L, Rolls ET (2005) Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. J Neurophysiol 93:1342-1357.

Booth MCA, Rolls ET (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb Cortex 8:510-523.

Buss DM (2015) Evolutionary Psychology: The New Science of the Mind, 5th Edition. New York: Pearson.

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

Cosmides I, Tooby J (1999) Evolutionary psychology. In: MIT Encyclopedia} of the Cognitive Sciences (Wilson R, Keil F, eds), pp 295-298. Cambridge, MA: MIT Press.

Dennett DC (1991) Consciousness Explained. London: Penguin.

Dunbar R (1996) Grooming, Gossip, and the Evolution of Language. London: Faber and Faber.

Franco L, Rolls ET, Aggelopoulos NC, Treves A (2004) The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons. Exp Brain Res 155:370-384.

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

Grabenhorst F, Rolls ET (2011) Value, pleasure, and choice in the ventral prefrontal cortex. Trends Cogn Sci 15:56-67.

Hobbes T (1651) Leviathan, or the Matter, Forme, and Power of a Commonwealth, Ecclesiasticall and Civil.

Lieberman DA (2000) Learning. Belmont, CA: Wadsworth.

Locke J (1689) The Two Treatises of Civil Government.

Millikan RG (1984) Language, Thought, and Other Biological Categories: New Foundation for Realism. Cambridge, MA: MIT Press.

Pearce JM (2008) Animal Learning and Cognition, 3rd Edition. Hove, UK: Psychology Press.

Rawls J (1971) A Theory of Justice. Oxford: Oxford University Press.

Ridley M (1993) The Red Queen: Sex and the Evolution of Human Nature. London: Penguin.

Ridley M (1996) The Origins of Virtue. London: Viking.

Rolls ET (1990) A theory of emotion, and its application to understanding the neural basis of emotion. Cognition and Emotion 4:161-190.

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

Rolls ET (1997a) Consciousness in neural networks? Neural Netw 10:1227-1240.

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

Rolls ET (1999) The Brain and Emotion. Oxford: Oxford University Press.

Rolls ET (2003) Consciousness absent and present: a neurophysiological exploration. Prog Brain Res 144:95-106.

Rolls ET (2004) A higher order syntactic thought (HOST) theory of consciousness. In: Higher-Order Theories of Consciousness: An Anthology (Gennaro RJ, ed), pp 137-172. Amsterdam: John Benjamins.

Rolls ET (2005a) Emotion Explained. Oxford: Oxford University Press.

Rolls ET (2005b) Consciousness absent or present: a neurophysiological exploration of masking. In: The First Half Second: The Microgenesis and Temporal Dynamics of Unconscious and Conscious Visual Processes (Ogmen H, Breitmeyer BG, eds), pp 89-108, chapter 106. Cambridge, MA: MIT Press.

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

Rolls ET (2007b) A computational neuroscience approach to consciousness. Neural Netw 20:962-982.

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

Rolls ET (2008b) Emotion, higher order syntactic thoughts, and consciousness. In: Frontiers of Consciousness (Weiskrantz L, Davies MK, eds), pp 131-167. Oxford: Oxford University Press.

Rolls ET (2009) The anterior and midcingulate cortices and reward. In: Cingulate Neurobiology and Disease (Vogt BA, ed), pp 191-206. Oxford: Oxford University Press.
Rolls ET (2011a) A neurobiological basis for affective feelings and aesthetics. In: The Aesthetic Mind: Philosophy and Psychology (E.Schellekens, P.Goldie, eds), pp 116-165 Oxford: Oxford University Press.
Rolls ET (2011b) Consciousness, decision-making, and neural computation. In: Perception-Action Cycle: Models, Algorithms and Systems (V.Cutsuridis, A.Hussain, J.G.Taylor, eds), pp 287-333. Berlin: Springer.
Rolls ET (2012a) Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. Front Comput Neurosci 6, 35:1-70.
Rolls ET (2012b) Neuroculture. On the Implications of Brain Science. Oxford: Oxford University Press.
Rolls ET (2013) What are emotional states, and why do we have them? Emot Rev 5:241-247.
Rolls ET (2014a) Emotion and Decision-Making Explained. Oxford: Oxford University Press.
Rolls ET (2014b) Emotion and Decision-Making Explained: Précis. Cortex 59:185-193.
Rolls ET (2016a) Cerebral Cortex: Principles of Operation. Oxford: Oxford University Press.
Rolls ET (2016b) Pattern separation, completion, and categorisation in the hippocampus and neocortex. Neurobiol Learn Mem 129:4-28.
Rolls ET (2016c) A non-reward attractor theory of depression. Neurosci Biobehav Rev 68:47-58.
Rolls ET, Grabenhorst F (2008) The orbitofrontal cortex and beyond: from affect to decision-making. Prog Neurobiol 86:216-244.
Rolls ET, Treves A (2011) The neuronal encoding of information in the brain. Prog Neurobiol 95:448-490.
Rolls ET, Critchley HD, Treves A (1996) The representation of olfactory information in the primate orbitofrontal cortex. J Neurophysiol 75:1982-1996.
Rolls ET, Treves A, Tovee MJ (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. Exp Brain Res 114:177-185.
Rolls ET, Aggelopoulos NC, Franco L, Treves A (2004) Information encoding in the inferior temporal cortex: contributions of the firing rates and correlations between the firing of neurons. Biol Cybern 90:19-32.
Rolls ET, Critchley HD, Verhagen JV, Kadohisa M (2010) The representation of information about taste and odor in the orbitofrontal cortex. Chemosensory Perception 3:16-33.
Rolls ET, Treves A, Robertson RG, Georges-François P, Panzeri S (1998) Information about spatial view in an ensemble of primate hippocampal cells. J Neurophysiol 79:1797-1813.
Rosenthal DM (1986) Two concepts of consciousness. Philosophical Studies 49:329-359.
Rosenthal DM (1990) A theory of consciousness. In: ZIF. Bielefeld, Germany: Zentrum für Interdisziplinaire Forschung.
Rosenthal DM (1993) Thinking that one thinks. In: Consciousness (Davies M, Humphreys GW, eds), pp 197-223. Oxford: Blackwell.
Rosenthal DM (2004) Varieties of Higher-Order Theory. In: Higher Order Theories of Consciousness (Gennaro RJ, ed), pp 17-44. Amsterdam: John Benjamins.
Rosenthal DM (2005) Consciousness and Mind. Oxford: Oxford University Press.
Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054-1069.