Published in final edited form as:
Annu Rev Neurosci. 2012 ; 35: 287–308. doi:10.1146/annurev-neuro-062111-150512.

Neural Basis of Reinforcement Learning and Decision Making

Daeyeol Lee[1,2], **Hyojung Seo**[1], and **Min Whan Jung**[3]

Daeyeol Lee: daeyeol.lee@yale.edu; Hyojung Seo: hyojun.seo@yale.edu; Min Whan Jung: min@ajou.ac.kr

[1]Department of Neurobiology, Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, Connecticut 06510, USA

[2]Department of Psychology, Yale University, New Haven, Connecticut 06520, USA

[3]Neuroscience Laboratory, Institute for Medical Sciences, Ajou University School of Medicine, Suwon 443-721, Republic of Korea

Abstract

Reinforcement learning is an adaptive process in which an animal utilizes its previous experience to improve the outcomes of future choices. Computational theories of reinforcement learning play a central role in the newly emerging areas of neuroeconomics and decision neuroscience. In this framework, actions are chosen according to their value functions, which describe how much future reward is expected from each action. Value functions can be adjusted not only through reward and penalty, but also by the animal's knowledge of its current environment. Studies have revealed that a large proportion of the brain is involved in representing and updating value functions and using them to choose an action. However, how the nature of a behavioral task affects the neural mechanisms of reinforcement learning remains incompletely understood. Future studies should uncover the principles by which different computational elements of reinforcement learning are dynamically coordinated across the entire brain.

Keywords

prefrontal cortex; neuroeconomics; reward; striatum; uncertainty

INTRODUCTION

Decision making refers to the process by which an organism chooses its actions, and has been studied in such diverse fields as mathematics, economics, psychology, and neuroscience. Traditionally, theories of decision making have fallen into two categories. On the one hand, normative theories in economics generate well-defined criteria for identifying best choices. Such theories, including expected utility theory (von Neumann and Morgenstern 1944), deal with choices in an idealized context, and often fail to account for actual choices made by humans and animals. On the other hand, descriptive psychological theories try to account for failures of normative theories by identifying a set of heuristic rules applied by decision makers. For example, prospect theory (Kahneman & Tversky 1979) can successfully account for the failures of expected utility theory in describing human decision making under uncertainty. These two complementary theoretical frameworks are essential for neurobiological studies of decision making. However, a fundamental question not commonly addressed by either approach is the role of learning.

Corresponding authors: Daeyeol Lee, Ph.D. (daeyeol.lee@yale.edu), Department of Neurobiology, Yale University School of Medicine, 333 Cedar Street, SHM B404, New Haven, CT 06510, USA, Min Whan Jung, Ph.D. (min@ajou.ac.kr), Institute for Medical Sciences, Ajou University School of Medicine, Suwon 443-721, Republic of Korea.

How do humans and animals acquire their preference for different actions and outcomes in the first place?

Our goal in this paper is to review and organize recent findings about the functions of different brain areas that underlie experience-dependent changes in choice behaviors. Reinforcement learning theory (Sutton & Barto 1998) has been widely adopted as the main theoretical framework in designing experiments as well as interpreting empirical results. Since this topic has been frequently reviewed (Dayan & Niv 2008, van der Meer & Redish 2011, Ito & Doya 2011, Bornstein & Daw, 2011), we only briefly summarize the essential elements of reinforcement learning theory and focus on the following questions. First, where and how in the brain are the estimates of expected reward represented and updated by the animal's experience? Converging evidence from a number of recent studies suggests that many of these computations are carried out in multiple interconnected regions in the frontal cortex and basal ganglia. Second, how are the value signals for potential actions transformed to the final behavioral response? Competitive interactions among different pools of recurrently connected neurons are a likely mechanism for this selection process (Usher & McClelland 2001, Wang 2002), but their neuroanatomical substrates remain poorly understood (Wang 2008). Finally, how is model-based reinforcement learning implemented in the brain? Humans and animals can acquire new knowledge about their environment without directly experiencing reward or penalty, and this knowledge can be used to influence subsequent behaviors (Tolman 1948). This is referred to as model-based reinforcement learning, whereas reinforcement learning entirely relying on experienced reward and penalty is referred to as model-free. Recent studies have begun to shed some light on how these two different types of reinforcement learning are linked in the brain. We conclude with some suggestions for future research on the neurobiological mechanisms of reinforcement learning.

REINFORCEMENT LEARNING THEORIES OF DECISION MAKING

Economic utilities and value functions

Economic theories of decision making focus on how numbers can be attached to alternative actions so that choices can be understood as selecting an action that has the maximum value among all possible actions. These hypothetical quantities are often referred to as utilities, and can be applied to all types of behaviors. By definition, behaviors chosen by an organism are those that maximize the organism's utility (Figure 1a). These theories are largely agnostic about how these utilities are determined, although they are presumably constrained by evolution and individual experience. By contrast, reinforcement learning theories describe how the animal's experience alters its value functions, which in turn influence subsequent choices (Figure 1b).

The goal of reinforcement learning is to maximize future rewards. Analogous to utilities in economic theories, value functions in reinforcement learning theory refer to the estimates for the sum of future rewards. However, since the animal cannot predict the future changes in its environment perfectly, value functions, unlike utilities, reflect the animal's empirical estimates for its future rewards. Rewards in the distant future are often temporally discounted so that more immediate rewards exert stronger influence on the animal's behavior. The reinforcement learning theory utilizes two different types of value functions. First, action value function refers to the sum of future rewards expected for taking a particular action in a particular state of the environment, and is often denoted by Q(s,a), where s and a refer to the state of the environment and the animal's action, respectively. The term action is used formally: it can refer to not only a physical action, such as reaching to a coffee mug in a particular location with a specific limb, but also an abstract choice, such as buying a particular guitar. Second, state value function, often denoted by V(s), refers to the

sum of future rewards expected from a particular state of the animal's environment. If the animal always chooses only one action in a given state, then its action value function would be equal to the state value function. Otherwise, the state value function would correspond to the average of action value functions weighted by the probability of taking each action in a given state. State value functions can be used to evaluate the action outcomes, but in general, action value functions are required for selecting an action.

Model-free vs. model-based reinforcement learning

Value functions can be updated according to two different types of information. First, they can be revised according to the reward or penalty received by the animal after each action. Value functions would not change and hence no learning would be necessary, if choice outcomes are always perfectly predicted from the current value functions. Otherwise, value functions must be modified to reduce errors in reward predictions. The signed difference between the actual reward and the reward expected by the current value functions is referred to as a reward prediction error (Sutton & Barto 1998). In a class of reinforcement learning algorithms, referred to as simple or model-free reinforcement learning, reward prediction error is the primary source of changes in value functions. More specifically, the value function for the action chosen by the animal or the state visited by the animal is updated according to the reward prediction error, while the value functions for all other actions and states remain unchanged or simply decay passively (Barraclough et al. 2004, Ito & Doya 2009).

In the second class of reinforcement learning, referred to as model-based reinforcement learning, value functions can be changed more flexibly. These algorithms can update the value functions on the basis of the animal's motivational state and its knowledge of the environment without direct reward or penalty. The use of cognitive models allows the animal to adjust its value functions immediately, whenever it acquires a new piece of information about its internal state or external environment. There are many lines of evidence that animals as well as humans are capable of model-based reinforcement learning. For example, when an animal is satiated for a particular type of reward, the subjective value of the same food would be diminished. However, if the animal relies entirely on simple reinforcement learning, the tendency to choose a given action would not change until it experiences the devalued reward through the same action. Previous work has shown that rats can change their behaviors immediately according to their current motivational states following the devaluation of specific food items. This is often used as a test for goal-directed behaviors, and indicates that animals are indeed capable of model-based reinforcement learning (Balleine & Dickinson 1998; Daw et al. 2005). Humans and animals can also simulate the consequences of potential actions that they could have chosen. This is referred to as counterfactual thinking (Roese & Olson 1995), and the information about hypothetical outcomes from unchosen actions can be incorporated into value functions when they are different from the outcomes predicted by the current value functions (Lee et al. 2005, Coricelli et al. 2005, Boorman et al. 2011). Analogous to reward prediction error, the difference between hypothetical and predicted outcomes is referred to as fictive or counterfactual reward prediction error (Lohrenz et al. 2007, Boorman et al. 2011).

Learning during social decision making

During social interaction, the outcomes of actions are often jointly determined by the actions of multiple players. Such strategic situations are referred to as games (von Neumann & Morgenstern 1944). Decision makers can improve the outcomes of their choices during repeated games by applying a model-free reinforcement learning algorithm. In fact, for relatively simple games, such as two-player zero-sum games, humans and animals gradually approximate optimal strategies using model-free reinforcement learning algorithms

(Mookherjee & Sopher 1994, Erev & Roth 1998, Camerer 2003, Lee et al. 2004). On the other hand, players equipped with a model-based reinforcement learning algorithm can adjust their strategies more flexibly according to the predicted behaviors of other players. The ability to predict the beliefs and intentions of other players is often referred to as the theory of mind (Premack & Woodruff 1978), and during social decision making, this may dramatically improve the efficiency of reinforcement learning. In game theory, this is also referred to as belief learning (Camerer 2003).

In pure belief learning, the outcomes of actual choices by the decision makers do not exert any additional influence on their choices, unless such outcomes can modify the beliefs or models of the decision makers about the other players. In other words, reward or penalty does not have any separate roles other than affecting the decision maker's belief about the other players. Results from studies in behavioral economics showed that such pure belief learning models do not account for human choice behaviors very well (Mookherjee & Sopher, 1997, Erev & Roth 1998, Feltovich 2000, Camerer 2003). Instead, human behaviors during repeated games are often consistent with hybrid learning models, such as the experience-weighted attraction model (Camerer & Ho 1999). In these hybrid models, value functions are adjusted by both real and fictive reward prediction errors. Learning rates for these different reward prediction errors can be set independently. Similar to the results from human studies, hybrid models also account for the behaviors of non-human primates performing a competitive game task against a computer better than either model-free reinforcement learning or belief learning model (Lee et al. 2005, Abe & Lee 2011).

NEURAL REPRESENTATION OF VALUE FUNCTIONS

Utilities and value functions are central to economic and reinforcement learning theories of decision making, respectively. In both theories, these quantities are assumed to capture all the relevant factors influencing choices. Thus, brain areas or neurons involved in decision making are expected to harbor signals related to utilities and value functions. In fact, neural activity related to reward expectancy has been found in many different brain areas (Schultz et al. 2000, Hikosaka et al. 2006; Wallis & Kennerley 2010), including sensory cortical areas (Shuler & Bear 2006, Serences 2008; Vickery et al. 2011). The fact that signals related to reward expectancy are widespread in the brain suggests that they are likely to subserve not only reinforcement learning, but also other related cognitive processes, such as attention (Maunsell 2004, Bromberg-Martin et al. 2010a, Litt et al. 2011). Neural signals related to reward expectancy can be divided into at least two different categories, depending on whether they are related to specific actions or states. Neural signals related to action value functions would be useful in choosing a particular action, especially if such signals are observed before the execution of a motor response. Neural activity related to state value functions may play more evaluative roles. In particular, during decision making, the state value function changes from the weighted average of action values for alternative choices to the action value function for the chosen action. The latter is often referred to as chosen value (Padoa-Schioppa & Assad 2006, Cai et al. 2011).

During experiments on decision making, choices can be made among alternative physical movements with different spatial trajectories, or among different objects regardless of the movements required to acquire them. Whereas these different options are all considered actions in the reinforcement learning theory, neural signals related to the corresponding action value functions may vary substantially according to the dimension in which choices are made. In most previous neurobiological studies, different properties of reward were linked to different physical actions. These studies have identified neural activity related to action value functions in numerous brain areas, including the posterior parietal cortex (Platt & Glimcher 1999, Sugrue et al. 2004, Dorris & Glimcher 2004, Seo et al. 2009), dorsolateral

prefrontal cortex (Barraclough et al. 2004, Kim et al. 2008), premotor cortex (Pastor-Bernier & Cisek 2011), medial frontal cortex (Seo & Lee 2007, 2009, Sul et al. 2010, So & Stuphorn 2010), and striatum (Samejima et al. 2005, Lau & Glimcher 2008; Kim et al. 2009; Cai et al. 2011). Despite the limited spatial and temporal resolutions available in neuroimaging studies, metabolic activity related to action value functions has been also identified in the supplementary motor area (Wunderlich et al. 2009). In contrast, neurons in the primate orbitofrontal cortex are not sensitive to spatial locations of targets associated with specific rewards (Tremblay & Schultz 1999, Wallis & Miller 2003), suggesting that they encode action value functions related to specific objects or goals. When animals chose between two different flavors of juice, neurons in the primate orbitofrontal cortex indeed signaled the action value functions associated with specific juice flavors rather than the directions of eye movements used to indicate the animal's choices (Padoa-Schioppa & Assad, 2006).

Neurons in many different brain areas often combine value functions for alternative actions and other decision-related variables. Precisely how multiple types of signals are combined in the activity of individual neurons can therefore provide important clues about how such signals are computed and utilized. For example, likelihood of choosing one of two alternative choices is determined by the difference in their action value functions, and therefore neurons encoding such signals may be closely involved in the process of action selection. During a binary choice task, neurons in the primate posterior parietal cortex (Seo et al. 2009), dorsolateral prefrontal cortex (Kim et al. 2008), premotor cortex (Pastor-Bernier & Cisek 2011), supplementary eye field (Seo & Lee 2009), and dorsal striatum (Cai et al. 2011), as well as the rodent secondary motor cortex (Sul et al. 2011) and striatum (Ito & Doya 2009), encode the difference between the action value functions for two alternative actions.

Signals related to state value functions are also found in many different brain areas. During a binary choice task, the sum or average of the action value functions for two alternative choices corresponds to the state value function before a choice is made. In the posterior parietal cortex and dorsal striatum, signals related to such state value functions and action value functions coexist (Seo et al. 2009, Cai et al. 2011). Neurons encoding state value functions are also found in the ventral striatum (Cai et al. 2011), anterior cingulate cortex (Seo & Lee, 2007), and amygdala (Belova et al. 2008). Neural activity related to chosen values that correspond to post-decision state value functions is also widespread in the brain, and has been found in the orbitofrontal cortex (Padoa-Schioppa & Assad, 2006, Sul et al. 2010), medial frontal cortex (Sul et al. 2010), dorsolateral prefrontal cortex (Kim & Lee 2011), and striatum (Lau & Glimcher 2008, Kim et al. 2009, Cai et al. 2011). Since reward prediction error corresponds to the difference between the outcome of a choice and chosen value, neural activity related to chosen values might be utilized to compute reward prediction errors and update value functions. In some of these areas, such as the dorsolateral prefrontal cortex (Kim & Lee 2011) and dorsal striatum (Cai et al. 2011), activity related to chosen value signals emerges later than the signals related to the sum of the value functions for alternative actions, suggesting that action selection might take place during this delay (Figure 2).

NEURAL MECHANISMS OF ACTION SELECTION

During decision making, neural activity related to action value functions must be converted to the signals related to a particular action and transmitted to motor structures. The precise anatomical location playing a primary role in action selection may vary with the nature of a behavioral task. For example, actions selected by fixed stimulus-action associations or well-practiced motor sequences might rely more on the dorsolateral striatum compared to flexible goal-directed behaviors (Knowlton et al. 1996, Hikosaka et al. 1999, Yin & Knowlton

2006). Considering that spike trains of cortical neurons are stochastic (Softky & Koch 1993), the process of action selection is likely to rely on a network of neurons temporally integrating the activity related to difference in action value functions (Soltani & Wang 2006, Krajbich et al. 2010). An analogous process has been extensively studied for action selection based on noisy sensory stimulus during perceptual decision making. For example, psychophysical performance during a two-alternative forced choice task is well described by the so-called random-walk or drift-diffusion model in which a particular action is selected when the gradual accumulation of noisy evidence reaches a threshold for that action (Laming 1968, Roitman & Shadlen 2002, Smith & Ratcliff 2004).

Neurons in multiple brain areas involved in motor control often build up their activity gradually prior to specific movements, suggesting that these areas might also be involved in action selection. Execution of voluntary movements are tightly coupled with phasic neural activity in a number of brain areas, such as the primary motor cortex (Georgopoulos et al. 1986), premotor cortex (Churchland et al. 2006), frontal eye field (Hanes & Schall 1996), supplementary eye field (Schlag & Schlag-Rey 1987), posterior parietal cortex (Andersen et al. 1987), and superior colliculus (Schiller & Stryker 1972, Wurtz & Goldberg 1972). All of these structures are closely connected with motor nuclei in the brainstem and spinal cord. In addition, neurons in these areas display persistent activity related to the metrics of upcoming movements when the desired movement is indicated before a go signal, suggesting that they are also involved in motor planning and preparation (Smyrnis et al. 1992, Glimcher & Sparks 1992, Weinrich & Wise 1982, Bruce & Goldberg 1985, Schall 1991, Gnadt & Andersen 1988). Such persistent activity has been often associated with working memory (Funahashi et al. 1989, Wang 2001), but may also subserve the temporal integration of noisy inputs (Shadlen & Newsome 2001, Roitman & Shadlen, 2002, Wang, 2002, Curtis & Lee 2010). In fact, neural activity in accordance with gradual evidence accumulation has been found in the same brain areas that show persistent activity related to motor planning, such as the posterior parietal cortex (Roitman & Shadlen 2002), frontal eye field (Ding & Gold 2011), and superior colliculus (Horwitz & Newsome 2001).

Computational studies have demonstrated that a network of neurons with recurrent excitation and lateral inhibition can perform temporal integration of noisy sensory inputs and produce a signal corresponding to an optimal action (Wang 2002, Lo & Wang, 2006, Beck et al. 2008, Furman & Wang 2008). Most of these models have been developed to account for the pattern of activity observed in the lateral intraparietal (LIP) cortex during a perceptual decision making task. Nevertheless, value-dependent action selection might also involve attractor dynamics in a similar network of neurons, provided that their input synapses are adjusted in a reward-dependent manner (Soltani & Wang 2006, Soltani et al. 2006). Therefore, neurons involved in evaluation of unreliable sensory information may also contribute to value-based decision making. Consistent with this possibility, neurons in the LIP tend to change their activity according to the value of rewards expected from alternative actions (Platt & Glimcher 1999, Sugrue et al. 2004, Seo et al. 2009, Louie & Glimcher 2010).

In contrast to the brain areas involved in selecting a specific physical movement, other areas might be involved in more abstract decision making. For example, the orbitofrontal cortex might play a particularly important role in making choices among different objects or goods (Padoa-Schioppa 2011). On the other hand, an action selection process guided by internal cues rather than external sensory stimuli might rely more on the medial frontal cortex. Activity related to action value functions have been found in the supplementary and presupplementary motor areas (Sohn & Lee 2007, Wunderlich et al. 2009), as well as the supplementary eye field (Seo & Lee 2009, So & Stuphorn 2010). More importantly, neural activity related to an upcoming movement appears in the medial frontal cortex earlier than in

other areas of the brain. For example, when human subjects are asked to initiate a movement voluntarily without any immediate sensory cue, scalp EEG displays the so-called readiness potential well before movement onset, and its source has been localized to the supplementary motor area (Haggard 2008, Nachev et al. 2008). The hypothesis that internally generated voluntary movements are selected in the medial frontal cortex is also consistent with the results from single-neuron recording and neuroimaging studies. Individual neurons in the primate supplementary motor area often begin to change their activity according to an upcoming limb movement earlier than those in the premotor cortex or primary motor cortex, especially when the animal is required to produce such movements voluntarily without immediate sensory cues (Tanji & Kurata 1985, Okano & Tanji 1987). Similarly, neurons in the supplementary eye field begin to modulate their activity according to the direction of an upcoming saccade earlier than similar activity recorded in the frontal eye field and LIP (Coe et al. 2002). In rodents performing a dynamic foraging task, signals related to the animal's choice appear in the medial motor cortex, presumably a homolog of the primate supplementary motor cortex, earlier than many other brain areas, including the primary motor cortex and basal ganglia (Sul et al. 2011). Furthermore, lesions in this area make the animal's choices less dependent on action value functions (Sul et al., 2011). Finally, analysis of BOLD activity patterns during a self-timed motor task has also identified signals related to an upcoming movement even several seconds before the movement onset in the human supplementary motor area (Soon et al. 2008).

In summary, neural activity potentially reflecting the process of action selection has been identified in multiple regions, including areas involved in motor control, orbitofrontal cortex, and medial frontal cortex. It would be therefore important to investigate in the future how these multiple areas interact cooperatively or competitively depending on the demands of specific behavioral tasks. The frame of reference in which different actions are represented varies across brain areas, and therefore, how the actions encoded in one frame of reference, for example, in objects space, are transformed to another, such as visual or joint space, needs to be investigated (Padoa-Schioppa 2011).

NEURAL MECHANISMS FOR UPDATING VALUE FUNCTIONS

Temporal credit assignment and eligibility trace

Reward resulting from a particular action is often revealed after a substantial delay, and an animal might carry out several other actions before collecting the reward resulting from a previous action. Therefore, it can be challenging to associate an action and its corresponding outcome correctly, and this is referred to as the problem of temporal credit assignment (Sutton & Barto 1998). Not surprisingly, loss of the ability to link specific outcomes to corresponding choices interferes with the process of updating value functions appropriately. While normal animals can easily alter their preferences between two objects when the probabilities of getting rewards from the two objects are switched, humans, monkeys, and rats with lesions in the orbitofrontal cortex are impaired in such reversal learning tasks (Iversen & Mishkin 1970, Schoenbaum et al. 2002; Fellows & Farah 2003, Murray et al. 2007). These deficits may arise due to failures in temporal credit assignment. For example, during a probabilistic reversal learning task, in which the probabilities of rewards from different objects were dynamically and unpredictably changed, deficits produced by the lesions in the orbitofrontal cortex were due to erroneous associations between the choices and their outcomes (Walton et al. 2010).

In reinforcement learning theory, the problem of temporal credit assignment can be resolved in at least two different ways. First, a series of intermediate states can be introduced during the interval between an action and a reward, so that they can propagate the information about the reward to the value function of the correct action (Montague et al. 1996). This

basic temporal difference model was initially proposed to account for the reward prediction error signals conveyed by dopamine neurons, but was shown to be inconsistent with the actual temporal profiles of dopamine neuron signals (Pan et al. 2005). Another possibility is to utilize short-term memory signals related to the states or actions selected by the animal. Such memory signals are referred to as eligibility traces (Sutton & Barto 1998), and can facilitate action-outcome association even when the outcome is delayed. Therefore, eligibility traces can account for the temporally discontinuous shift in the phasic activity of dopamine neurons observed during classical conditioning (Pan et al. 2005). Signals related to the animal's previous choices have been observed in a number of brain areas, including the prefrontal cortex and posterior parietal cortex in monkeys (Barraclough et al. 2004, Seo & Lee 2009, Seo et al. 2009), as well as many regions in the rodent frontal cortex and striatum (Kim et al. 2007, 2009, Sul et al. 2010, 2011; Figure 3), and they might provide eligibility traces necessary to form associations between actions and their outcomes (Curtis & Lee 2010). In addition, neurons in many of these areas, including the orbitofrontal cortex, often encode specific conjunctions of chosen actions and their outcomes, for example, by increasing their activity when a positive outcome is obtained from a specific action (Barraclough et al. 2004, Seo & Lee 2009, Kim et al. 2009, Roesch et al. 2009, Sul et al. 2010, Abe & Lee 2011). Such action-outcome conjunction signals may also contribute to the resolution of the temporal credit assignment problem.

Integration of chosen value and reward prediction error

In model-free reinforcement learning, the value function for a chosen action is revised according to reward prediction error. Therefore, signals related to chosen value and reward prediction error must be combined in the activity of individual neurons involved in updating value functions. Signals related to reward prediction error were first identified in the midbrain dopamine neurons (Schultz 2006), but later found to exist in many other areas, including the lateral habenula (Matsumoto & Hikosaka 2007), globus pallidus (Hong & Hikosaka 2008), anterior cingulate cortex (Matsumoto et al. 2007, Seo & Lee 2007), orbitofrontal cortex (Sul et al. 2010), and striatum (Kim et al. 2009, Oyama et al. 2010). Thus, the extraction of reward prediction error signals might be gradual and implemented through a distributed network of multiple brain areas. Dopamine neurons might then play an important role in relaying these error signals to update the value functions broadly represented in different brain areas. Signals related to chosen values are also distributed in multiple brain areas, including the medial frontal cortex, orbitofrontal cortex, and striatum (Padoa-Schioppa & Assad 2006, Lau & Glimcher 2008, Kim et al. 2009, Sul et al. 2010, Cai et al. 2011). The areas in which signals related to chosen value and reward prediction error converge, such as the orbitofrontal cortex and striatum, might therefore play an important role in updating value functions (Kim et al. 2009, Sul et al. 2010).

It is often hypothesized that the primary site for updating and storing action value functions is at the synapses between axons from cortical neurons and dendrites of medium spiny neurons in the striatum (Reynolds et al. 2001, Hikosaka et al. 2006, Lo & Wang 2006, Hong & Hikosaka 2011). Signals related to reward prediction error arrive at these synapses via the terminals of dopamine neurons in the substantia nigra (Levey et al. 1993, Schultz 2006, Haber et al. 2000, Haber & Knutson 2010), and multiple types of dopamine receptors in the striatum can modulate the plasticity of corticostriatal synapses accordingo the relative timing of presynaptic vs. postsynaptic action potentials (Shen et al. 2008, Gerfen & Surmeier 2011). However, the nature of specific information stored by these synapses remains poorly understood. In addition, whether the corticostriatal circuit carries appropriate signals related to eligibility traces for chosen actions and other necessary state information at the right time needs to be tested in future studies. Given broad dopaminergic projections to various cortical

areas (Lewis et al. 2001), value functions might be updated in many of the same cortical areas encoding the value functions at the time of decision making.

Uncertainty and learning rate

The learning rate, which controls the speed of learning, must be adjusted according to uncertainty and volatility of the animal's environment. In natural environments, decision makers face many different types of uncertainty. When the probabilities of different outcomes are known, as when flipping a coin or during economic experiments, uncertainty about outcomes is referred to as risk (Kahneman & Tversky 1979) or expected uncertainty (Yu & Dayan 2005). In contrast, when the exact probabilities are unknown, this is referred to as ambiguity (Ellsberg 1961) or unexpected uncertainty (Yu & Dayan, 2005). Ambiguity or unexpected uncertainty is high in a volatile environment, in which the probabilities of different outcomes expected from a given action change frequently (Behrens et al. 2007), and this requires a large learning rate so that value functions can be modified quickly. By contrast, if the environment is largely known and stable, then the learning rate should be close to 0 so that value functions are not too easily altered by stochastic variability in the environment. Human learners can change their learning rates almost optimally when the rate of changes in reward probabilities for alternative actions is manipulated (Behrens et al. 2007). The level of volatility, and hence the learning rate, is reflected in the activity of the anterior cingulate cortex, suggesting that this region of the brain might be important for adjusting the learning rate according to the stability of the decision maker's environment (Behrens et al. 2007; Figure 4). Similarly, the brain areas known to increase their activity during decision making under ambiguity, such as the lateral prefrontal cortex, orbitofrontal cortex, and amygdala, might also be involved in optimizing the learning rate (Hsu et al. 2005, Huettel et al. 2006). Single-neuron recording studies have also implicated the orbitofrontal cortex in evaluating the amount of uncertainty in choice outcomes (Kepecs et al. 2008, O'Neill & Schultz 2010).

NEURAL SYSTEMS FOR MODEL-FREE VS. MODEL-BASED REINFORCEMENT LEARNING

Model-based value functions and reward prediction errors

During model-based reinforcement learning, decision makers utilize their knowledge of the environment to update the estimates of outcomes expected from different actions, even without actual reward or penalty. A wide range of algorithms can be used to implement model-based reinforcement learning. For example, decision makers may learn the configuration and dynamics of their environment separately from the values of outcomes at different locations (Tolman 1948). This enables the animal to re-discover an optimal path of travel quickly whenever the location of a desired item changes. Flexibly combining these two different types of information might rely on the prefrontal cortex (Daw et al. 2005, Pan et al. 2008, Gläscher et al. 2010) and hippocampus (Womelsdorf et al. 2010, Simon & Daw 2011). For example, activity in the human lateral prefrontal cortex increases when unexpected state transitions are observed, suggesting that this area is involved in learning the likelihood of state transitions (Gläscher et al. 2010). The hippocampus might play a role in providing the information about the layout of the environment and other contextual information necessary for updating the value functions. The integration of information about the behavioral context and current task demands encoded in these two areas, especially at the time of decision making, may rely on rhythmic synchronization of neural activity in the theta frequency range (Sirota et al. 2008, Benchenane et al. 2010, Hyman et al. 2010, Womelsdorf et al. 2010).

Whether and to what extent model-free and model-based forms of reinforcement learning are supported by the same brain areas remains an important area of research. Value functions estimated by model-free and model-based algorithms might be updated or represented separately in different brain areas (Daw et al. 2005), but they might be also combined to produce a unique estimate for the outcomes expected from chosen actions. For example, when decision makers are required to combine the information about reward history and social information, reliability of predictions based on these two different types of information is reflected separately in two different regions of the anterior cingulate cortex (Behrens et al. 2008). On the other hand, signals related to reward probability predicted by both types of information were found in the ventromedial prefrontal cortex (Behrens et al. 2008). Similarly, human ventral striatum might represent the chosen values and reward prediction errors regardless of how they are computed (Daw et al. 2011, Simon & Daw 2011). This is also consistent with the finding that reward prediction error signals encoded by the midbrain dopamine neurons, as well as the neurons in the globus pallidus and lateral habenula, are in accordance with both model-free and model-based reinforcement learning (Bromberg-Martin et al. 2010b).

Hypothetical outcomes and mental simulation

From the information that becomes available after completing chosen actions, decision makers can often deduce what alternative outcomes would have been possible from other actions. They can then use this information about hypothetical outcomes to update the action value functions for unchosen actions. In particular, the observed behaviors of other decision makers during social interaction are a rich source of information about such hypothetical outcomes (Camerer & Ho 1999, Camerer 2003, Lee et al. 2005, Lee 2008). Results from lesion and neuroimaging studies have demonstrated that the information about hypothetical or counterfactual outcomes might be processed in the same brain areas that are also involved in evaluating the actual outcomes of chosen actions, such as the prefrontal cortex (Camille et al. 2004, Coricelli et al. 2005, Boorman et al. 2011) and striatum (Lohrenz et al. 2007). The hippocampus might also be involved in simulating the possible outcomes of future actions (Hassabis & Maguire 2007, Schacter et al. 2007, Johnson & Redish 2007, Luhmann et al. 2008). Single-neuron recording studies have also shown that neurons in the dorsal anterior cingulate cortex respond similarly to actual and hypothetical outcomes (Hayden et al. 2009). More recent neurophysiological experiments in the dorsolateral prefrontal cortex and orbitofrontal cortex further revealed that neurons in these areas tend to encode actual and hypothetical outcomes for the same action, suggesting that they might provide an important substrate for updating the action value functions for chosen and unchosen actions simultaneously (Abe & Lee 2011; Figure 5).

CONCLUSIONS

In recent decades, reinforcement learning theory has become a central framework in the newly emerging areas of neuroeconomics and decision neuroscience. This is hardly surprising, because unlike abstract decisions analyzed in economic theories, biological organisms seldom receive complete information about the likelihoods of different outcomes expected from alternative actions. Instead, they face the challenge of learning how to predict the outcomes of their actions by trial and error, which is the essence of reinforcement learning.

The field of reinforcement learning theory has yielded many different algorithms, which provide neurobiologists with exciting opportunities to test whether they can successfully account for the actual behaviors of humans and animals and how different computational elements are implemented in the brain. In some cases, particular theoretical components closely correspond to specific brain structures, as in the case of reward prediction errors and

midbrain dopamine neurons. However, in general, a relatively well circumscribed computational step in a given algorithm is often implemented in multiple brain areas, and this relationship might change with the animal's experience and task demands. The challenge that lies ahead is therefore to understand whether and how the signals in different brain areas related to various components of reinforcement learning, such as action value functions and chosen value, make different contributions to the overall behaviors of the animal. An important example discussed in this article is the relationship between model-free and model-based reinforcement learning algorithms. Neural machinery of model-free reinforcement learning might be phylogenetically older, and for simple decision-making problems, it may be more robust. By contrast, for complex decision-making problems, model-based reinforcement learning algorithms can be more efficient, since it can avoid the need to re-learn appropriate stimulus-action associations repeatedly by exploiting the regularities in the environment and the current information about the animal's internal state. Failures in applying appropriate reinforcement learning algorithms can lead to a variety of maladaptive behaviors observed in different mental disorders. Therefore, it would be crucial for future studies to elucidate the mechanisms that allow the brain to coordinate different types of reinforcement learning and their individual elements.

Acknowledgments

We are grateful to Soyoun Kim, Jung Hoon Sul, and Hoseok Kim for their help with the illustrations, and Jeansok Kim, Matthew Kleinman, and Tim Vickery for helpful comments on the manuscript. The research of the authors was supported by the grants from the National Institute of Drug Abuse (DA024855 and DA029330 to D. L.) and the Korea Ministry of Education, Science and Technology (the Brain Research Center of the 21st Century Frontier Research Program, NRF grant 2011-0015618 and the Original Technology Research Program for Brain Science 2011-0019209 to M.W.J.).

LITERATURE CITED

Abe H, Lee D. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. Neuron. 2011; 70:731–741. [PubMed: 21609828]

Andersen RA, Essick GK, Siegel RM. Neurons of area 7 activated by both visual stimuli and oculomotor behavior. Exp. Brain Res. 1987; 67:316–322. [PubMed: 3622691]

Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998; 37:407–419. [PubMed: 9704982]

Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. 2004; 7:404–410. [PubMed: 15004564]

Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A. Probabilistic population codes for Bayesian decision making. Neuron. 2008; 60:1142–1152. [PubMed: 19109917]

Belova MA, Paton JJ, Salzman CD. Moment-to-moment tracking of state value in the amygdala. J. Neurosci. 2008; 48:10023–10030. [PubMed: 18829960]

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nat. Neurosci. 2007; 10:1214–1221. [PubMed: 17676057]

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. Associative learning of social value. Nature. 2008; 456:245–249. [PubMed: 19005555]

Benchenane K, Peyrache A, Khamassi M, Tierney PL, Gioanni Y, Battaglia FP, et al. Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. Neuron. 2010; 66:921–936. [PubMed: 20620877]

Boorman ED, Behrens TE, Rushworth MF. Counterfactual choicer and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biol. 2011; 9 e1001093.

Bornstein AM, Daw ND. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. Curr. Opin. Neurobiol. 2011; 21:374–380. [PubMed: 21429734]

Bromberg-Martin ES, Matsumoto M, Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron. 2010a; 68:815–834. [PubMed: 21144997]

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol. 2010b; 104:1068–1076. [PubMed: 20538770]

Bruce CJ, Goldberg ME. Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. 1985; 53:603–635. [PubMed: 3981231]

Cai X, Kim S, Lee D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron. 2011; 69:170–182. [PubMed: 21220107]

Camerer, CF. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton: Princeton Univ. Press; 2003.

Camerer C, Ho TH. Experience-weighted attraction learning in normal form games. Econometrica. 1999; 67:827–874.

Camille N, Coricelli G, Sallet J, Pradat-Diehl P, Duhamel JR, Sirigu A. The involvement of the orbitofrontal cortex in the existence of regret. Science. 2004; 304:1167–1170. [PubMed: 15155951]

Churchland MM, Yu BM, Ryu SI, Santhnam G, Shenoy KV. Neural variability in premotor cortex provides a signature of motor preparation. J. Neurosci. 2006; 26:3697–3712. [PubMed: 16597724]

Coe B, Tomihara K, Matsuzawa M, Hikosaka O. Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J. Neurosci. 2002; 22:5081–5090. [PubMed: 12077203]

Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ. Regret and its avoidence: a neuroimaging study of choice behavior. Nat. Neurosci. 2005; 8:1255–1262. [PubMed: 16116457]

Curtis CE, Lee D. Beyond working memory: the role of persistent activity in decision making. Trends Cogn. Sci. 2010; 14:216–222. [PubMed: 20381406]

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69:1204–1215. [PubMed: 21435563]

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 2005; 8:1704–1711. [PubMed: 16286932]

Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 2008; 18:185–196. [PubMed: 18708140]

Ding L, Gold JI. Neural correlates of perceptual decision making before, during, and after decision commitment in monkey frontal eye field. Cereb. Cortex. 2011 In press.

Dorris MC, Glimcher PW. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron. 2004; 44:365–378. [PubMed: 15473973]

Ellsberg D. Risk, ambiguity, and the Savage axioms. Q. J. Econ. 1961; 61:643–669.

Erev I, Roth AE. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. Am. Econ. Rev. 1998; 88:848–881.

Fellows LK, Farah MJ. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. Brain. 2003; 126:1830–1837. [PubMed: 12821528]

Feltovich R. Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. Econometrica. 2000; 68:605–641.

Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. 1989; 61:331–349.

Furman M, Wang X-J. Similarity effect and optimal control of multiple-choice decision making. Neuron. 2008; 60:1153–1168. [PubMed: 19109918]

Georgopoulos AP, Schwartz AB, Kettner RE. Neural population coding of movement direction. Science. 1986; 233:1416–1419. [PubMed: 3749885]

Gerfen CR, Surmeier DJ. Modulation of striatal projection systems by dopamine. Annu. Rev. Neurosci. 2011; 34:441–466. [PubMed: 21469956]

Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66:585–595. [PubMed: 20510862]

Glimcher PW, Sparks DL. Movement selection in advance of action in the superior colliculus. Nature. 1992; 355:542–545. [PubMed: 1741032]

Gnadt JW, Andersen RA. Memory related motor planning activity in posterior parietal cortex of macaque. Exp. Brain Res. 1988; 70:216–220. [PubMed: 3402565]

Haber SN, Fudge JL, McFarland R. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J. Neurosci. 2000; 20:2369–2382. [PubMed: 10704511]

Haber SN, Knutson B. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology. 2010; 35:4–26. [PubMed: 19812543]

Haggard P. Human volition: towards a neuroscience of will. Nat. Rev. Neurosci. 2008; 9:934–946. [PubMed: 19020512]

Hanes DP, Schall JD. Neural control of voluntary movement initiation. Science. 1996; 274:427–430. [PubMed: 8832893]

Hassabis D, Maguire EA. Deconstructing episodic memory with construction. Trends Cogn. Sci. 2007; 11:299–306. [PubMed: 17548229]

Hayden BY, Pearson JM, Platt ML. Fictive reward signals in the anterior cingulate cortex. Science. 2009; 324:948–950. [PubMed: 19443783]

Hikosaka O, Nakamura K, Nakahara H. Basal ganglia orient eyes to reward. J. Neurophysiol. 2006; 95:567–584. [PubMed: 16424448]

Hikoaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, et al. Parallel neural networks for learning sequential procedures. Trends Neurosci. 1999; 22:464–471. [PubMed: 10481194]

Hong S, Hikosaka O. The globus pallidus sends reward-related signals to the lateral habenula. Neuron. 2008; 60:720–729. [PubMed: 19038227]

Hong S, Hikosaka O. Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. Front. Behav. Neurosci. 2011; 5:15. [PubMed: 21472026]

Horwitz GD, Newsome WT. Target selection for saccadic eye movements: prelude activity in the superior colliculus during a direction-discrimination task. J. Neurophysiol. 2001; 86:2543–2558. [PubMed: 11698541]

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF. Neural systems responding to degrees of uncertainty in human decision-making. Science. 2005; 310:1680–1683. [PubMed: 16339445]

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML. Neural signatures of economic preferences for risk and ambiguity. Neuron. 2006; 49:765–775. [PubMed: 16504951]

Hyman JM, Zilli EA, Paley AM, Hasselmo ME. Working memory performance correlates with prefrontal-hippocampal theta interactions but not with prefrontal neuron firing rates. Front. Integr. Neurosci. 2010; 4:2. [PubMed: 20431726]

Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J. Neurosci. 2009; 29:9861–9874. [PubMed: 19657038]

Ito M, Doya K. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr. Opin. Neurobiol. 2011; 21:368–373. [PubMed: 21531544]

Iversen SD, Mishkin M. Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. Exp. Brain Res. 1970; 11:376–386. [PubMed: 4993199]

Johnson A, Redish AD. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. 2007; 27:12176–12189. [PubMed: 17989284]

Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica. 1979; 47:263–291.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008; 455:227–231. [PubMed: 18690210]

Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J. Neurosci. 2009; 29:14701–14712. [PubMed: 19940165]

Kim S, Hwang J, Lee D. Prefrontal coding of temporally discounted values during intertemporal choice. Neuron. 2008; 59:161–172. [PubMed: 18614037]

Kim S, Lee D. Prefrontal cortex and impulsive decision making. Biol. Psychiatry. 2011; 69:1140–1146. [PubMed: 20728878]

Kim Y, Huh N, Lee H, Baeg E, Lee D, Jung MW. Encoding of action history in the rat ventral striatum. J. Neurophysiol. 2007; 98:3548–3556. [PubMed: 17942629]

Knowlton BJ, Mangels JA, Squire LR. A neostriatal habit learning system in humans. Science. 1996; 273:1399–1402. [PubMed: 8703077]

Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nat. Neurosci. 2010; 13:1292–1298. [PubMed: 20835253]

Laming, DRJ. Information Theory of Choice-Reaction Times. London: Academic Press; 1968.

Lau B, Glimcher PW. Value representations in the primate striatum during matching behavior. Neuron. 2008; 58:451–463. [PubMed: 18466754]

Lee D. Game theory and neural basis of social decision making. Nat. Neurosci. 2008; 11:404–409. [PubMed: 18368047]

Lee D, Conroy ML, McGreevy BP, Barraclough DJ. Reinforcement learning and decision making in monkeys during a competitive game. Cogn. Brain Res. 2004; 22:45–58.

Lee D, McGreevy BP, Barraclough DJ. Learning and decision making in monkeys during a rock-paper-scissors game. Cogn. Brain Res. 2005; 25:416–430.

Levey AI, Hersch SM, Rye DB, Sunahara RK, Niznik HB, et al. Localization of D_1 and D_2 dopamine receptors in brain with subtype-specific antiboties. Proc. Natl. Acad. Sci. USA. 1993; 90:8861–8865. [PubMed: 8415621]

Lewis DA, Melchitzky DS, Sesack SR, Whitehead RE, Auh S, Sampson A. Dopamine transporter immunoreactivity in monkey cerebral cortex: regional, laminar, and ultrastructural localization. J. Compar. Neurol. 2001; 432:119–136.

Litt A, Plassmann H, Shiv B, Rangel A. Dissociating valuation and saliency signals during decision making. Cereb. Cortex. 2011; 21:95–102. [PubMed: 20444840]

Lo C-C, Wang X-J. Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat. Neurosci. 2006; 9:956–963. [PubMed: 16767089]

Lohrenz T, McCabe K, Camerer CF, Montague PR. Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. U.S.A. 2007; 104:9493–9498. [PubMed: 17519340]

Louie K, Glimcher PW. Separating value from choice: delay discounting activity in the lateral intrapreital area. J. Neurosci. 2010; 30:5498–5507. [PubMed: 20410103]

Luhmann CC, Chun MM, Yi D-J, Lee D, Wang X-J. Neural dissociation of delay and uncertainty in intertemporal choice. J. Neurosci. 2008; 28:14459–14466. [PubMed: 19118180]

Matsumoto M, Hikosaka O. Lateral habenula as a source of negative reward signals in dopamine neurons. Nature. 2007; 447:1111–1115. [PubMed: 17522629]

Matsumoto M, Matsumoto K, Abe H, Tanaka K. Medial prefrontal cell activity signaling prediction errors of action values. Nat. Neurosci. 2007; 10:647–656. [PubMed: 17450137]

Maunsell JHR. Neuronal representations of cognitive state: reward or attention? Trends Cogn. Sci. 2004; 8:261–265. [PubMed: 15165551]

Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 1996; 16:1936–1947. [PubMed: 8774460]

Mookherjee D, Sopher B. Learning behavior in an experimental matching pennies game. Games Econ. Behav. 1994; 7:62–91.

Mookherjee D, Sopher B. Learning and decision costs in experimental constant sum games. Games Econ. Behav. 1997; 19:97–132.

Murray EA, O'Doherty JP, Schoenbaum G. What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. J. Neurosci. 2007; 27:8166–8169. [PubMed: 17670960]

Nachev P, Kennard C, Husain M. Functional role of the supplementary and pre-supplementary motor areas. Nat. Rev. Neurosci. 2008; 9:856–869. [PubMed: 18843271]

Okano K, Tanji J. Neuronal activities in the primate motor fields of the agranular frontal cortex preceding visually triggered and self-paced movement. Exp. Brain Res. 1987; 66:155–166. [PubMed: 3582529]

O'Neill M, Schultz W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. Neuron. 2010; 68:789–800. [PubMed: 21092866]

Oyama K, Hernádi I, Iijima T, Tsutsui K-I. Reward prediction error coding in dorsal striatal neurons. J. Neurosci. 2010; 30:11447–11457. [PubMed: 20739566]

Padoa-Schioppa C. Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. 2011; 34:333–359. [PubMed: 21456961]

Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature. 2006; 441:223–226. [PubMed: 16633341]

Pan W-X, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci. 2005; 25:6235–6242. [PubMed: 15987953]

Pan X, Sawa K, Tsuda I, Tsukada M, Sakagami M. Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. Nat. Neurosci. 2008; 11:703–712. [PubMed: 18500338]

Pastor-Bernier A, Cisek P. Neural correlates of biased competition in premotor cortex. J. Neurosci. 2011; 31:7083–7088. [PubMed: 21562270]

Platt ML, Glimcher PW. Neural correlates of decision variables in parietal cortex. Nature. 1999; 400:233–238. [PubMed: 10421364]

Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1978; 4:515–526.

Reynolds JNJ, Hyland BI, Wickens JR. A cellular mechanism of reward-related learning. Nature. 2001; 413:67–70. [PubMed: 11544526]

Roesch MR, Singh T, Brown PL, Mullins SE, Schoenbaum G. Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. J. Neurosci. 2009; 29:13365–13376. [PubMed: 19846724]

Roese, NJ.; Olson, JM. What might have been: the social psychology of counterfactual thinking. New York: Psychology Press; 1995.

Roitman JD, Shadlen MN. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. 2002; 22:9475–9489. [PubMed: 12417672]

Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005; 310:1337–1340. [PubMed: 16311337]

Schacter DL, Addis DR, Buckner RL. Remembering the past to imagine the future: the prospective brain. Nat. Rev. Neurosci. 2007; 8:657–661. [PubMed: 17700624]

Schall JD. Neuronal activity related to visually guided saccadic eye movements in the supplementary motor area of rhesus monkeys. J. Neurophysiol. 1991; 66:530–558. [PubMed: 1774585]

Schiller PH, Stryker M. Single-unit recording and stimulation in superior colliculus of the alert rhesus monkey. J. Neurophysiol. 1972; 35:915–924. [PubMed: 4631839]

Schlag J, Schlag-Rey M. Evidence for a supplementary eye field. J. Neurophysiol. 1987; 57:179–200. [PubMed: 3559671]

Schoenbaum G, Nugent SL, Saddoris MP, Setlow B. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. Neuroreport. 2002; 13:885–890. [PubMed: 11997707]

Schultz W. Behavioral theories and the neurophysiology of reward. Annu. Rev. Psychol. 2006; 57:87–115. [PubMed: 16318590]

Schultz W, Tremblay L, Hollerman JR. Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb. Cortex. 2000; 10:272–284. [PubMed: 10731222]

Seo H, Barraclough DJ, Lee D. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J. Neurosci. 2009; 29:7278–7289. [PubMed: 19494150]

Seo H, Lee D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J. Neurosci. 2007; 27:8366–8377. [PubMed: 17670983]

Seo H, Lee D. Cortical mechanisms for reinforcement learning in competitive games. Phil. Trans. R. Soc. B. 2008; 363:3845–3857. [PubMed: 18829430]

Seo H, Lee D. Behavioral and neural changes after gains and losses of conditioned reinforcers. J. Neurosci. 2009; 29:3627–3641. [PubMed: 19295166]

Serences JT. Value-based modulations in human visual cortex. Neuron. 2008; 60:1169–1181. [PubMed: 19109919]

Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex of the rhesus monkey. J. Neurophysiol. 2001; 86:1916–1936. [PubMed: 11600651]

Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 2008; 321:848–851. [PubMed: 18687967]

Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci. 2011; 31:5526–5539. [PubMed: 21471389]

Shuler MG, Bear MF. Reward timing in the primary visual cortex. Science. 2006; 311:1606–1609. [PubMed: 16543459]

Sirota A, Montgomery S, Fujisawa S, Isomura Y, Zugaro M, Buzsáki G. Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. Neuron. 2008; 60:683–697. [PubMed: 19038224]

Smith PL, Ratcliff R. Psychology and neurobiology of simple decisions. Trends Neurosci. 2004; 27:161–168. [PubMed: 15036882]

Smyrnis N, Taira M, Ashe J, Georgopoulos AP. Motor cortical activity in a memorized delay task. Exp. Brain Res. 1992; 92:139–151. [PubMed: 1486948]

So NY, Stuphorn V. Supplementary eye field encodes option and action value for saccades with variable reward. J. Neurophysiol. 2010; 104:2634–2653. [PubMed: 20739596]

Softky WR, Koch C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J. Neurosci. 1993; 13:334–350. [PubMed: 8423479]

Soltani A, Lee D, Wang X-J. Neural mechanism for stochastic behaviour during a competitive game. Neural Netw. 2006; 19:1075–1090. [PubMed: 17015181]

Soltani A, Wang X-J. A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J. Neurosci. 2006; 26:3731–3744. [PubMed: 16597727]

Sohn J-W, Lee D. Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. J. Neurosci. 2007; 27:13655–13666. [PubMed: 18077677]

Soon CS, Brass M, Heinze H-J, Haynes J-D. Unconscious determinants of free decisions in the human brain. Nat. Neurosci. 2008; 11:543–545. [PubMed: 18408715]

Sugrue LP, Corrado GS, Newsome WT. Matching behavior and the representation of value in the parietal cortex. Science. 2004; 304:1782–1787. [PubMed: 15205529]

Sul JH, Kim H, Huh N, Lee D, Jung MW. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron. 2010; 66:449–460. [PubMed: 20471357]

Sul JH, Jo S, Lee D, Jung MW. Role of rodent secondary motor cortex in value-based action selection. Nat. Neurosci. 2011 In press.

Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.

Tanji J, Kurata K. Contrasting neuronal activity in supplementary and precentral motor cortex of monkeys. I. Responses to instructions determining motor responses to forthcoming signals of different modalities. J. Neurophysiol. 1985; 53:129–141. [PubMed: 3973654]

Tolman EC. Cognitive maps in rats and men. Psychol. Rev. 1948; 55:189–208. [PubMed: 18870876]

Tremblay L, Schultz W. Relative reward preference in primate orbitofrontal cortex. Nature. 1999; 398:704–708. [PubMed: 10227292]

Usher M, McClelland J. On the time course of perceptual choice: the leaky, competing accumulator model. Psychol. Rev. 2001; 108:550–592. [PubMed: 11488378]

van der Meer MAA, Redish AD. Ventral striatum: a critical look at models of learning and evaluation. Curr. Opin. Neurobiol. 2011; 21:387–392. [PubMed: 21420853]

Vickery TJ, Chun MM, Lee D. Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron. 2011 In press.

von Neumann, J.; Morgenstern, O. Theory of Games and Economic Behavior. Princeton: Princeton Univ. Press; 1944.

Wallis JD, Kennerley SW. Heterogeneous reward signals in prefrontal cortex. Curr. Opin. Neurobiol. 2010; 20:191–198. [PubMed: 20303739]

Wallis JD, Miller EK. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. Eur. J. Neurosci. 2003; 8:2069–2081. [PubMed: 14622240]

Walton ME, Behrens TE, Buckley MJ, Rudebeck PH, Rushworth MF. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron. 2010; 65:927–939. [PubMed: 20346766]

Wang X-J. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 2001; 24:455–463. [PubMed: 11476885]

Wang X-J. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 2002; 36:955–968. [PubMed: 12467598]

Wang X-J. Decision making in recurrent neuronal circuits. Neuron. 2008; 60:215–234. [PubMed: 18957215]

Weinrich M, Wise SP. The premotor cortex of the monkey. J. Neurosci. 1982; 2:1329–1345. [PubMed: 7119878]

Womelsdorf T, Vinck M, Leung LS, Everling S. Selective theta-synchronization of choice-relevant information subserves goal-directed behavior. Front. Hum. Neurosci. 2010; 4:210. [PubMed: 21119780]

Wunderlich K, Rangel A, O'Doherty JP. Neural computations underlying action-based decision making in the human brain. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:17199–17204. [PubMed: 19805082]

Wurtz RH, Goldberg ME. Activity of superior colliculus in behaving monkey. 3. Cells discharging before eye movements. J. Neurophysiol. 1972; 35:575–586. [PubMed: 4624741]

Yin HH, Knowlton BJ. The role of the basal ganglia in habit formation. Nat. Rev. Neurosci. 2006; 7:464–476. [PubMed: 16715055]

Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005; 46:681–692. [PubMed: 15944135]

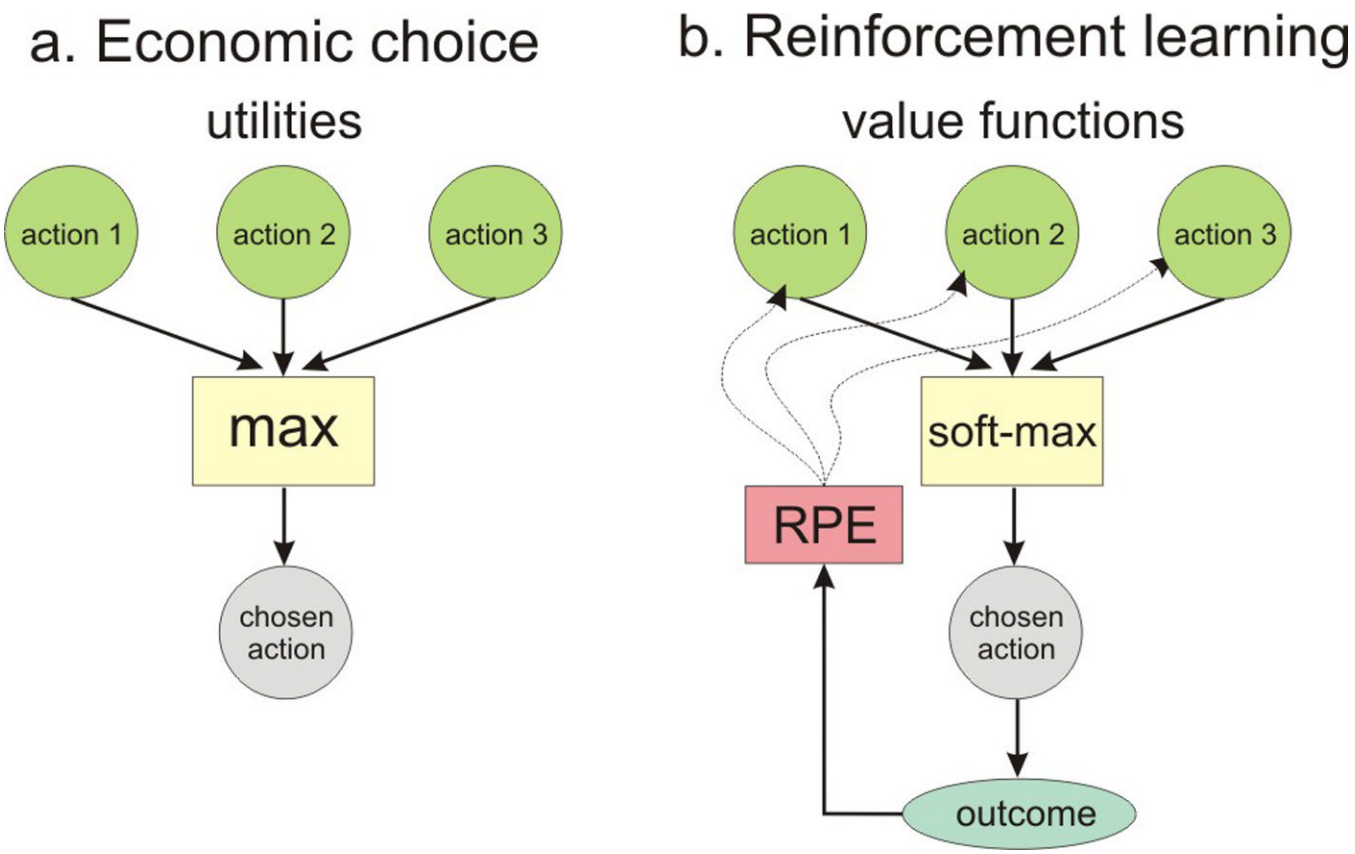

Figure 1. Economic and reinforcement learning theories of decision making

(a) In economic theories, decision making corresponds to selecting an action with the maximum utility. (b) In reinforcement learning, actions are chosen probabilistically (i.e., softmax) on the basis of their value functions. In addition, value functions are updated on the basis of the outcome (reward or penalty) resulting from the action chosen by the animal. RPE, reward prediction error.

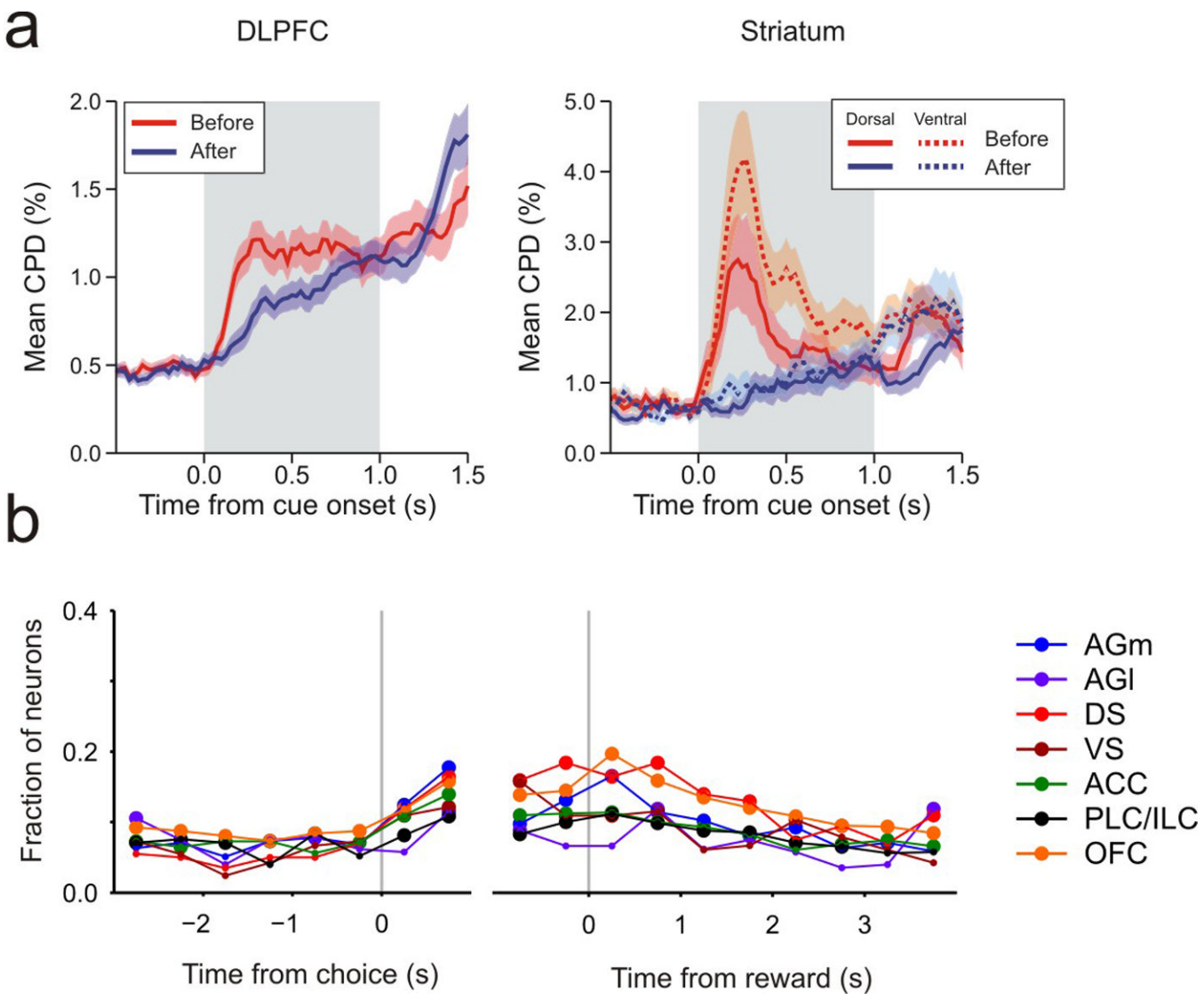

Figure 2. Time course of signals related to different state value functions during decision making

(a) Signals related to the state value functions before (red) and after (blue) decision making in the dorsolateral prefrontal cortex (DLPFC; Kim et al. 2008, Kim & Lee 2011) and striatum (Cai et al. 2011) during an intertemporal choice task. These two state value functions correspond to the average of the action value functions for two options and the chosen value, respectively. During these studies, monkeys chose between a small immediate reward and a large delayed reward, and the magnitude of neural signals related to different value functions were estimated by the coefficient of partial determination (CPD). Lines correspond to the mean CPD for all the neurons recorded in each brain area with the shaded area corresponding to the standard error of the mean. (b) Proportion of neurons carrying chosen value signals in the rodent lateral (AGl) and medial (AGm) agranular cortex, corresponding to the primary and secondary motor cortex, respectively, dorsal (DS) and ventral (VS) striatum, anterior cingulate cortex (ACC), prelimbic (PLC)/infralimbic (ILC) cortex, and orbitofrontal cortex (OFC). During these studies (Kim et al. 2009, Sul et al. 2010, 2011), the rats performed a dynamic foraging task. Large symbols indicate that the proportions are significantly ($p<0.05$) above the chance level.

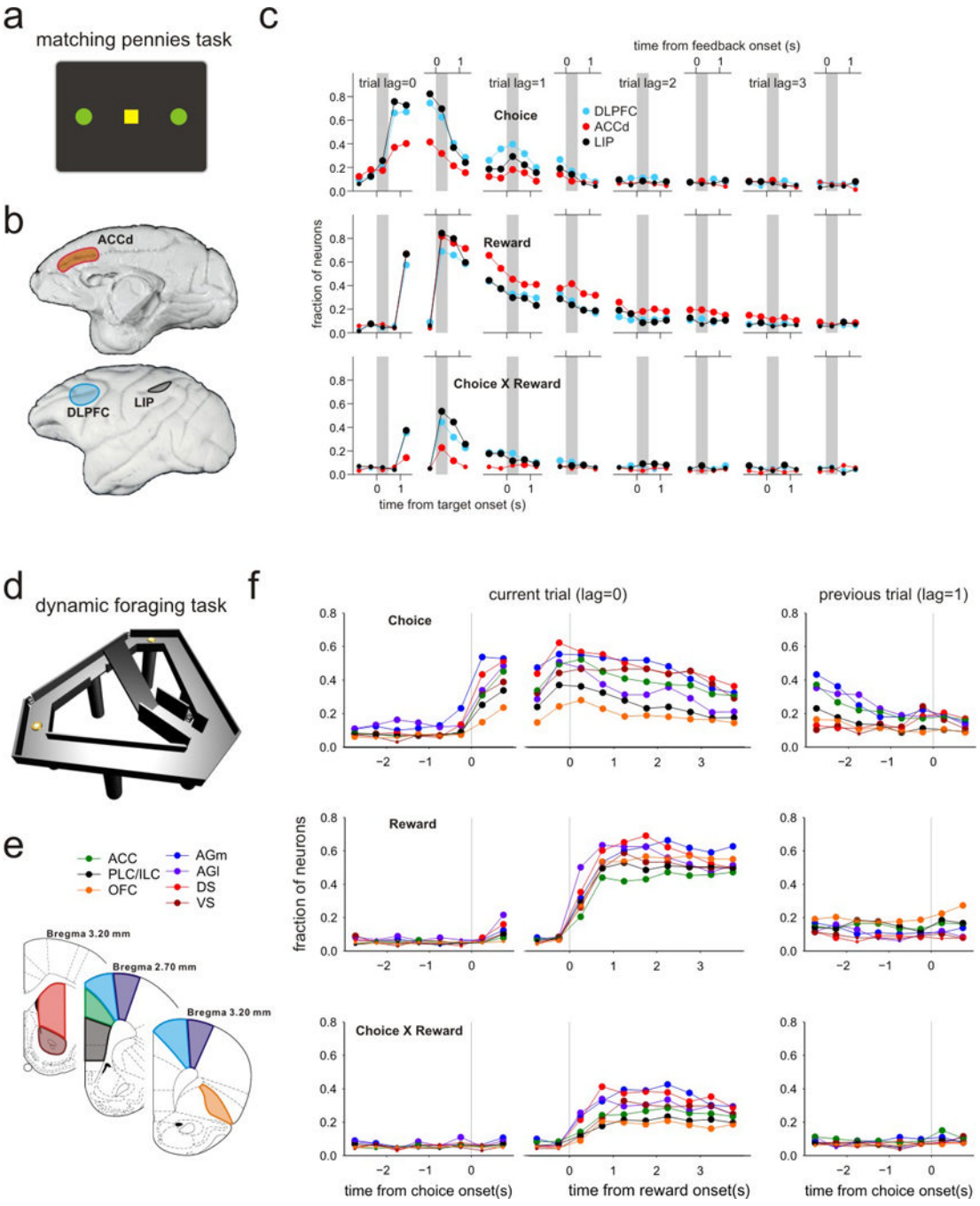

Figure 3. Time course of signals related to the animal's choice, its outcome, and action-outcome conjunction in multiple brain areas of primates and rodents

(a) Spatial layout of the choice targets during a matching pennies task used in single-neuron recording experiments in monkeys. (b) Brain regions tested during the studies on monkeys (Barraclough et al. 2004, Seo & Lee 2007, Seo et al. 2009). ACCd, dorsal anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex; LIP, lateral intraparietal cortex. (c) Fraction of neurons significantly modulating their activity according to the animal's choice (top), its outcome (middle), and choice-outcome conjunction (bottom) during the current (trial lag =0) and 3 previous trials (trial lags =1~3). (d) Modified T-maze used in a rodent dynamic foraging task. (e) Anatomical areas tested in single-neuron recording experiments in rodents

(Kim et al. 2009, Sul et al. 2010, 2011). Same abbreviations as in Figure 2b. (f) Fraction of neurons significantly modulating their activity according to the animal's choice (top), its outcome (middle), and choice-outcome conjunction (bottom) during the current (lag =0) and previous trials (lag =1). Large symbols indicate that the proportions are significantly ($p<0.05$) above the chance level.

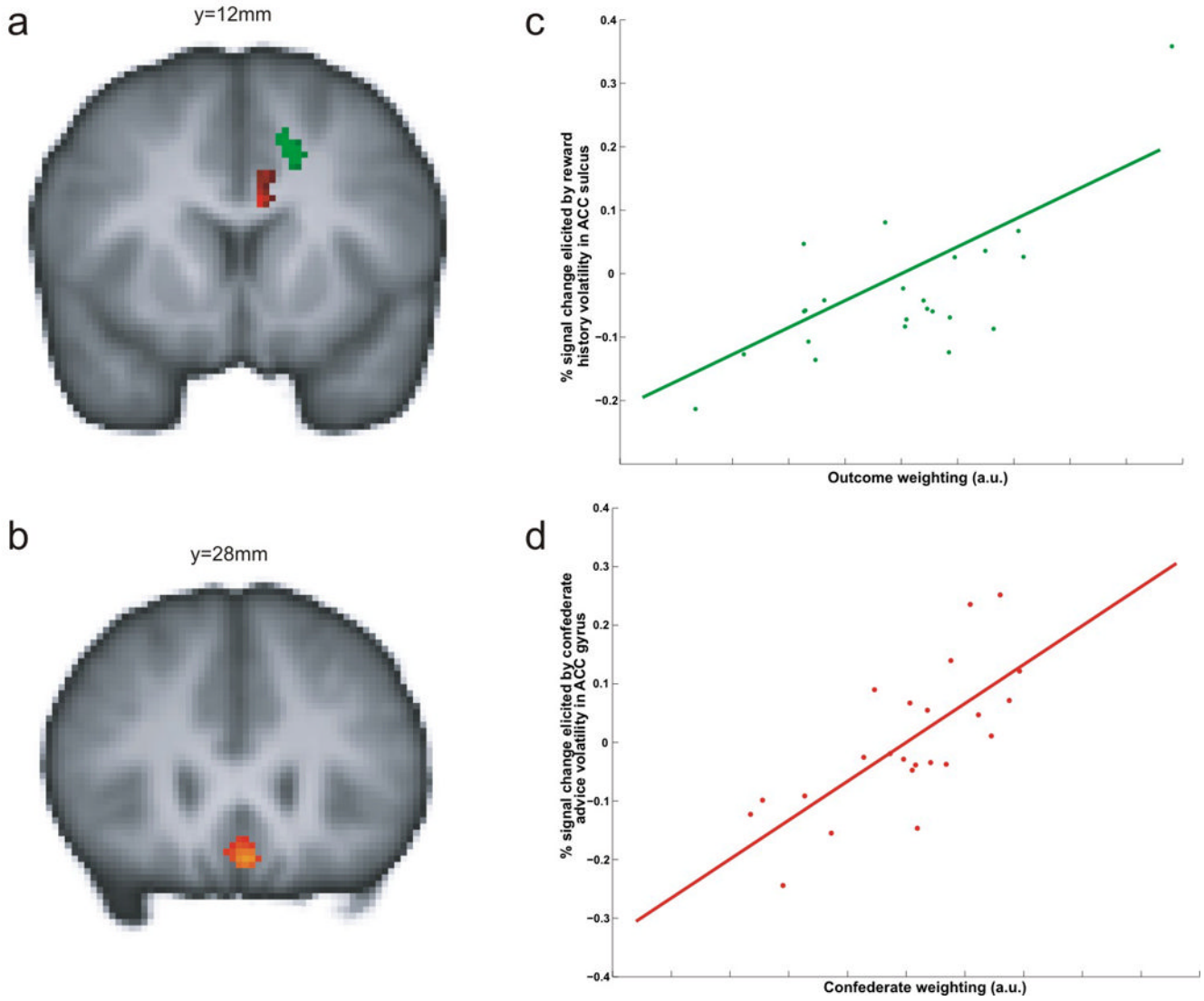

Figure 4. Areas in the human brain involved in updating model-free and model-based value functions (Behrens et al. 2008)

(a) Regions in which the activity is correlated with the volatility in estimating the value functions based on reward history (green) and social information (red). (b) Activity in the ventromedital prefrontal cortex was correlated with the value functions regardless of whether they were estimated from reward history or social information. (c) Subjects more strongly influenced by reward history (ordinate) tended to show greater signal change in the anterior cingulate cortex in association with reward history (abscissa; green region in a). (d) Subjects more strongly influenced by social information (ordinate) showed greater signal changes in the anterior cingulate cortex in association with social information (abnscissa; red region in a).

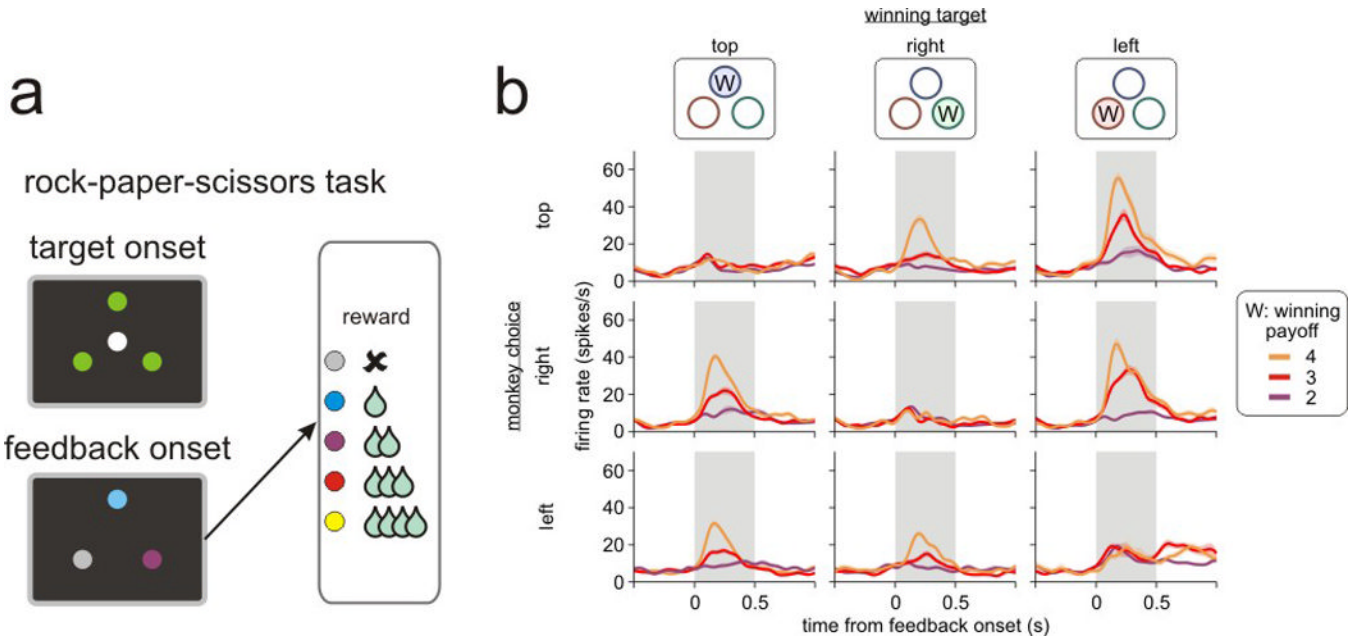

Figure 5. Neuronal activity related to hypothetical outcomes in the primate orbitofrontal cortex

(a) Rock-paper-sciossors task used for single-neuron recording studies in monkeys (Abe & Lee 2011). (b) An example neuron recorded in the orbitofrontal cortex that modulated its activity according to the magnitude of reward that was available from the unchosen winning target (indicated by 'W' in the top panels). The spike density function of this neuron was estimated separately according to the position of the winning target (columns), the position of the target chosen by the animal (rows), and the magnitude of the reward available from the winning target (colors).