

Multiple Forms of Value Learning and the Function of Dopamine

Bernard W. Balleine¹, Nathaniel D. Daw² and John O'Doherty³

1. Department of Psychology and the Brain Research Institute, UCLA
2. Center for Neural Science and Department of Psychology, New York University
3. Computation and Neural Systems Program, California Institute of Technology.

Correspondence

Bernard Balleine
Department of Psychology, UCLA
BOX 951563
Los Angeles, CA 90095-1563

Tel: 310.825.7560
Fax: 310.206.5895
Email: balleine@psych.ucla.edu

1. Introduction

Among the key findings in the behavioral psychology and systems neuroscience of decision-making is that the same behavior – for instance, a rat's lever press – can arise from multiple influences that are both neurally and psychologically dissociable. In this article, we review recent theory and research about these decision mechanisms and their neural bases, focusing particularly on the idea that they embody distinct evaluative or motivational processes: that is, different sorts of 'value'.

Although there is an extensive literature linking the control of executive functions to the prefrontal cortex (Goldman-Rakic, 1995; Fuster, 2000), more recent studies suggest that these functions depend on reward-related circuitry linking cortex with the striatum (Chang et al., 2002; Lauwereyns et al., 2002; Tanaka et al., 2006). Evidence from a range of species suggests that discrete cortico-striatal networks control two functionally distinct decision processes. The first involves actions that are more flexible or *goal-directed*, sensitive to reward-related feedback and that involve regions of association cortices particularly medial, orbitomedial, premotor and anterior cingulate cortices together with their efferent targets in caudate/dorsomedial striatum (Haruno and Kawato, 2006; Levy and Dubois, 2006). The second involves actions that are relatively automatic or *habitual* and that depend on sensorimotor cortices and dorsolateral striatum/putamen (Jog et al., 1999; Poldrack et al., 2001). These two sorts of processes have been argued to employ different learning rules (Dickinson, 1994), different forms of plasticity (Partridge et al., 2000; Smith et al., 2001), and different computational principles (Dayan and Balleine, 2002; Daw et al., 2005). Furthermore, degeneration of these two cortico-striatal circuits has been argued to result in two distinct forms of pathology such as Huntington's disease, obsessive compulsive disorder and Tourette's syndrome on the one hand (Robinson et al., 1995; Bloch et al., 2005; Hodges et al., 2006) and Parkinson's disease and multiple system atrophy on the other (Antonini et al., 2001; Seppi et al., 2006).

Interestingly, these distinct mechanisms of action control appear to be related to distinct motivational or evaluative processes. Goal-directed actions are so named

because they are sensitive to feedback about what we shall call goal values – that is, the rewarding values of the particular outcomes (for instance, the specific sort of food delivered for a lever press). Further demonstrating their goal sensitivity, such actions are also sensitive to predictive cues that signal the likelihood of those outcomes independent of any actions—referred to here as Pavlovian values. In contrast, habitual actions have long been argued to arise from a generalized propensity for emitting a particular response, which is acquired from and grounded in a history of reinforcement but divorced from any representation of the specific reinforcing goal. We refer to these generalized action propensities as habit values.

This chapter will review the behavioral evidence for these distinct value functions, computational models that relate these functions to distinct aspects of adaptive behavior, and the neural systems that mediate these processes. Of particular interest, will be the role of the midbrain dopamine system, and the extent to which it is involved, separately or integratively, in these valuation processes.

2. Reward, prediction and reinforcement

A classical economic notion of choice assumes that people evaluate options according to their expected utilities. Human and animal actions, however, are not nearly so unitary.

Psychologists (see Box) have long attempted to distinguish bona fide decisions from other behaviors, such as reflexes, that might only appear to be choice-like. On one view, for an action to qualify as truly volitional or “goal-directed,” it should depend on two factors, which echo the probability and utility of an outcome from the standard formula for expected utility. First is knowledge of the contingency between the action and some outcome, and second, the valuation of that outcome as a desirable goal.

As detailed below, these two criteria have been operationalized into behavioral tests. These tests reveal a clear dissociation between behaviors that are demonstrably goal-directed in this sense, and a second class of “habitual” behaviors, which pass neither test. As sketched more mathematically in Section 3, these behaviors can be

viewed as arising from different processes for evaluating an action: either through something like an explicit computation of expected utility, or a shortcut that simply assumes that previously reinforced behaviors will be valuable.

We begin by describing evidence for yet a third class of behaviors, and of underlying valuations, which also fail to satisfy the criteria for goal-directed actions. These are the conditioned reflexes studied by Pavlov (1927).

2.1. Reflex versus action

Although most aspects of our behavioral repertoire can be described with respect to attaining some goal or other, many of these activities are actually simple reflexes that are elicited without deliberation by environmental cues. As discussed above, a critical distinction between reflexive responses and goal-directed actions is that the latter are controlled by knowledge of their relationship to their consequences whereas the former are not. Classic examples of reflexive responses that have a misleading veneer of choice-like goal-directedness about them can be found in simple conditioning situations, such as that made popular by Pavlov (1927). He studied salivary responses to food delivery in dogs and the conditioning of those reflexes produced by pairing a neutral stimulus (such as a tone) with food.

The key feature of this sort of task is that food has a contingent (“Pavlovian”) relationship with the stimulus but its delivery is not contingent on any action the animal takes. The fact that the stimulus often acquires the ability to elicit anticipatory salivation in this situation is usually thought to reflect the transfer of control over the salivary reflex from the food to the stimulus based on the Pavlovian association between stimulus and food (Figure 3A).

From a decision-making perspective, however, it is possible to argue that, in fact, in this situation dogs control their salivation and decide to produce this response, perhaps to facilitate digestion or to improve the taste of food. To assess this explanation of conditioned salivation, Sheffield (1965) arranged a standard Pavlovian conditioning experiment in which he paired a tone with food delivery, but with a twist: if the dog salivated during the tone, then the food was not delivered on that trial. This arrangement

maintains a Pavlovian relationship between the tone and food but abolishes any positive relationship between salivation and food. Sheffield reasoned that, if the salivation was an action controlled by its relationship to food then arranging that salivation omitted food delivery should ensure that the dogs stop salivating; indeed having never had the opportunity to learn that salivating improved the reward value of the food by enhancing its flavor or improving its ingestion, they should never acquire salivation to the tone at all. Sheffield found that it was clearly the Pavlovian tone-food relationship that controlled salivary performance; during the course of over 800 tone-food pairings the dogs acquired and maintained salivation to the tone even though this resulted in them losing most of the food they could have obtained by withholding their salivary response.

Although salivation might in principle be the exception, numerous other studies over the last forty years have established, in a range of species including humans (Pithers, 1985) and other animals (Williams and Williams, 1969; Holland, 1979), that a large variety of conditioned responses are maintained by the relationship between the Pavlovian predictive cue and the outcome (food, shock etc) with which it is associated, rather than by the relationship between the response and the outcome.

2.2. Pavlovian values versus goal values

Though Pavlovian responses do not satisfy the criteria for goal-directed action, the same is not true of behaviors acquired in other conditioning preparations, notably instrumental conditioning (Thorndike, 1911). Here, animals readily learn to produce new and often quite arbitrary responses (such as lever presses) to gain access to food, which unlike in Pavlovian conditioning is delivered contingent on the action. In contrast to the salivation response studied by Sheffield, putting these responses on an omission contingency, in which responding leads to the omission of an otherwise freely delivered food, rapidly reduces their performance (Davis and Bitterman, 1971; Dickinson et al., 1998). Furthermore, numerous studies have demonstrated the exquisite sensitivity of the performance of instrumental actions to changes in the net probability of outcome delivery given the action (i.e. the difference between the probability of food given a response and the probability of that food given no response). These changes can be

highly selective; for instance, in situations in which different actions lead to different outcomes, degrading a particular action-outcome contingency by delivering its associated outcome non-contingently often has no effect on the performance of other actions (Colwill and Rescorla, 1986; Dickinson and Mulatero, 1989; Balleine and Dickinson, 1998). Like the omission test, this contingency degradation test exercises the first of the criteria for goal-directed action discussed above, that it be based on knowledge of the action-outcome contingency (Figure 3A).

Instrumental responses are sensitive not only to the action-outcome contingency, but also to the second criterion for a goal-directed action: the value of the goal. Again, experiments in rats have provided some of the clearest demonstrations of this effect. For example, in a number of studies, illustrated in Figure 1, hungry rats have been trained to carry out two actions (presses of different levers, or a lever press and a chain pull), with one response earning, say, food pellets and the other a sucrose solution. After this training, the desirability of one of the two outcomes is reduced, either by specific satiety (in which the rats are allowed to eat a large quantity of one outcome) or, in other studies, by taste aversion learning (in which consumption of one outcome is followed by drug-induced illness; see Figure 1). Both treatments reduce the consumption of a specific food relative to other foods. After the devaluation treatment, the rats are given a choice test between the two actions in *extinction* (in the absence of the delivery of either food outcome). If – as expected for a goal-directed behavior – performance of the actions is maintained by the values of their respective outcomes (their *goal values*), then devaluation of one outcome should reduce subsequent performance of its associated action in the extinction test, relative to the other action. This is exactly the result that is commonly reported; devaluing one of the food rewards selectively and profoundly reduces the performance of the action that in training delivered that outcome, compared to the action trained with the non-devalued outcome (Dickinson and Balleine, 1994; Balleine, 2001).

Instrumentally trained actions, unlike Pavlovian conditioned reflexes, therefore satisfy both criteria for goal-directed decisions. However, this should not be taken to indicate that Pavlovian learning has no effect on decision-making. To the contrary,

instrumental behaviors are also potentially affected by Pavlovian cue-outcome associations, as demonstrated in *Pavlovian-instrumental transfer* studies. These show that choice between instrumental actions, which as reviewed above is sensitive to the goal value and contingency, can also be substantially modified by the presentation of reward-related Pavlovian cues.

For example, in a recent study, illustrated in Figure 2, Corbit and Balleine (2005) exposed hungry rats to Pavlovian training in which two distinct auditory cues, a tone and white noise, were used to predict the delivery of two different but equally valued foods, grain food pellets and a sucrose solution. After this phase, the rats were trained to push two levers, one paired with the grain pellets and the other with the sucrose solution.

In short, the first phase trains rats on a Pavlovian contingency (cues with outcomes, and no leverpressing) and the second phase an instrumental contingency (leverpressing for outcomes, without the tone or noise). The third phase tests the combined effect of both contingencies, by giving the rats a choice test in which the rats were allowed freely to choose between the two levers during presentations of the tone and noise stimuli. During this test, neither the grain pellet nor the sucrose outcomes were delivered, so any performance could only be attributable to the previous learning. Although the rats' performance on the two levers was the same in the absence of the auditory stimuli, presentation of the tone was found to increase the rats' choice of the lever that, in training, had delivered the same outcome as the tone whereas presentation of the noise was found to increase the rats' choice of the other lever (see Figure 2). Numerous demonstrations of this effect have been reported, confirming that choice between actions is determined not just by goal values but by the specific predictions, what we refer to here as the *Pavlovian values*, provided by Pavlovian, reward-related cues (Colwill and Rescorla, 1988; Colwill and Motzkin, 1994; Corbit et al., 2001; Corbit and Balleine, 2003).

Note that from a pure utility-based decision-making perspective, this effect is somewhat puzzling, because in training the Pavlovian cues predict outcome delivery not contingent on any action, and it is therefore unclear why their predictions should be relevant to the valuation of actions. One clue comes from an experiment by Delamater

(1995), who trained a Pavlovian cue (e.g. noise predicting sucrose), but then selectively degraded its contingent predictive relationship by presenting the sucrose both during the noise and during a period when the noise was not presented. The degraded stimulus no longer showed Pavlovian-instrumental transfer; that is, its presentation failed to produce a selective increase in pressing a lever that had previously been paired with sucrose. This result is important in demonstrating that it is not just the pairing of a stimulus with an outcome that mediates its effects on choice; it is the *information* that the stimulus provides as a net predictor of a specific rewarding event. That the effect of a Pavlovian cue on instrumental choices depends on its predictive validity suggests that animals interpret it as providing information regarding the availability or likelihood of achieving a specific goal.

We have reviewed evidence that instrumental choice is biased both by goal values, as demonstrated in outcome devaluation studies, and also by Pavlovian cue-outcome values in Pavlovian-instrumental transfer. One question is whether these two effects really reflect distinct evaluative processes or if they instead are somehow actually mediated by the same mechanism. This might be true, for instance, if sensory cues related to the actions themselves play a role analogous to the lights and tones in the Pavlovian case. Although views of this sort have been formalized in a number of versions (Rescorla and Solomon, 1967); (Trapold and Overmier, 1972), there is now considerable evidence confirming that goal values and Pavlovian values are mediated by distinct processes (Balleine and Ostlund, 2007). For instance, the strength of the two effects – the effect of Pavlovian cues in transfer experiments and the effect of outcome devaluation – can be independently, differentially modulated by several behavioral manipulations, and one can be observed in circumstances when the other is not present (Corbit and Balleine, 2003; Holland (2004).

2.3. Goal values versus habit values

Importantly, only some instrumentally trained actions, such as newly acquired lever presses, demonstrate the two defining features of goal-directed actions in the manner detailed above. Other instrumental behaviors, such as the same responses following

overtraining, can become more automatic, involuntary or impulsive and fail to satisfy these tests; i.e., they become *habits* (Dickinson, 1994). There has been considerable interest in habits as a putative counterpart in normal function to the loss of behavioral control in various neurodegenerative conditions and in drug addiction (Dickinson et al., 2002; Robbins and Everitt, 2002; Miles et al., 2003; Cardinal and Everitt, 2004). These ideas are rooted in classic theories of stimulus-response (S-R) reinforcement learning (Hull, 1943). According to these, rewarding events reinforce or create direct, causal associations between contiguous sensations and responses, allowing the stimulus directly to elicit the response in a manner that is no longer dependent on the response-outcome contingency or the outcome's value (Holman, 1975; Adams and Dickinson, 1981; Dickinson et al., 1983; Dickinson et al., 1995) (Figure 3A).

Although it is straightforward to apply these ideas to drug addiction, only relatively recently was direct evidence found to suggest that these mechanisms also apply to activities associated with natural rewards like food. For example, to test contingency sensitivity, Dickinson et al (1996) trained hungry rats to press two levers for food pellets before delivering a sugar solution freely and periodically. Responding on one lever had no effect on sugar delivery, but responding on the other delayed it; in other words, to maximize their access to both food pellets and sugar, the rats had to withhold responding on one lever but not the other. Consistent with the results discussed in the previous section, this proved to be a relatively easy task for undertrained animals. However, animals who had been overtrained on the initial lever pressing task did not adjust readily to this omission contingency and kept responding similarly on both levers even though this lost them significant access to the sugar (Dickinson et al., 1998). This same effect has recently been replicated in mice (Frankland et al., 2004).

After overtraining, instrumental actions can also fail to satisfy the second criterion for goal-directed behavior: they can be insensitive to changes in goal value. Holman (Holman, 1975) showed that overtrained lever presses in thirsty rats reinforced by access to a saccharin solution persisted even after the saccharin had been devalued by pairing its consumption with illness. (The test was performed in extinction, that is, without further delivery of the saccharin.) It is important to recognize how maladaptive

the lever pressing was in Holman's rats. Although the pairing with illness resulted in the rats no longer consuming or even contacting the previously palatable (but now poisonous) saccharin, they continued to work on the lever at a rate comparable to that of rats for which the saccharin was not devalued. This effect was replicated several times over the next decade (Adams, 1981; Adams and Dickinson, 1981; Dickinson et al., 1983). These findings provided direct evidence that, in addition to control by a goal-directed process, the performance of instrumental actions can also become habitual. The later experiments also show that either process can be engaged depending not just on the degree of training but also on the relationship between instrumental performance and reward delivery. In particular, when rewards are delivered on a timed schedule so that changes in the rate of lever pressing have little effect on the rate of reward, actions tend to become habitual. When rewards are delivered so that the experienced rate of reward is proportional to the rate of lever pressing, however, actions tend to remain goal-directed.

To review, the experiments above suggest that there exists a second class of "habitual" instrumental actions that are distinguished from goal-directed actions by failing to meet the criteria of contingency or outcome sensitivity. As mentioned above, these habitual responses are classically envisioned to arise from stimulus-response associations, which lack any representation of the rewarding outcome and are instead valued in the sense of having been "stamped in" by a history of reinforcement (Hull, 1943). We refer to this propensity to emit a habitual response as a *habit value*. Finally, and interestingly, animals not only acquire new actions in situations where they are reinforced by primary rewards such as food, but also when they are reinforced by stimuli associated with those primary rewards. This latter effect, called *conditioned reinforcement*, has also been argued to be mediated by a process of S-R association, in this case reinforced by the stimulus signaling reward rather than the reward itself. In confirmation of this suggestion, recent studies have found evidence that the actions acquired through this process are insensitive to outcome devaluation (Parkinson et al., 2005).

2.4. Expectancy, reinforcement and reward

As described in this section, behavioral research provides considerable evidence for at least three distinct forms of evaluative process, which bias decision-making based on expectancy, reinforcement and reward (Figure 3B). Expectancy induced by stimuli that predict reward – Pavlovian values – can exert selective effects on action selection; these stimuli selectively elevate the performance of actions associated with the same outcome as that predicted by the stimulus relative to other actions. Only cues that are valid predictors of an outcome associated with an action affect its performance in choice tests, suggesting that this effect is based on the predictive information that the stimuli provide about the likelihood of specific outcomes.

This effect of Pavlovian values differs from the biasing of action selection induced by reinforcement, whether it is an effect of primary or of conditioned reinforcement. In this case the bias toward the performance of an action that has proven to be successful in the past is based on an association between the action and the stimulus or state in which the action was rewarded. Hence, this is considered a reinforcement function of reward delivery; it is not the anticipation of the reward but the strengthening of the stimulus-response association that produces the change in performance of the response and establishes the strength of the habit value.

Finally, the reinforcing function of reward delivery is to be contrasted with its function as a valued goal of an action. The reward value of goals clearly biases the animals' tendency to select and initiate a particular course of action in a way that does not depend upon the state or stimulus with which that goal is associated or on the function of the goal as a reinforcer of a specific stimulus-response connection. Rather reward value appears to modulate the tendency to initiate an action that is selected based on the encoding of the action-outcome contingency. These distinct functions are best illustrated in the differential effects of reward devaluation and contingency degradation on undertrained and relatively more overtrained actions.

Generally, therefore, as illustrated in Figure 3B, behavioral evidence points to three dissociable 'value' functions that guide the choice and decision-making process. But how deep does this analysis run? In the following sections we will assess this question

from the perspective of both contemporary computational models and their neural implementation within the functional systems that current evidence suggests mediate the neural bases of decision making.

3. Reinforcement learning

In contrast to psychologists studying natural agents, computer scientists have also long studied issues of prediction and action in the context of controlling artificial agents such as robots. In fact, there have been important relationships forged between the approaches taken to the study of adaptive behavior in natural and artificial systems. Reinforcement learning (Sutton & Barto, 1997), for instance, is the study of how to learn, by trial and error, to make efficient decisions. Apart from its practical engineering uses, reinforcement learning has provided an apt set of models both for conditioning behavior and the underlying neurophysiology.

Here, we formally define the problem of reinforcement learning and review some basic approaches to solving it. (For more details, see also the chapter by Niv and Montague in this volume.) In the process, we will note striking analogies between these computational tools and several of the psychological constructs outlined in section 2; notably those of Pavlovian, habitual, and goal values.

3.1 The Markov decision process

Most work in reinforcement learning targets a class of model tasks, the Markov decision process (MDP), which is simplified enough to admit of formal analysis but still embodies many features of nontrivial real-world decisions. Here we lay out the model.

In an MDP, the counterpart to a stimulus in conditioning is a *state*. At each timestep t , the model world takes on a discrete state s_t . Also, the agent chooses some action a_t . Together, these probabilistically determine the next state, s_{t+1} . Formally, the structure of a particular MDP is defined by the *transition function*

$$T(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a) \quad [1]$$

that specifies the probability distribution over the new state (s') that follows any state (s), action (a) pair.

In selecting actions, the goal of the agent is to obtain rewards. The variable r_t measures (as a real number) the reward value that the subject receives on trial t . We assume that this is a function of the state: $r_t = R(s_t)$. Together, the *reward function* R and the transition function T define an MDP.

For instance, a typical conditioning experiment might be modeled with states corresponding to the different controlled stimuli and outcomes; actions such as lever presses; a transition function detailing how actions bring about outcomes; and finally a reward function mapping some of these states (e.g. those corresponding to consuming various foods like pellets and sucrose) to their rewarding value. Note that this formalism respects the distinction between the predicted sensory and the motivational aspects of an outcome (here, represented by the outcome state itself and its reward value). The following sections describe how different sorts of value can be built up from these elements. As with their psychological counterparts, goal and habit values, these can be differentially sensitive to manipulations of reward value, e.g. by illness or satiety treatments. In the present setting, these revaluation manipulations can be viewed as changing the reward function (Daw et al, 2005; Niv et al 2006).

3.2 Action selection in MDPs

The central problem in reinforcement learning is being dropped into an unknown MDP – like a rat into a Skinnerbox – and learning, by trial and error, how best to gather rewards. A major difficulty in this setting is that each action influences the next state and reward, but this state also influences the next one and its reward, and so on. We must therefore evaluate actions in terms of their long-term consequences.

To this end, we can define the *state-action value function*, Q , as the cumulative reward that follows a particular action in a particular state:

$$Q(s, a) = E[r_t + r_{t+1} + r_{t+2} + r_{t+3} + \dots | s_t = s, a_t = a] \quad [2]$$

Here the expectation $E[\square]$ averages over randomness in the state transitions, and we are being temporarily ambiguous about what actions the agent takes at times $t+1$ and thereafter.

The motivation for this definition is that since $Q(s, a)$ measures the long-term

reward following an action, it ameliorates the principal difficulty of decision-making in an MDP. If we knew it, then at any state optimal action selection would be as simple as choosing whichever action had maximal value. The strategy is to guide choice by predicting future rewards.

We can rewrite the function in a more useful form by noting that the bulk of the sum (all the terms starting with r_{t+1}) is just the value of the successor state, $Q(s_{t+1}, a_{t+1})$. This motivates the following *recursive* definition of the long-term value Q in terms of itself:

$$Q(s, a) = R(s) + \sum_{s'} T(s, a, s') \max_{a'} [Q(s', a')] \quad [3]$$

Here, the first reward r_t is determined by the MDP reward function R , and the average over the successor state is written explicitly, according to the probabilities given by the transition function T . Also, this equation resolves the ambiguity about what action the agent takes in the successor state s' by assuming she takes the best action there, which has value $\max_{a'} [Q(s', a')]$, that is the value of the best action a' in that state.

Formally, the definition of long-term value in Equation 3 specifies what it means to choose optimally in an MDP. More practically, it serves as the basis for a number of different approaches to reinforcement learning. These simply correspond to different methods of using experience to solve Equation 3 for the action value function $Q(s, a)$.

Returning finally to the psychological context, we may ask to which of our notions of value this long-term value corresponds. In fact, we will suggest, it corresponds to *both* habit value and goal value. On this view, the difference between the two is not in what information they carry, but rather in how this information is inferred from experience.

In particular, estimates of Q may be derived using different reinforcement learning procedures that either do, or do not, make use of an intermediate representation of the action-outcome transition function and the identity of the goal outcome. These two sorts of value suggest computational counterparts to goal and habit values, respectively, in that the first representation satisfies both of criteria for goal-directedness from the previous section, whereas the latter does not (Daw et al. 2005).

3.3 Model-based reinforcement learning and goal values

A straightforward, though indirect, approach to predicting action values in an

initially unknown task is simply to attempt to figure out what the task is: formally, to estimate the reward function R and transition function T that together define the MDP. These functions form a complete description, or *model*, of the MDP, and methods that work by recovering them are accordingly known as *model-based* reinforcement learning. Note that these functions are easy to learn from experience; each state transition observed or reward received is a direct sample from one of these functions; one need only average this raw experience to estimate the underlying functions.

Given R and T , it is possible, though computationally laborious, to estimate the action value function $Q(s,a)$ simply by evaluating Equation 3. The standard procedure for doing this is called value iteration. It requires plugging in these functions, then iteratively expanding the future value term on the right hand side of Equation 3, adding up expected rewards over different anticipated trajectories of potential future states and actions. The process is rather like how a novice evaluates chess moves, by exhaustively considering series of future states (board positions) and actions (moves and countermoves).

Values derived this way are a computational counterpart to goal values (Daw et al 2005). This is because the transition function T embodies knowledge, analogous to response-outcome associations, about how particular actions in particular states lead to other states, including those representing particular outcomes like various foods. Similarly, the reward function R represents the rewarding value of those outcomes. Decisions derived from these functions will therefore be sensitive to outcome contingency and value, the two criteria for goal directed actions. For instance, if the reward value of these outcomes changes (as represented by the reward function R), then Q values and choices derived from this knowledge will immediately reflect the change.

3.4 Model-free reinforcement learning and habit values

Instead of estimating future value indirectly, via the transition and reward functions, an alternative approach is to estimate it directly, by averaging samples of the right hand side. One version of this idea, called Q-learning (Watkins, 1989), works as follows.

Suppose we have some initial estimate of the value function, which we denote $\hat{Q}(s, a)$ to distinguish it from the true function. On timestep t , starting in state s_t , we receive reward r_t and take action a_t . This leads us to a new state s_{t+1} , drawn from the transition distribution $T(s_t, a_t, \cdot)$. This small snatch of experience gives us a new estimate of the future value $\hat{Q}(s_t, a_t)$, to wit: $r_t + \max_{a'} [\hat{Q}(s_{t+1}, a')]$, and we may update our previous estimate $\hat{Q}(s_t, a_t)$ by averaging this in. In the next step, we may go on to update our estimate $\hat{Q}(s_{t+1}, a_{t+1})$ using the subsequent reward and state transition, and so on.

A standard approach is to mix the new sample with the old prediction in proportions determined by a “learning rate” parameter η . The resulting update rule for $\hat{Q}$ can then be written:

$$\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) + \eta \delta_t \quad [4]$$

where the “prediction error” δ_t measures the difference between the old prediction and the newly observed sample:

$$\delta_t = r_t + \gamma \max_{a'} [\hat{Q}(s_{t+1}, a')] - \hat{Q}(s_t, a_t) \quad [5]$$

Algorithms of this sort are known as *temporal-difference* algorithms, because the last two terms of the prediction error are the difference between estimates $\hat{Q}$ made at successive timesteps. They are also known as *model-free*, because in contrast to the approach discussed in the previous section, they do not rely on any representation of the MDP.

For this reason, like habit values, values predicted this way will be insensitive to changes in the reward value (Daw et al. 2005; Dayan & Balleine 2002). Since the learned function $\hat{Q}(s, a)$ is not grounded in any information about the *identities* of the future outcomes (such as food) contributing to the predicted value, it is analogous to the formation of a stimulus-response association. Also like habit values, values learned this way are grounded in a history of reinforcement rather than a prediction of a particular future reward.

3.4 The actor/critic and Pavlovian values

One variation on temporal-difference methods is particularly relevant to the study of conditioning because it decomposes value learning into learning about states (“Pavlovian” learning) and, separately, learning about actions (“instrumental” learning). These methods are known as actor/critic algorithms (Barto, 1995).

A reinforcement learning counterpart to Pavlovian value is a function measuring future rewards expected from a state, while ignoring actions. (That is, strictly speaking, averaging out the agent’s action choices as though they were just another source of randomness in state-state transitions. The values derived are denoted as “ V ” values rather than “ Q ” values)

The state value function has its own recursive definition, analogous to Equation 3:

$$V(s) = R(s) + \sum_a \sum_{s'} T(s, \pi(a), s') V(s') \quad [6]$$

(where $\pi(a)$ is the probability that action a will be chosen). It also has its own temporal-difference learning rule with error signal

$$\delta_t^V = r_t + \hat{V}(s_{t+1}) - \hat{V}(s_t) \quad [7]$$

Given a Pavlovian module that learns to predict $\hat{V}(s)$, called a *critic*, we can define a separate *actor* module that uses this information to learn how to select actions. One version of this - called “advantage learning” (Baird, 1994, Dayan & Balleine 2002) - defines the *advantage* A of action a in state s as

$$A(s, a) = Q(s, a) - V(s) \quad [8]$$

that is, the difference between the future value of taking action a and the future value averaged over actions according to their usual frequencies π . As it turns out, this function can be estimated simply by averaging TD errors from Equation 6 (which are samples of the advantages). Here, like Q , $\hat{A}(s, a)$ serves as a habit value implicitly defining an action selection policy (take whichever action has the highest advantage at a state); the Pavlovian predictions from the critic help to learn this policy by defining the error signal δ_t^V that trains it.

Actor/critic algorithms like this one work by learning Pavlovian predictions $\hat{V}(s)$ and

then learning to choose those actions that tend to lead to states with a high Pavlovian value. The psychological counterpart is conditioned reinforcement, in which responses can be reinforced not just by primary rewards but also by Pavlovian cues predicting reward. As discussed in Section 2, this is only one of the ways in which Pavlovian and instrumental conditioning interact. Other, more selective, effects of Pavlovian values on choice (e.g. that illustrated in Figure 2) go beyond the scope of the reinforcement learning models discussed here (but see Niv et al., 2007).

In summary, we have suggested that goal and habit values may correspond to two computational approaches to the problem of predicting value for guiding action choice, and that the learning of habit values can be further subdivided into Pavlovian and instrumental subproblems (Figure 4A).

4. Neural basis of reinforcement learning

The dissociation between multiple valuation mechanisms is also evident in their underlying neural substrates. Here we review parallel evidence from animals and humans pointing to particular and distinct, neural circuits supporting different forms of valuation (Figure 4B).

4. 1 Expected reward: Value signals and Pavlovian values

As described above, the reinforcement learning counterpart for Pavlovian values is the state value function. Studies of the neural basis of state or Pavlovian values have focused on three brain regions in particular: (1) the amygdala in the medial temporal lobes, (2) the orbitofrontal cortex on the ventral surface of the frontal lobes, and (3) the ventral striatum in the basal ganglia.

Single-unit recording studies in both rodents and nonhuman primates have implicated neurons in both amygdala and orbitofrontal cortex in encoding stimulus-reward associations (Schoenbaum et al., 1998; Paton et al., 2006). Cue-related and anticipatory responses have also been found in orbitofrontal cortex that relate to the monkeys' behavioral preference for the associated outcomes (Tremblay and Schultz, 1999).

Although amygdala lesions generally result in mild or no impairment on acquisition of Pavlovian or indeed instrumental associations (Baxter and Murray, 2002), lesions of either the amygdala and orbitofrontal cortex or crossed unilateral lesions of both structures, result in impairments in the modulation of conditioned Pavlovian responses following changes in the value of the associated outcome, induced by an outcome devaluation procedure (e.g. pairing the outcome with a malaise inducing substance such as lithium chloride, or feeding the animal to satiety on that outcome) (Hatfield et al., 1996; Malkova et al., 1997; Baxter et al., 2000; Pickens et al., 2003; Ostlund and Balleine, 2007).

Another brain region implicated in Pavlovian valuation is the ventral striatum, which receives afferent input from both the amygdala and orbitofrontal cortex (Fudge et al., 2002; Haber et al., 2006). Activity of some neurons in the ventral striatum have, like those in the amygdala and orbitofrontal cortex, been found to reflect expected reward in relation to the onset of a stimulus presentation, and activity of neurons in this area have been found to track progression through a task sequence ultimately leading to reward (Shidara et al., 1998; Cromwell and Schultz, 2003; Day et al., 2006). Some studies report that lesions of a part of the ventral striatum, the nucleus accumbens core, can impair learning or expression of Pavlovian approach behavior (Parkinson et al., 1999). Lesion of the dopaminergic input into accumbens can also impair such learning (Parkinson et al., 2002), suggesting an important role for ventral striatum in mediating learning of Pavlovian conditioned associations.

Functional neuroimaging studies in humans have also revealed activations in some of these areas during both appetitive and aversive Pavlovian conditioning in response to CS presentation (e.g. Gottfried et al., 2002, 2003). When taken together, the above findings implicate a network of brain regions involving the amygdala, ventral striatum and orbitofrontal cortex in learning and expression of Pavlovian values, which may correspond to the state-value component of some reinforcement learning models.

4.2 Learning of state value representations

The finding of Pavlovian value signals in the brain raises the question of how such

signals are learned in the first place. As we saw in section 3, a key feature of most reinforcement learning models is the use of a prediction error signal to update expectations of future reward based on differences between expected and actual outcomes.

Initial evidence for prediction error signals in the brain emerged from the work of Wolfram Schultz and colleagues who observed such signals by recording from the phasic activity of dopamine neurons in awake behaving non-human primates undergoing simple Pavlovian or instrumental conditioning tasks with reward (Mirenowicz and Schultz, 1994; Schultz et al., 1997; Hollerman and Schultz, 1998; Schultz, 1998). These neurons, particularly those present in the ventral tegmental area in the midbrain, project strongly to the ventral striatum, amygdala and orbitofrontal cortex (Oades and Halliday, 1987), the three regions most closely identified as playing a role in encoding representations of stimulus-bound reward expectancies discussed earlier. The response profile of these dopamine neurons closely resembles a specific form of prediction error derived from the temporal difference learning rule, in which predictions of future reward are computed at each time interval within a trial, and the error signal is generated by computing the difference in successive predictions (Montague et al., 1996; Schultz et al., 1997). Just like the temporal difference prediction error signal, these neurons increase their firing when a reward is presented unexpectedly; decrease their firing from baseline when a reward is unexpectedly omitted; respond initially at the time of the US before learning is established but shift back in time within a trial to respond instead at the time of presentation of the CS once learning has taken place. (For more details, see Niv and Montague, this volume.).

In order to test for evidence of a temporal difference prediction error signal in the human brain, O'Doherty et al., (2003) scanned human subjects while they underwent a classical conditioning paradigm in which associations were learned between visual stimuli and a pleasant sweet taste reward. Significant correlations were found between a temporal difference prediction error signal and BOLD responses in a number of brain regions, most notably the ventral striatum (ventral putamen bilaterally) (Figure 5) and orbitofrontal cortex, both prominent target regions of the midbrain dopamine neurons

believed to carry a reward prediction error. These results suggest that prediction error signals are present in the human brain during Pavlovian reward-learning.

The evidence discussed so far supports the presence of prediction error signals during learning involving appetitive or rewarding stimuli. The next obvious question is whether such signals can be found to underlie learning about punishing as well as rewarding events. Evidence of a role for dopamine neurons in responding during aversive learning is mixed. Single-unit studies have generally failed to observe strong dopaminergic activity in response to aversive events (Schultz, 1998), and indeed it has been found that dopamine neurons may in fact be inhibited in responding during aversive stimulation such as tail pinch in rats (Ungless et al., 2004). On the other hand, a number of studies measuring dopamine release in the striatum in rats using microdialysis techniques have found evidence for increased dopamine levels during aversive as well as appetitive conditioning (Pezze and Feldon, 2004). However, as termination of an aversive stimulus can in itself be rewarding, the implications of these studies for understanding the role of dopamine in aversive learning are still debated. Irrespective of whether dopamine will turn out to play a role in aversive learning or not, existing evidence appears to rule out the suggestion that phasic dopamine encodes prediction errors for aversive events in the same way that it does for appetitive events. This then raises the question of whether prediction error signals for aversive learning are present anywhere else in the brain, such as by another neuromodulatory system. The proposal has been made that serotonin neurons in the dorsal raphe might fulfill this role (Daw et al., 2002).

In fact, neuroimaging studies have revealed strong evidence for the presence of prediction error signals in the human brain during aversive as well as appetitive learning (Seymour et al., 2004; Seymour et al., 2005). Given that such aversive signals in striatum are unlikely to depend on the afferent input of dopamine neurons, these findings also show that BOLD activity in ventral striatum should not be considered to be a pure reflection of the afferent input of dopamine neurons, an interpretation implicitly assumed in some of the reward imaging literature. Rather, activity in striatum is likely to also reflect the influence of a number of different neuromodulatory systems in addition to

dopamine, input from other cortical and sub-cortical areas, as well as intrinsic computations within this region.

The studies discussed above demonstrate that prediction error signals are present during learning to predict both appetitive and aversive events, a finding consistent with the tenets of a prediction error-based account of associative learning. However, merely demonstrating the presence of such signals in the striatum during learning does not establish whether these signals are causally related to learning or merely an epiphenomenon. The first study aiming to uncover a causal link was that of Pessiglione et al., (2006) who manipulated systemic dopamine levels by delivering a dopamine agonist and antagonist while subjects were being scanned with fMRI during performance of a reward-learning task. Prediction error signals in striatum were boosted following administration of the dopaminergic agonist, and diminished following administration of the dopaminergic antagonist. Moreover, behavioral performance followed the changes in striatal activity and was increased following administration of the dopamine agonist and decreased following administration of the antagonist. These findings support, therefore, a causal link between prediction error activity in striatum and the degree of behavioral learning for reward.

4.3 The actor/critic in the brain

Next we turn to the mechanisms underlying action-selection for reward in the brain. As outlined in section 3, in order to drive action-selection, reinforcement learning models need to learn about the expected future reward that follows from taking individual actions in order to use these learned action-values to guide choice. The key question then, is where and how are action-related values represented in the brain? We will consider evidence to suggest the possibility that different types of action-related value representations may be encoded in different brain systems, depending on the nature of the associative relationship on which that action-value depends, i.e. whether the action selection is being controlled by the goal-directed or habitual system.

First, we will consider habit values. As outlined in section 3.2, these action-values can be learned by model-free reinforcement learning algorithms such as the actor-critic

(Barto, 1992), in which prediction error signals generated by a critic that evaluates the expected reward available in each state of the world are used to update the policy (or stimulus-response associations) stored in a separate actor module. Thus the implementation of an actor/critic like architecture in the brain would suggest the existence of two separate modules, a critic concerned with learning about reward expectations more generally, and an actor, which stores the action-values and/or policy. Given the connections and anatomy of the basal ganglia, it has been proposed that an actor/critic like process might be implemented in the striatum. Houk et al., (1995) suggested that the actor and critic could be implemented within the patch/striosome and matrix compartments distributed throughout the striatum. However, more relevant to the present discussion was the proposal by Montague et al., (1996) that the ventral and dorsal striatum corresponds to the critic and actor respectively. The evidence reviewed above would appear to implicate a circuit involving the ventral striatum alongside the orbitofrontal cortex and amygdala in implementing a function rather similar to that proposed by the critic and involving the encoding of state values.

Evidence is also beginning to accumulate implicating at least a part of the dorsal striatum in implementing the actor. More specifically, lesions of the dorsolateral striatum have been found to abolish the onset of habitual behavior in rats, rendering the lesioned animals permanently goal-directed (Yin et al., 2004). This study therefore suggests a critical involvement of the dorsolateral striatum in implementing stimulus-response learning in the rodent brain. (O'Doherty et al., 2004), used fMRI to compare responses on an instrumental choice task to that elicited during a control Pavlovian task in which subjects experienced the same stimulus-reward contingencies but did not actively choose which action to select. This comparison was designed to isolate the actor, which was hypothesized to be engaged only in the instrumental task, from the critic, which was hypothesized to be engaged in both the instrumental and Pavlovian control tasks. Consistent with the proposal of a dorsal vs ventral actor/critic architecture, activity in dorsal striatum was found to be specifically correlated with prediction error signals in the instrumental task but not in the Pavlovian task, whereas ventral striatum was correlated with prediction error signals in both tasks (Figure 6), a finding that has received support

from a number of other fMRI studies (Haruno et al., 2004; Tricomi et al., 2004).

It is important to note that although the architecture within the ventral and dorsal striatum is consistent with an actor/critic like process, it is not yet clear whether learning of habit values in the dorsal striatum is governed by state-value prediction errors generated from the critic, as should be the case for the *literal implementation* of the full actor/critic model. It is also plausible that habit values within the dorsal striatum might be learned and updated directly, as instantiated in other variants of reinforcement learning such as Q-learning or SARSA (Watkins and Dayan, 1992; Sutton and Barto, 1998).

4.4 Brain systems underlying goal-directed learning in rats and primates

In addition to the evidence reviewed above implicating dorsolateral striatum in habitual aspects of instrumental responding, evidence is emerging to implicate a different set of brain regions in mediating the goal-directed component of instrumental responding. In rodents, two brain regions in particular have been implicated in goal-directed learning: the prelimbic cortex in the frontal lobe, and the area of striatum to which this region of cortex projects: the dorsomedial striatum. Lesions of either of these regions in rats prevents the acquisition of goal-directed learning, rendering animals habitual even during the early stages of training (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Yin et al., 2005). Notably, prelimbic cortex although necessary for initial acquisition does not appear to be necessary for the expression of goal-directed behavior as lesions of this area do not impair goal-directed behavior once the lesions are induced after initial training (Ostlund and Balleine, 2005). On the other hand, dorsomedial striatum does appear to be critical for both learning and expression of goal-directed behavior, as lesions of this area impair such behavior if induced either before or after training (Yin et al., 2005).

This raises the question of whether there exist homologous regions of the primate prefrontal cortex that contribute to similar functions. A number of fMRI studies in humans have implicated parts of the frontal cortex, especially its ventral aspects, in encoding the value of chosen actions (Daw et al., 2006; Kim et al., 2006). Taken together, these

findings suggest that orbital and medial prefrontal cortex are involved in keeping track of the expected future reward associated with chosen actions, and that these areas show a response profile consistent with an expected value signal generated by reinforcement learning models. However, the above studies do not differentiate between action-related value representations based on stimulus-response associations from those based on action-outcome associations.

To determine whether such signals reflect goal values or habit values, Valentin et al., (2007) performed an outcome devaluation paradigm in humans, similar to the manipulation used in rodents to determine whether instrumental behavior is goal-directed or habitual. By testing for regions of the brain showing a change in activity during selection of an action associated with a devalued outcome (devalued by feeding subjects to satiety on that outcome), it was possible to test for regions showing sensitivity to the learned action-outcome associations. The regions found to show such a response profile were medial and central OFC (Figure. 7). These findings suggest that action-outcome information is present in OFC alongside stimulus-outcome representations, indicative of a role for OFC particularly its medial aspects in encoding expectations of reward tied to specific actions above and beyond its role in encoding stimulus bound predictions. However, in apparent contradiction of the above findings, it has been found that lesions of OFC in rats do not produce impairments at goal-directed learning in contrast to the effects of lesions of the prelimbic area that do produce robust deficits on this capacity (Ostlund and Balleine, 2007).

The source of such discrepancies between studies remains to be determined, but one intriguing possibility is that rat and human OFC may not be entirely homologous. It is interesting to note that in a previous human stimulus-based devaluation study by Gottfried et al. (2003), modulatory effects of reinforcer devaluation were found in central, but not medial OFC areas, whereas in the Valentin et al. study, evidence was found of instrumental devaluation effects in both central and medial areas. This raises the possibility that medial OFC may be more involved in the goal-directed component of instrumental conditioning by encoding goal values whereas central OFC may contribute more to Pavlovian stimulus-outcome learning by encoding Pavlovian values (as this area

was found in both the Valentin et al. study and in the previous Pavlovian devaluation study). This speculation is consistent with the known anatomical connectivity of these areas in which central areas of OFC (Brodmann areas 11 and 13) receive input primarily from sensory areas, consistent with a role for these areas in learning the relationship between stimuli, whereas the medial OFC (areas 14 and 25) receives input primarily from structures on the adjacent medial wall of prefrontal cortex such as cingulate cortex, an area often implicated in response selection and/or reward-based action choice (Carmichael and Price, 1995, 1996). It is also notable that although the majority of single-unit studies in monkeys have reported stimulus-related activity and not response-related selectivity in the OFC (e.g., Thorpe et al., 1983; Tremblay and Schultz, 1999; Padoa-Schioppa and Assad, 2006) these studies have typically recorded from more lateral and central areas of the OFC (Brodmann areas 12/47 and 13, respectively), and not from more medial areas. It is therefore plausible that more medial sectors of the OFC in humans correspond to regions considered part of medial prefrontal cortex in rats that have been more conclusively linked to goal-directed learning (Corbit and Balleine, 2003; Balleine and Dickinson, 1998; Ostlund and Balleine, 2005). Indeed, consistent with this hypothesis, Ostlund & Balleine (2007) found that, whereas medial prefrontal cortex was involved in goal-directed action, lesion-induced damage to the more central and lateral regions of the OFC, while leaving goal-directed action intact, abolished the ability of Pavlovian cues to selectively elevate instrumental performance in an outcome-selective Pavlovian-instrumental transfer task (such as that illustrated in Figure 2).

As outlined in section 3.3, a computational framework for goal-directed learning is to propose that this form of learning is mediated by a form of 'model-based' inference in which the agent uses a full model of the decision problem to iteratively evaluate the future consequences of each action in order to compute action-values. Evidence that model-based inference may be implemented in the human ventromedial prefrontal cortex has emerged from a study by Hampton et al., (2006), wherein subjects were scanned with fMRI while participating in a decision problem called probabilistic reversal learning which has a hidden structure in that rewards available following choice of two different actions are anti-correlated, and the values of the two actions reverse from time

to time. Hampton et al., compared two computational models in terms of how well they could account for human choice behavior on the task and for the pattern of neural activity in vmPFC during task performance. One of these algorithms incorporated the rules or structure of the decision problem as would be expected for a model-based inference mechanism, whereas the other model was a simple model-free reinforcement learning algorithm that did not incorporate the structure and thus would only learn to slowly and incrementally update values following successive reinforcements. Consistent with a role for vmPFC in model-based inference, predicted reward signals in this region were found to reflect the structure of the decision problem, such that activity was updated instantly following a reversal, rather than being updated incrementally as might be expected for a model-free reinforcement learning mechanism.

5. Summary and Conclusions

We have reviewed evidence that the values underpinning decision making are not unitary but instead are fractionated psychologically, neurally, and computationally. In particular, we have distinguished goal values from habit values: the former more cognitive and grounded in knowledge about particular expected outcomes, the latter reflecting more generalized motivation divorced from any particular goal identity. Computationally, this distinction parallels one between different methods for bringing experience to bear on the problem of learned value predictions. The influences of these values on behavior are dissociable through manipulations such as reward devaluation, which affects goal values but not habit values. Neurally, lesions localize the two sorts of value in discrete cortico-striatal networks, comprising medial vs. lateral portions of dorsal striatum, together with their associated areas of prefrontal cortex and thalamus. Candidate human analogues of many parts of these networks have been identified using functional neuroimaging.

We have also identified a third sort of value – the Pavlovian value associated with stimuli or states. Though such values are not associated with actions, they nonetheless can impact behavior in a number of ways; most particularly these values can affect

choice apparently by providing information as to the likelihood that an action will pay off with a specific outcome or consequence. Yet a third region of striatum, the ventral part, is an important locus for Pavlovian values and their influences on instrumental behavior.

Two major issues remain. First, what role does dopamine play in each of these systems? Computational models such as the actor-critic identify dopamine in ventral and dorsolateral striatum as a signal for training Pavlovian and habitual values, a role the system is anatomically well situated and physiologically well suited to play; particularly given that, according to the models, a single prediction error signal can train both sorts of value. But the involvement of dorsomedial striatum in goal values raises the likelihood of dopaminergic participation in these as well, since this area's tight interrelationship with dopaminergic midbrain parallels that of the other striatal subterritories (Calabresi et al., 2007). From the perspective of computational models of goal-directed action, this is a puzzling prospect, because the learning underlying goal values is envisioned to be of a rather different sort, and because characteristic features of the dopaminergic response (such as the transfer of responses from rewards to stimuli predicting them) seem closely related to schemes (such as "caching"; Daw et al 2005) for learning habit values. Clearly, more data is required about dopamine's involvement with goal values. The single most crucial question is how dopamine neurons behave following reward devaluation. Does the dopaminergic response to a state or action persist when its associated reward is devalued, as would be expected if the response is driven by habit values? Any indications of devaluation sensitivity or outcome specificity in the dopamine response would motivate a rethinking of computational accounts of dopamine, and an investigation of hybrid models (like the successor representation; Dayan 1992) that use habit-style methods to learn outcome-specific goal values.

A second open issue is to what extent our framework of dissociable value systems, whose roots lie largely in animal conditioning, overlaps other dual- or multi-system models popular in human decision-making. Such models are particularly common in behavioral economic approaches to self-control issues such as temporal discounting (Thaler, 1981; Loewenstein and Prelec, 1992). To develop one particular example, several models (e.g. (Laibson, 1997) hold that inter-temporal choice is mediated by two

competing systems, an impulsive system in which future rewards are highly discounted, and a reflective system which discounts future rewards only shallowly. While an obvious possibility is to identify these influences with habitual and goal-directed systems, respectively, it has also been suggested that Pavlovian values might produce impulsive effects in many time-discounting paradigms (Dayan et al 2006). Neurally, (McClure et al., 2004) have proposed on the basis of an fMRI study of temporal discounting behavior that a reflective system is localized in lateral parts of prefrontal cortex, whereas an impulsive system is suggested to be present in the striatum. The evidence behind this putative dissociation has been hotly disputed (Kable and Glimcher, 2007). In any case, this proposed anatomical breakdown does not map perfectly onto the goal-directed vs habitual framework outlined here, because as discussed above, different parts of dorsal striatum are involved in both goal-directed and habitual processes. Also, though the goal-directed system, like the reflective system, is suggested to depend on prefrontal cortex, our imaging and lesion work have implicated medial rather than lateral aspects of these cortices. An important area for future research will be to categorize similarities and differences between these different possible models of human choice, in order to ultimately develop a more parsimonious and unitary model that accounts for each of the different behaviors and most closely reflects the true structure (both computationally and neuranatomically) of the brain's systems for guiding choice.

Length reduction = $9650/12653 = 76\%$ of original submission.

References

- Baxter MG, Murray EA (2002) The amygdala and reward. *Nat Rev Neurosci* 3:563-573.
- Baxter MG, Parker A, Lindner CC, Izquierdo AD, Murray EA (2000) Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *J Neurosci* 20:4311-4319.
- Bloch MH, Leckman JF, Zhu H, Peterson BS (2005) Caudate volumes in childhood predict symptom severity in adults with Tourette syndrome. *Neurology* 65:1253-1258.
- Calabresi P, Picconi B, Tozzi A, Di Filippo M (2007) Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci* 30:211-219.
- Cardinal RN, Everitt BJ (2004) Neural and psychological mechanisms underlying appetitive learning: links to drug addiction. *Curr Opin Neurobiol* 14:156-162.
- Carmichael ST, Price JL (1995) Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. *J Comp Neurol* 363:642-664.
- Carmichael ST, Price JL (1996) Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *J Comp Neurol* 371:179-207.
- Chang JY, Chen L, Luo F, Shi LH, Woodward DJ (2002) Neuronal responses in the frontal cortico-basal ganglia system during delayed matching-to-sample task: ensemble recording in freely moving rats. *Exp Brain Res* 142:67-80.
- Colwill RC, Rescorla RA (1986) Associative structures in instrumental learning. *The Psychology of Learning and Motivation* 20:55-104.
- Colwill RM, Rescorla RA (1988) Associations between the discriminative stimulus and the reinforcer in instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes* 14:155-164.
- Colwill RM, Motzkin DK (1994) Encoding of the unconditioned stimulus in Pavlovian conditioning. *Animal Learning and Behavior* 22:384-394.
- Corbit LH, Balleine BW (2003) The role of prelimbic cortex in instrumental conditioning. *Behav Brain Res* 146:145-157.
- Corbit LH, Muir JL, Balleine BW (2001) The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *J Neurosci* 21:3251-3260.
- Cromwell HC, Schultz W (2003) Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J Neurophysiol* 89:2823-2838.
- Davis J, Bitterman ME (1971) Differential reinforcement of other behavior (DRO): A yoked-control comparison. *Journal of the Experimental Analysis of Behavior* 15:237-241.
- Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Netw* 15:603-616.
- Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441:876-879.
- Day JJ, Wheeler RA, Roitman MF, Carelli RM (2006) Nucleus accumbens neurons encode Pavlovian approach behaviors: evidence from an autoshaping paradigm. *Eur J Neurosci* 23:1341-1351.
- Delamater AR (1995) Outcome-selective effects of intertrial reinforcement in Pavlovian

- appetitive conditioning with rats. *Animal Learning and Behavior* 23:31-39.
- Dickinson A (1994) Instrumental conditioning. In: *Animal cognition and learning* (Mackintosh NJ, ed), pp 4-79. London: Academic Press.
- Dickinson A, Mulatero CW (1989) Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule. *Behavioural Processes* 19.
- Dickinson A, Balleine BW (1994) Motivational control of goal-directed action. *Animal Learning & Behavior* 22:1-18.
- Dickinson A, Nicholas DJ, Adams CD (1983) The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *Quarterly Journal of Experimental Psychology* 35B:35-51.
- Dickinson A, Wood N, Smith JW (2002) Alcohol seeking by rats: action or habit? *Q J Exp Psychol B* 55:331-348.
- Dickinson A, Balleine BW, Watt A, Gonzales F, Boakes RA (1995) Overtraining and the motivational control of instrumental action. *Animal Learning & Behavior* 22:197-206.
- Frankland PW, Wang Y, Rosner B, Shimizu T, Balleine BW, Dykens EM, Ornitz E, Silva AJ (2004) Sensory-gating abnormalities in young males with fragile X syndrome and *Fmr1*-knockout mice. *Molecular Psychiatry* 9:417-425.
- Fudge JL, Kunishio K, Walsh P, Richard C, Haber SN (2002) Amygdaloid projections to ventromedial striatal subterritories in the primate. *Neuroscience* 110:257-275.
- Fuster JM (2000) Executive frontal functions. *Experimental Brain Research* 133.
- Goldman-Rakic PS (1995) Architecture of the prefrontal cortex and the central executive. *Ann N Y Acad Sci* 769:71-83.
- Gottfried JA, O'Doherty J, Dolan RJ (2002) Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *J Neurosci* 22:10829-10837.
- Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301:1104-1107.
- Haber SN, Kim KS, Mailly P, Calzavara R (2006) Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J Neurosci* 26:8368-8376.
- Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci* 26:8360-8367.
- Haruno M, Kawato M (2006) Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Netw* 19:1242-1254.
- Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M (2004) A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci* 24:1660-1665.
- Hatfield T, Han JS, Conley M, Gallagher M, Holland P (1996) Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order

- conditioning and reinforcer devaluation effects. *J Neurosci* 16:5256-5265.
- Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, Hughes G, Elliston LA, Hartog C, Goldstein DR, Thu D, Hollingsworth ZR, Collin F, Synek B, Holmans PA, Young AB, Wexler NS, Delorenzi M, Kooperberg C, Augood SJ, Faull RL, Olson JM, Jones L, Luthi-Carter R (2006) Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet* 15:965-977.
- Holland PC (1979) Differential effects of omission contingencies on various components of Pavlovian appetitive conditioned responding in rats. *J Exp Psychol Anim Behav Process* 5:178-193.
- Holland PC, Gallagher M (2004) Amygdala-frontal interactions and reward expectancy. *Curr Opin Neurobiol* 14:148-155.
- Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci* 1:304-309.
- Holman EW (1975) Some conditions for the dissociation of consummatory and instrumental behavior in rats. *Learning & Motivation* 6:358-366.
- Houk JC, Adams JL, Barto AG (1995) A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In: *Models of Information Processing in the Basal Ganglia* (Houk JC, Davis JL, Beiser BG, eds), pp 249-270. Cambridge: MIT Press.
- Hull CL (1943) *Principles of Behavior*. New York: Appleton.
- Jog MS, Kubota Y, Connolly CI, Hillegaart V, Graybiel AM (1999) Building neural representations of habits. *Science* 286:1745-1749.
- Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. *Nat Neurosci* 10:1625-1633.
- Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol* 4:e233.
- Laibson D (1997) Golden eggs and hyperbolic discounting. *Q J Econ* 112:801-834.
- Lauwereyns J, Watanabe K, Coe B, Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. *Nature* 418:413-417.
- Levy R, Dubois B (2006) Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits. *Cereb Cortex* 16:916-928.
- Loewenstein G, Prelec D (1992) Anomalies in intertemporal choice: evidence and an interpretation. *Q J Econ* 107:573-597.
- Malkova L, Gaffan D, Murray EA (1997) Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *J Neurosci* 17:6011-6020.
- McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306:503-507.
- Miles FJ, Everitt BJ, Dickinson A (2003) Oral cocaine seeking by rats: action or habit? *Behav Neurosci* 117:927-938.
- Mirenowicz J, Schultz W (1994) Importance of unpredictability for reward responses in primate dopamine neurons. *J Neurophysiol* 72:1024-1027.
- Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936-1947.

- O'Doherty J, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38:329-337.
- O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452-454.
- Oades RD, Halliday GM (1987) Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. *Brain Res* 434:117-165.
- Ostlund SB, Balleine BW (2005) Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *J Neurosci* 25:7763-7770.
- Ostlund SB, Balleine BW (2007) Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. *J Neurosci* 27:4819-4825.
- Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441:223-226.
- Parkinson JA, Roberts AC, Everitt BJ, Di Ciano P (2005) Acquisition of instrumental conditioned reinforcement is resistant to the devaluation of the unconditioned stimulus. *Q J Exp Psychol B* 58:19-30.
- Parkinson JA, Olmstead MC, Burns LH, Robbins TW, Everitt BJ (1999) Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity by D-amphetamine. *J Neurosci* 19:2401-2411.
- Parkinson JA, Dalley JW, Cardinal RN, Bamford A, Fehner B, Lachenal G, Rudarakanchana N, Halkerston KM, Robbins TW, Everitt BJ (2002) Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behav Brain Res* 137:149-163.
- Partridge JG, Tang KC, Lovinger DM (2000) Regional and postnatal heterogeneity of activity-dependent long-term changes in synaptic efficacy in the dorsal striatum. *J Neurophysiol* 84:1422-1429.
- Paton JJ, Belova MA, Morrison SE, Salzman CD (2006) The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* 439:865-870.
- Pavlov IP (1927) Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. xv, 430.
- Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442:1042-1045.
- Pezze MA, Feldon J (2004) Mesolimbic dopaminergic pathways in fear conditioning. *Prog Neurobiol* 74:301-320.
- Pickens CL, Saddoris MP, Setlow B, Gallagher M, Holland PC, Schoenbaum G (2003) Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. *J Neurosci* 23:11078-11084.
- Pithers RT (1985) The roles of event contingencies and reinforcement in human autoshaping and omission responding. *Learning and Motivation* 16:210-237.
- Poldrack RA, Clark J, Pare-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck

- MA (2001) Interactive memory systems in the human brain. *Nature* 414:546-550.
- Rescorla RA, Solomon RL (1967) Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychol Rev* 74:151-182.
- Robbins TW, Everitt BJ (2002) Limbic-striatal memory systems and drug addiction. *Neurobiol Learn Mem* 78:625-636.
- Robinson D, Wu H, Munne RA, Ashtari M, Alvir JM, Lerner G, Koreen A, Cole K, Bogerts B (1995) Reduced caudate nucleus volume in obsessive-compulsive disorder. *Arch Gen Psychiatry* 52:393-398.
- Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat Neurosci* 1:155-159.
- Schultz W (1998) Predictive reward signal of dopamine neurons. *J Neurophysiol* 80:1-27.
- Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593-1599.
- Seppi K, Schocke MF, Prenschiuetz-Schuetzenau K, Mair KJ, Esterhammer R, Kremser C, Muigg A, Scherfler C, Jaschke W, Wenning GK, Poewe W (2006) Topography of putaminal degeneration in multiple system atrophy: a diffusion magnetic resonance study. *Mov Disord* 21:847-852.
- Seymour B, O'Doherty J P, Koltzenburg M, Wiech K, Frackowiak R, Friston K, Dolan R (2005) Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat Neurosci* 8:1234-1240.
- Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429:664-667.
- Sheffield FD (1965) Relation between classical and instrumental conditioning. In: *Classical conditioning* (Prokasy WF, ed), pp 302-322. New York, NY: Appelton-Century-Crofts.
- Shidara M, Aigner TG, Richmond BJ (1998) Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J Neurosci* 18:2613-2625.
- Smith R, *et al.* (2001) Regional differences in the expression of corticostriatal synaptic plasticity. *Neuroscience* 106:95-101.
- Sutton RS, Barto AG (1998) Reinforcement Learning. Cambridge, MA MIT Press.
- Tanaka SC, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S, Doya K (2006) Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Netw* 19:1233-1241.
- Thaler RH (1981) Some empirical evidence on time inconsistency. *Rev Econ Stud* 23:165-180.
- Thorpe SJ, Rolls ET, Maddison S (1983) The orbitofrontal cortex: neuronal activity in the behaving monkey. *Exp Brain Res* 49:93-115.
- Trapold MA, Overmier JB (1972) The second learning process in instrumental conditioning. In: *Classical Conditioning: II. Current research and theory* (Black AA, Prokasy WF, eds), pp 427-452. New York: Appleton-Century-Crofts.
- Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* 398:704-708.

- Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. *Neuron* 41:281-292.
- Ungless MA, Magill PJ, Bolam JP (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science* 303:2040-2042.
- Valentin VV, Dickinson A, O'Doherty JP (2007) Determining the neural substrates of goal-directed learning in the human brain. *J Neurosci* 27:4019-4026.
- Watkins CJ, Dayan P (1992) Q-learning *Machine Learning* 8:279-292.
- Williams DR, Williams H (1969) Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior* 12:511–520.
- Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur J Neurosci* 19:181-189.
- Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci* 22:513-523.

|

BOX: Basic learning processes and some terminology:

Psychologists have long studied fundamental learning processes. To date these have been isolated as being those related to (i) *stimulus habituation*, i.e. the reduction in the response elicited by a stimulus that is induced by its repeated presentation; (ii) *Pavlovian or classical conditioning*, in which the delivery of a biologically potent event (or US; i.e. food water, a predator and so on) is predicted by, or made conditional upon, a relatively neutral stimulus (or CS). As a consequence, the reflexive unconditioned response, or UR, induced by exposure to the US comes to be elicited by the CS (e.g. contact with food can elicit salivation and so can a bell that has been reliably paired with food); and (iii) *instrumental conditioning*, in which access to a motivationally valuable commodity (e.g. food when hungry; safety when apprehensive and so on) is made conditional on the performance of the animals' own actions, such as pressing a lever or pulling a chain.

Various aspects of these types of learning should be recognized. First it is possible to talk descriptively about learning in terms of the acquisition of a response. It is also possible to talk in terms of the mechanism that mediates that change in response. These have resulted in both behavioral and cognitive definitions and theories of learning. On a behavioral definition, all learning involves the association of stimuli (S) and behavioral responses (R) of one kind or another (i.e. is all S-R). On cognitive views, learning involves the formation of a novel cognitive structure or association; e.g. between stimuli (e.g. CS-US) or between responses and outcomes (e.g. R-O) that are then manifest in performance (of the CR in the case of CS-US, and of R in the case of R-O associations). As we discuss in this chapter, much evidence supports the cognitive perspective, but not all. Some training conditions appear to be particularly apt for producing S-R learning.

Finally, a number of general terms are used here that refer to procedures that are commonly used:

Contingency: The net probability of one event given another (e.g. of a US given a CS or of an outcome given a response); i.e. the difference between the probability of event 2 given event 1 ($p(E2/E1)$) and of event 2 given no event 1 ($p(E2/\text{no}E1)$). There is a positive contingency when $p(E2/E1) > p(E2/\text{no}E1)$ and a negative contingency when $p(E2/E1) < p(E2/\text{no}E1)$.

Outcome revaluation: A procedure that serves to change the motivational or reward value of an event: E.g. if rats are trained to respond for sugar, the sugar could be revalued by increasing its value by, say, an increase in food deprivation, or decreasing its value by a decrease in food deprivation. Sugar can also be devalued using taste aversion procedures; i.e. pairing the sugar with the induction of illness (see Figure 1)

Extinction: Refers to the reduction in the performance of a CR or of an instrumental response when the CS or the instrumental action is no longer followed by the US or outcome with which it was previously paired.

Figure 1

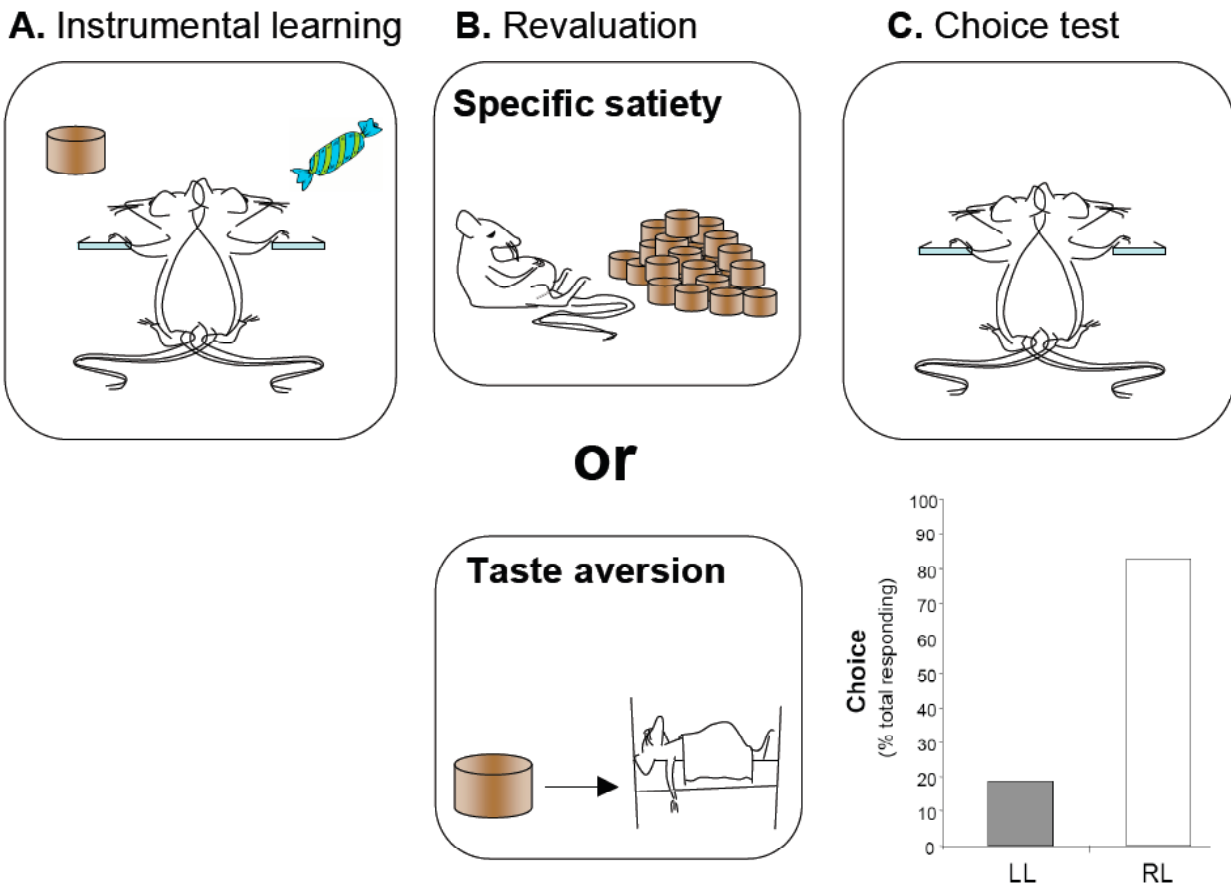


Figure 1. Assessing influence of goal values on choice using outcome devaluation. (A) Rats are first trained to perform two actions each earning one of the two foods. (B) One of the two foods is then revalued; here we illustrate either by taste aversion learning (lower panel) or by a specific satiety treatment (upper panel). (C) The influence of these treatments on choice is then assessed in a choice test conducted in extinction. (D) Typically, devaluation biases choice away actions that, in training, earned the devalued outcome (Data redrawn from Corbit & Balleine, 2003).

Figure 2

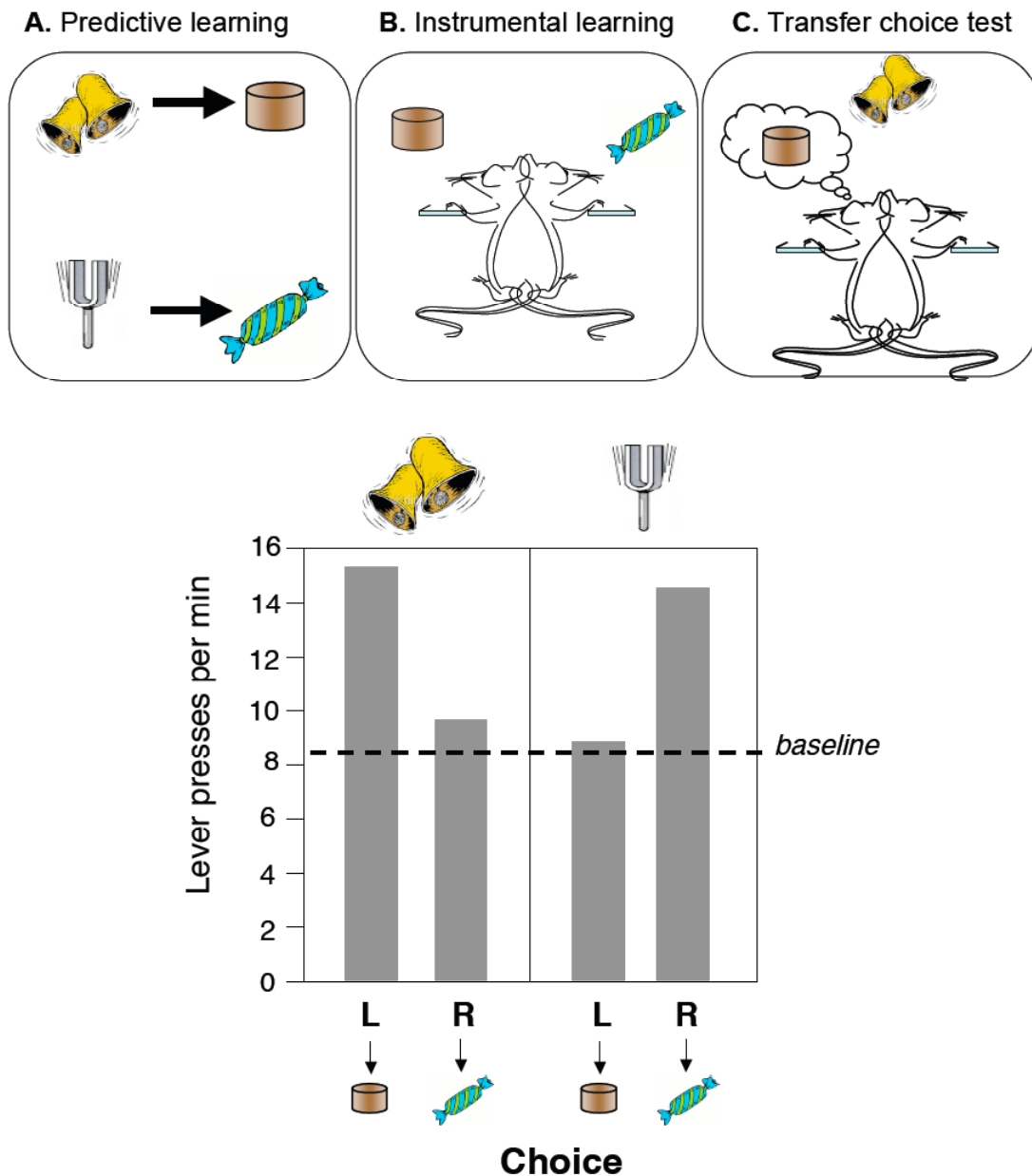


Figure 2. Assessing the influence of Pavlovian values on choice using Pavlovian-instrumental transfer. (A) Rats are first trained to predict two food outcomes based on distinct auditory cues and then (B) to perform two actions each earning one of the two foods. (C) The influence of the predictive cues on choice is assessed in a transfer choice test conducted in extinction. (D) Typically a stimulus biases choice towards actions that earn the outcome predicted by the stimulus (Data redrawn from Corbit & Balleine).

Figure 3

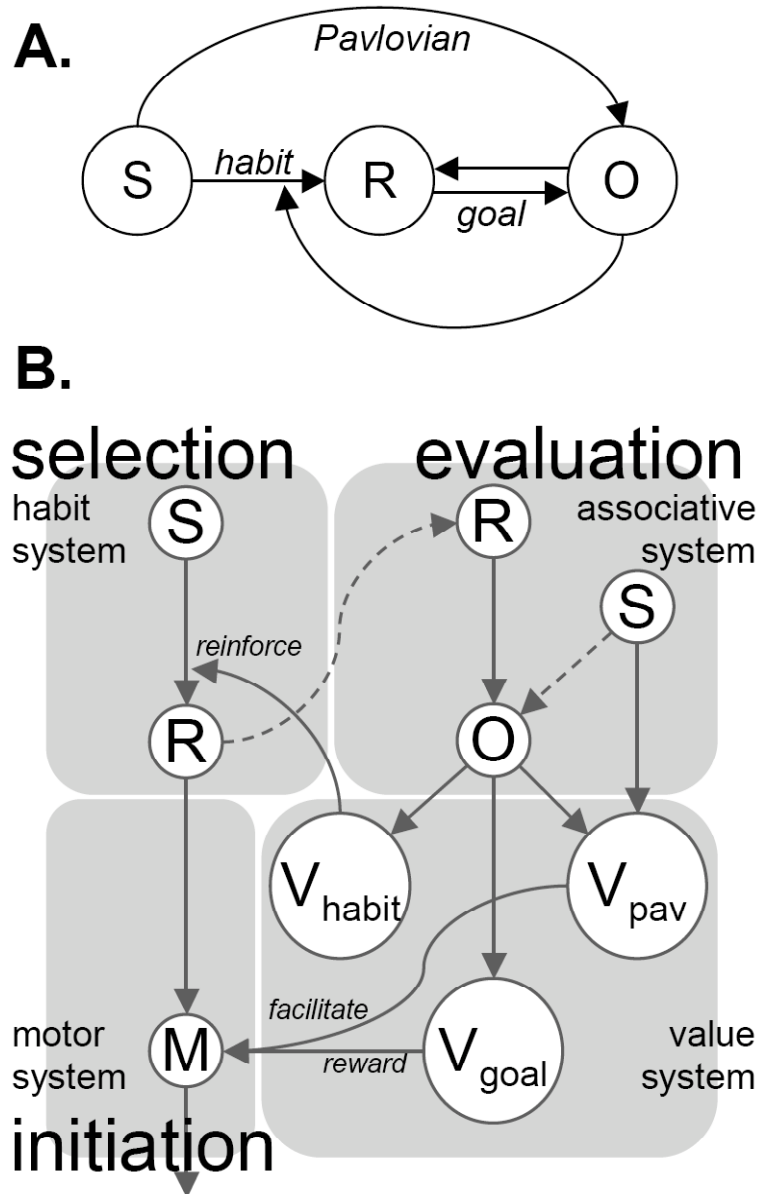


Figure 3: The associative structure of instrumental action. (A) A schematic overview summarizing goal, habit and Pavlovian values and their influence on action through associations between stimuli, responses, and outcomes. (B) Process view of instrumental action: the associative basis of action selection, evaluation and initiation. On this model action selection is modulated by outcome value and inputs from both selection and evaluative processes are required for actions to be performed. Performance can be facilitated by Pavlovian values and, after a period of invariant training, the increase habit value can render it sufficiently strong for selection to result immediately in action initiation without feedback from the goal value.

Figure 4

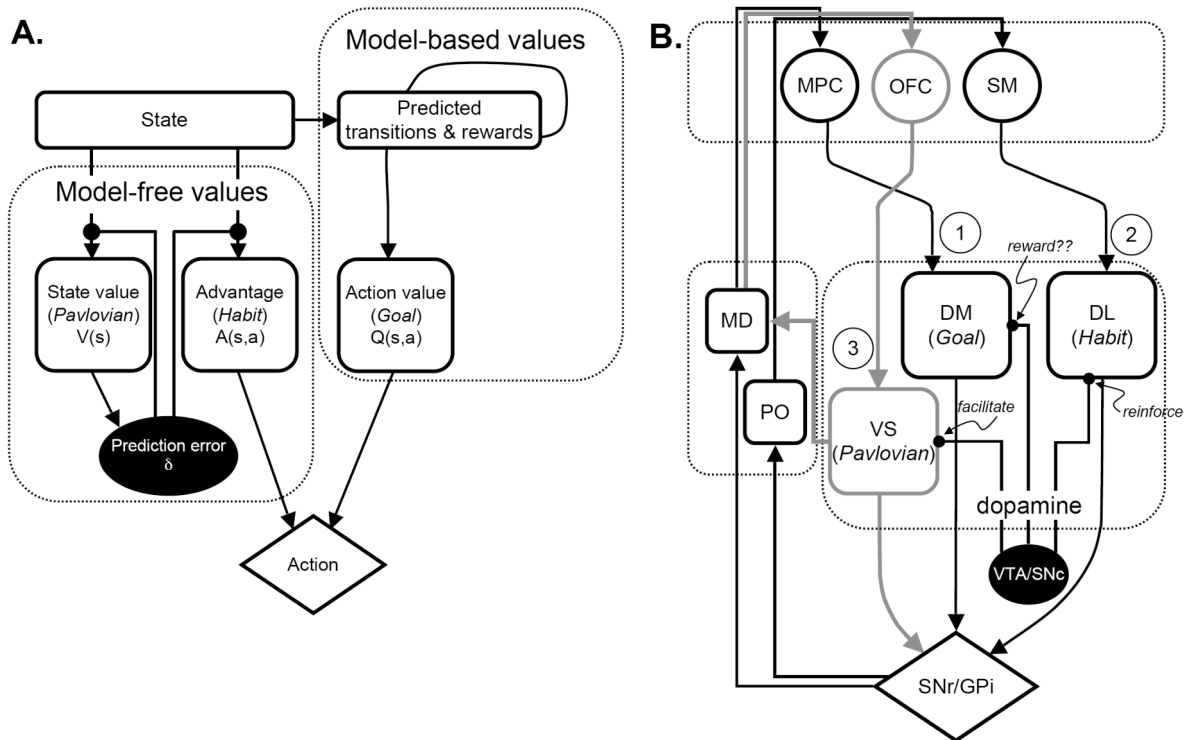


Figure 4: Computational and neural substrates of valuation. (A) Goal, habit and Pavlovian values from model-based and model-free reinforcement learning. (B) Neural circuits underlying valuation in conditioning. Evidence reviewed in text suggests that distinct neural networks mediate the influence of goal values, habit values and Pavlovian values on action selection and initiation. On this view, habits are encoded in a loop-like network involving sensory-motor (SM) cortical inputs to dorsolateral striatum (DL), with feedback to cortex via substantial nigra reticulata/internal segment of the globus pallidus (SNr/GPi) and posterior thalamus (PO) and are motivated by midbrain dopaminergic inputs from ventral tegmental area/substantia nigra pars compacta (VTA/SNc). A parallel circuit linking medial prefrontal cortex (MPC), dorsomedial striatum (DM), SNr and mediodorsal thalamus (MD) mediates goal-directed actions that may speculatively involve a dopamine-mediated reward process. Finally, actions can be facilitated by Pavlovian values mediated by a parallel ventral circuit mediated by orbitofrontal cortex (OFC) and ventral striatal (VS) inputs into the habit and goal-directed loops.

Figure 5

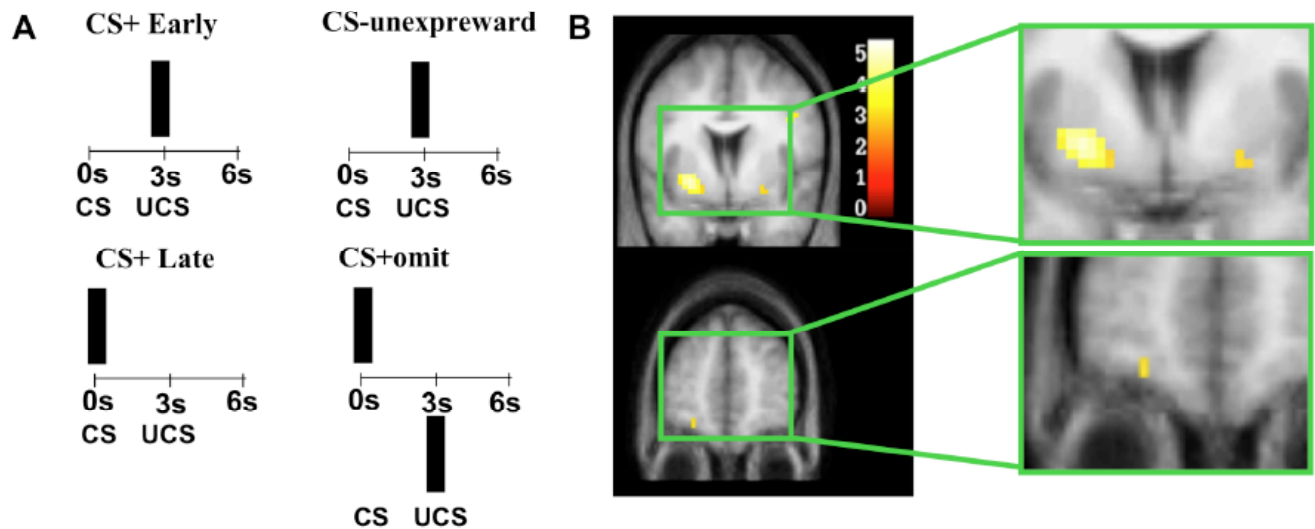


Figure 5. Prediction error signals in the human brain during appetitive Pavlovian conditioning.

A. Illustration of model prediction error signals elicited in an experiment in which in one trial type (CS+) an arbitrary visual cue is associated 3 s later with delivery of a taste reward (1M glucose), and in another trial type (CS) a different cue is followed by no taste. In addition, occasional surprise trials occur in which the CS+ is presented but the reward is omitted (CS+omit), and the CS is presented but a reward is unexpectedly delivered (CS_unexpreward). During early CS+ trials (before learning is established) the PE signal should occur at the time of delivery of the reward, whereas by late CS+ trials (post-learning) the signal should have switched to the time of presentation of the CS. On CS+omit trials a positive PE signal should occur at the time of presentation of the CS, but a negative PE signal should occur at the time the reward was expected (CS omit). CS-unexpreward trials should be associated with a positive signal at the time the reward is presented.

B. Parts of human ventral striatum (top) and orbitofrontal cortex (bottom) showing a significant correlation with the temporal difference prediction error signal. Data from O'Doherty et al., (2003).

Figure 6

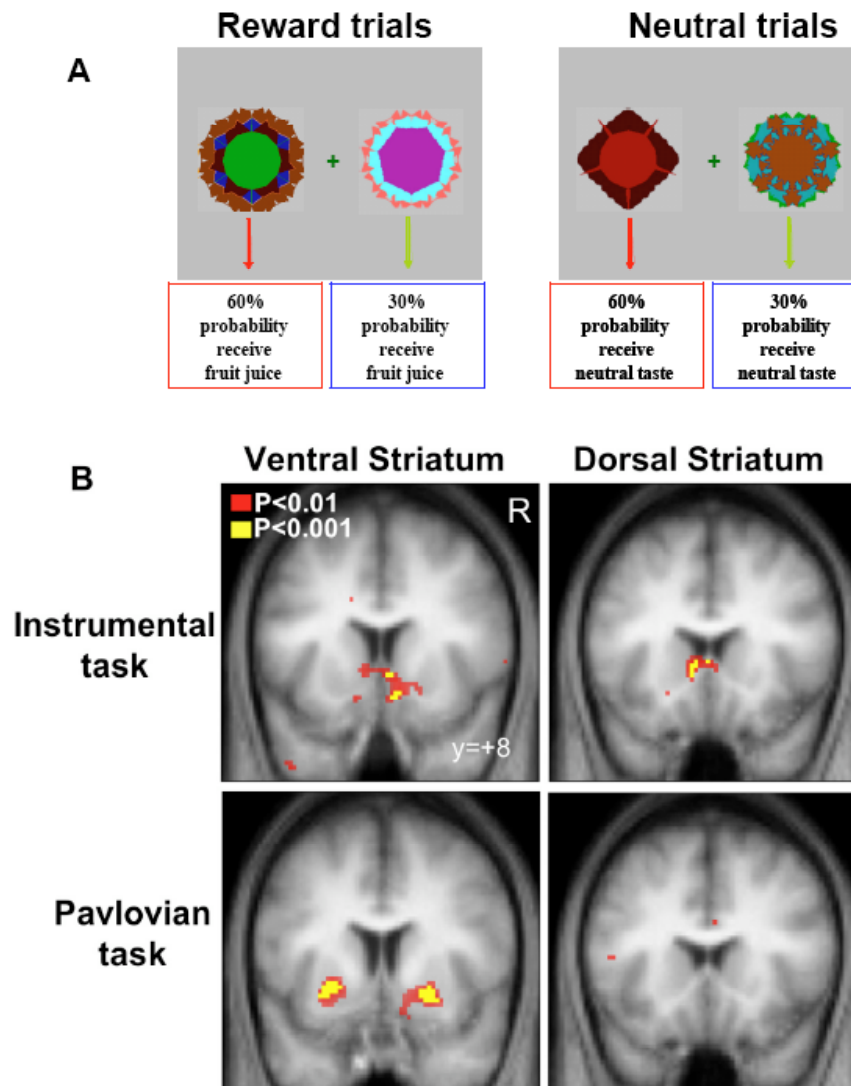


Figure 6. The actor/critic in the human striatum.

A. Instrumental choice task used by O'Doherty et al. (2004). On each trial of the reward condition subject chooses between two possible actions, one associated with a high probability of obtaining juice reward (60%), the other a low probability (30%). In a neutral condition subjects also choose between actions with similar probabilities but in this case they receive an affectively neutral outcome (tasteless solution). Prediction error responses during the reward condition of the instrumental choice task were compared to prediction error signals during a yoked Pavlovian control task.

B. Significant correlations with the reward prediction error signal generated by an actor/critic model were found in ventral striatum (ventral putamen extending into nucleus accumbens proper) in both the Pavlovian and instrumental tasks, suggesting that this region is involved in stimulus-outcome learning. By contrast a region of dorsal striatum (anteromedial caudate nucleus) was found to be correlated with prediction error signals only during the instrumental task, suggesting that this area is involved in the habitual aspects of instrumental learning. Data from O'Doherty et al, (2004).

Figure 7

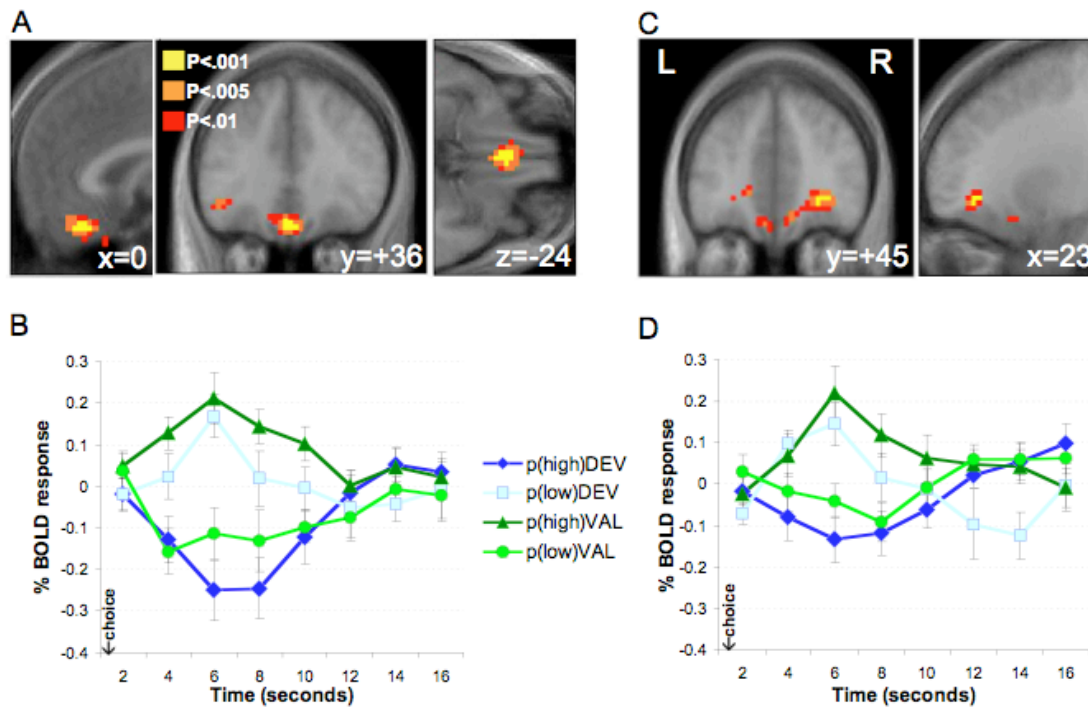


Figure 7. Regions of human OFC exhibiting response properties consistent with action-outcome learning.

Neural activity during action selection for reward in OFC showing a change in response properties as a function of the value of the outcome with each action. Choice of an action leading to a high probability of obtaining an outcome that had been devalued (p(high)DEV) led to a decrease in activity in these areas whereas choice of an action leading to a high probability of obtaining an outcome that was still valued by led to an increase in activity in the same areas. Devaluation was accomplished by means of feeding the subject to satiety on that outcome prior to the test period.

- (A) A region of medial OFC showing a significant modulation in its activity during instrumental action selection as a function of the value of the associated outcome
- (B) Time-course plots derived from the peak voxel (from each individual subject) in the mOFC during trials in which subjects chose each one of the four different actions (choice of the high vs low probability action in either the Valued or Devalued conditions).
- (C) A region of right central OFC also showing a significant interaction effect
- (D) Time-course plots from the peak voxel (from each individual subject) in the right central OFC.

Data from Valentin et al. 39. Copyright 2007 The Society for Neuroscience.