CHAPTER

15

Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning

Nathaniel D. Daw and Philippe N. Tobler

OUTLINE

Introduction	283
Learning: Prediction and Prediction Errors	283
Functional Anatomy of Dopamine and Striatum	285
Responses of Dopamine Neurons to Outcomes	287
Sequential Predictions: From Rescorla–Wagner to Temporal Difference Learning	289
Temporal Difference Learning and the Dopamine Response	293
From Error-Driven Learning to Choice	294
Conclusions	296
References	296

INTRODUCTION

This chapter provides an overview of reinforcement learning and temporal difference learning and relates these topics to the firing properties of midbrain dopamine neurons. First, we review the Rescorla–Wagner learning rule and basic learning phenomena, such as *blocking*, which the rule explains. Then we introduce the basic functional anatomy of the dopamine system and review studies that reveal a close correspondence between responses emitted by dopamine neurons and signals predicted by reinforcement learning. Finally, we introduce the generalization of the Rescorla–Wagner rule to sequential predictions as provided by temporal difference learning, and discuss its application to phasic activation changes of dopamine neurons. Subsequent chapters in this section deal with more advanced topics in reinforcement learning and presume that the reader is familiar with material covered in this chapter.

LEARNING: PREDICTION AND PREDICTION ERRORS

An important problem facing decision makers is learning, by trial and error, which decisions to make, so as best to obtain reward or to avoid punishment. In computer science, this problem is known as reinforcement learning (RL; for a more thorough introduction, see Sutton and Barto, 1998), and algorithms to accomplish it have been studied extensively. This chapter reviews the rather striking correspondence between theoretical algorithms and evidence from neuroscience and psychology about how the brain solves the RL problem. The prime correspondence between these two areas of research centers around the dopaminergic neurons of the midbrain (reviews can also be found in Glimcher, 2011; Niv, 2009; Schultz, 2007; Schultz *et al.*, 1997; Tobler, 2009).

To understand the role these neurons play, we first review research in learning, decision, and reward. We

begin with evidence from classic experiments in psychology using an experimental preparation – *classical conditioning* (also known as pavlovian conditioning) – which involves learning, but not decisions. This is an important subcomponent of the full RL problem, because choice between actions can be based on predicting how much reward they will produce.

Pavlov (1927/1960) famously exposed dogs to repeated pairings whereby an initially neutral stimulus, such as a bell, accompanied food, such as meat powder. He observed that following such training, the dogs would salivate to the sound of the bell even if it was presented without the food, by virtue of the bell's predictive relationship with the food. This *conditioned response* offers a direct window on how organisms use experience to learn to predict reward. Variations of this basic experiment have been conducted with a variety of species, from molluscs to humans, using a variety of appetitive and aversive outcomes as rewards and a variety of anticipatory behaviors as responses, and many basic phenomena are widely preserved across this range of species.

One popular view of the learning process that emerges from these experiments is that learning in classical conditioning is based on a comparison between what reward the organism experiences on a particular trial, and what reward it had expected on the basis of its previous learning (Bush and Mosteller, 1951). The difference between these two quantities is known as a *prediction error*: if the difference is large, predictions did not match observations, and there is a need for more learning to update those predictions.

More formally, assume that an animal maintains a set of predictions of the reward associated with each stimulus, s, called $V(s)$ (for value). Also assume that these predictions determine the animal's conditioned response to whichever stimulus is observed. Then upon observing stimulus s_k (e.g., the bell on trial k) and receiving a reward on that trial, r_k, the prediction error is

$$\delta_k = r_k - V_k(s_k) \tag{15.1}$$

As we will see below, this prediction error (with further refinements) appears to be carried by dopaminergic neurons (Houk *et al.*, 1995; Montague et al., 1996; Schultz et al., 1997).

The animal then updates the prediction in the direction of the prediction error, so as to reduce it. Thus, the predicted value on the next trial, $k+1$, of the stimulus s_k is:

$$V_{k+1}(s_k) = V_k(s_k) + \alpha \cdot \delta_k \tag{15.2}$$

(The value of stimuli that aren't observed remains the same, i.e. $V_{k+1}(s) = V_k(s)$, for all $s \neq s_k$.) In Equation 15.2, α is a *learning rate* parameter, between 0 and 1, which

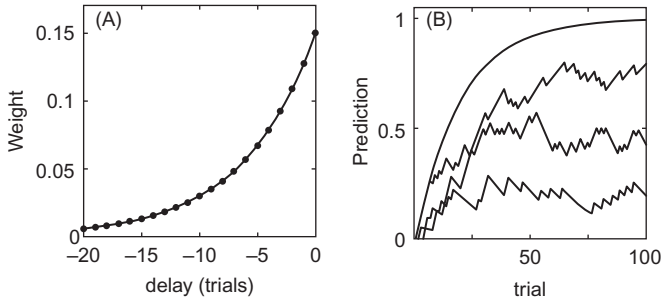

FIGURE 15.1 (A) The weights on rewards received at different past trials, according to the Rescorla/Wagner model. Weights decline exponentially into the past, with a steepness that depends on the learning rate parameter. (B) Simulation of Rescorla/Wagner model learning about four different cues, which are reinforced (from top to bottom) 100%, 75%, 50%, and 25% of the time. Learning curves grow to asymptote; for the stochastically rewarded stimuli, the prediction is noisy (driven by random patterns of reward and non-reward) around the underlying average reward.

determines the size of the update step. Its interpretation is clearer in an algebraically rearranged form of the update rule, $V_{k+1}(s_k) = (1-\alpha)V_k(s_k) + \alpha r_k$. This form reveals that the error-driven update accomplishes a weighted average between the observed reward (with weight α) and the previous reward prediction (with weight $(1-\alpha)$). Thus a larger learning rate updates the value prediction to look more like the current reward and a smaller learning rate relies more on older estimates than on the current reward.

A related way to understand this model, resulting from further algebraic manipulation, is to realize that it computes a weighted running average of all rewards received previously in the presence of the stimulus, with the most recent reward weighted most heavily and the weight for prior rewards declining exponentially in their lag. Here, the learning rate can be equivalently seen as controlling the steepness of the decay, with higher learning rates producing averages more sharply weighted toward the most recent rewards. Such an exponential pattern (Figure 15.1a) is a key hallmark of this sort of error-driven updating, which we will see verified in both behavioral and neural data later in this chapter.

Accordingly, applied to a simulated conditioning experiment (in which a bell is repeatedly paired with meat powder), the error-driven learning model described above nudges the prediction toward the observation on each trial, producing a gradual, asymptoting learning curve that ultimately predicts the actual magnitude of the average reward (Figure 15.1b). If rewards are stochastic (if meat powder is delivered based on the flip of a fair coin), then positive and negative prediction errors will be interleaved, and the net effect of all of these is that the prediction will climb

more sporadically to oscillate around the average reward (Figure 15.1b).

A further question (Rescorla and Wagner, 1972) is how animals learn stimulus–reward (for example: light–meat powder) relationships, when the experience with that stimulus is accompanied by other stimuli (the light is accompanied by a bell) that may themselves have previous reward associations. Kamin (1969) found behaviorally that such previous learning (about the bell) can attenuate (or block) new learning (about the light). Imagine that one of Pavlov's dogs has learned that a bell predicts meat powder and reliably salivates upon presentation of the bell. Now a light is presented simultaneously with the bell, and both of them are followed by meat powder. When the light is tested on its own, the dog's salivation to it is reduced (e.g., relative to a control situation in which the bell was also novel). Previous learning about the bell has blocked learning about the light's relationship with reward. The blocking phenomenon suggests that stimuli interact or compete with each other to explain the same rewards.

The Rescorla–Wagner (1972) model captures this effect by specifying that when multiple stimuli are observed (light and tone), the animal makes a single net prediction that is the sum of all of their predictions. Formally, $V_k^{net} = \sum_i V_k(s_i)$, where the sum is over all stimuli present on trial k. This leads to a net prediction error $\delta_k^{net} = r_k - V_k^{net}$, which is then used to update each of the observed stimuli as before, using Equation 15.2.

In the above example, the pre-trained bell already predicts the meat powder, whereas the added light predicts nothing initially, over and above what the bell already predicted. Thus, the sum of both predictions predicts the meat powder. Accordingly, no prediction error ensues and nothing is learned about the light even though it is reliably paired with the meat powder. According to this model (though this is not the only explanation for blocking) the blocking effect demonstrates that learning is driven by prediction errors.

The Rescorla–Wagner model successfully explained many basic learning phenomena and has made new predictions borne out by subsequent experiments. However some phenomena do not find a straightforward explanation with the Rescorla–Wagner model. One example is second-order conditioning, which is relevant here because it has an elegant explanation in terms of an elaborated model (temporal difference learning) that we introduce below, and also is closely related to important features of dopaminergic responses.

In second-order conditioning, if one stimulus (for example, a click) is consistently paired with another stimulus (the bell) that itself had previously been trained to predict reward, then the animal can learn to salivate to the click, even though the click has never itself been directly paired with reward. Such an effect is not predicted under the Rescorla–Wagner model, because the prediction error on a trial with the click and bell, but no reward, is negative. Before we treat this in greater detail, let us first consider how dopamine neurons and their target structures process reward prediction errors.

FUNCTIONAL ANATOMY OF DOPAMINE AND STRIATUM

The majority of dopamine neurons reside in the midbrain and form three cell groups, the retrorubral nucleus (RRN; cell group A8 in the rat), the substantia nigra pars compacta (SNpc; A9), and the ventral tegmental area (VTA; A10). These cell groups are contiguous, such that there are no clear boundaries between them. From these small nuclei, the dopamine neurons send widespread, ascending projections to regions such as the striatum (caudate and putamen), the amygdala and the (primarily frontal) cerebral cortex (Figure 15.2). The diffuse nature of these projections makes them well suited for broadcasting a scalar signal like Rescorla and Wagner's net prediction error.

The basal ganglia are a group of several subcortical nuclei that interact with cortex. In the striatum, dopamine axons target mostly medium spiny neurons (inset of Figure 15.2b; Freund *et al.*, 1984; Groves *et al.*, 1994), which are also the recipient neurons for the projection to striatum from cortex, the primary input to the basal ganglia. The dopamine axons make multiple synapses onto spines and shafts of one or several dendrites (Groves *et al.*, 1994). Each of the about 100,000 dopamine neurons in the macaque has an extensive and branching axon with about 500,000 synaptic and nonsynaptic release sites. As a consequence, each dopamine neuron innervates a large proportion of the 31 million striatal neurons, reflecting a strong divergence of the dopamine projection (Andén *et al.*, 1966; German *et al.*, 1988).

Dopamine neurons are electrically coupled to one another (electrical currents pass directly from cell to cell ensuring an unusually high degree of interneuronal synchrony; Grace and Bunney, 1983; Vandecasteele *et al.*, 2005), which may at least partly explain why they tend to show a homogenous response profile in electrophysiological recordings. Taken together with the divergence, this homogeneity implies that most of the target regions receive a similar message from dopamine neurons, again, consistent with the idea that they report a scalar signal, a single numerical quantity,

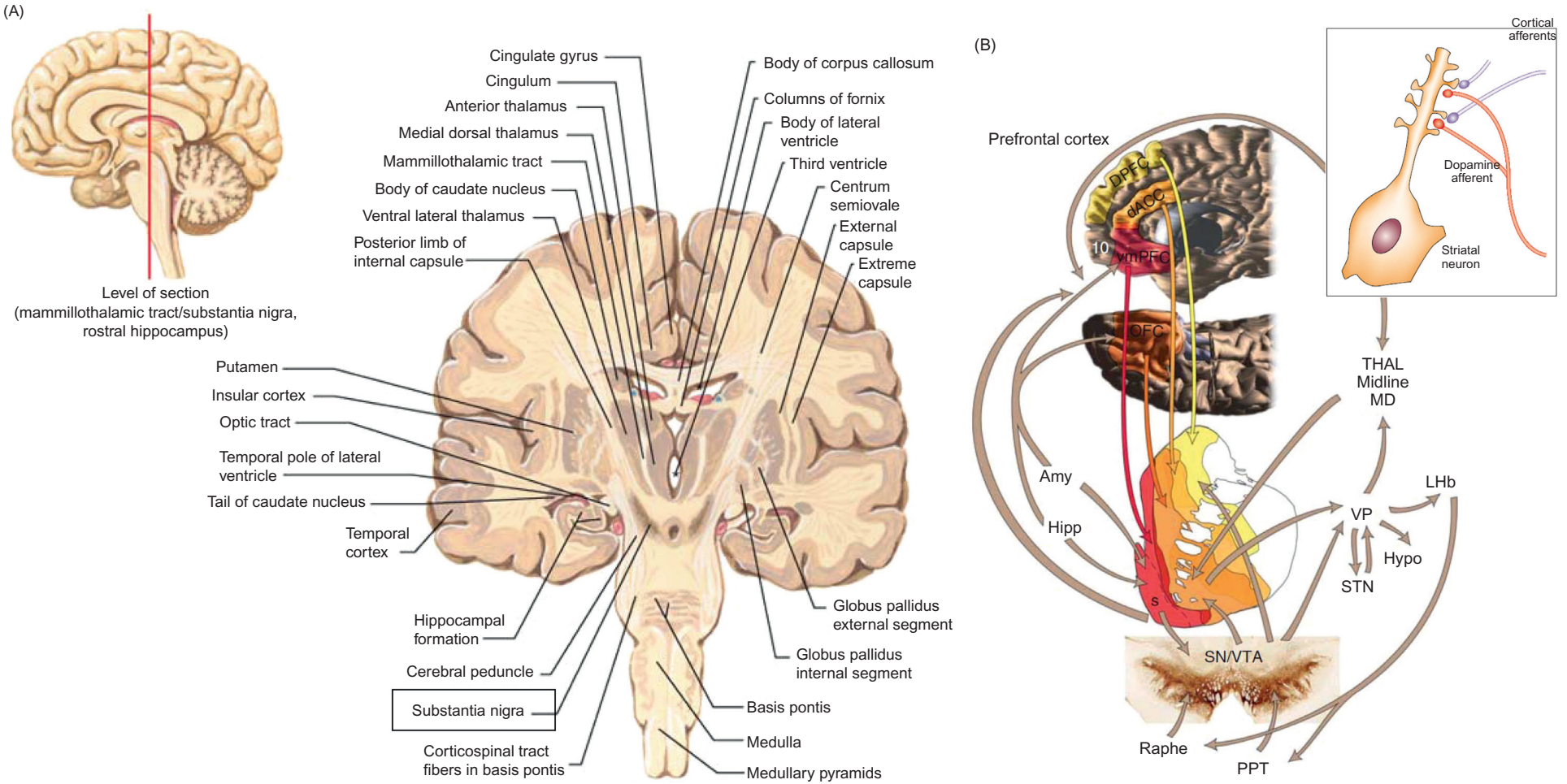

FIGURE 15.2 Anatomy and connectivity of the dopamine system. (A) Anterior–posterior location of coronal section on the right is shown in inset on the left. Dopamine neurons are located in the substantia nigra and the medially adjacent ventral tegmental area (not shown). (B) Connectivity of dopamine regions with striatum and cortex. Dopamine neurons (bottom) project to the ventral and dorsal striatum and other regions. Together with the striatum, the pallidum and the subthalamic nucleus (STN), the substantia nigra forms the basal ganglia. Inset on top right: dopamine and cortical inputs converge on neurons with spines in the striatum ("medium spiny neurons"). Abbreviations: VP, ventral pallidum; THAL, thalamus; LHb, lateral habenula; Amy, amygdala; Hipp, hippocampus. *(A) Adapted with permission from. Felten and Shetty (2009); (B) adapted from Haber and Knutson (2010); inset of (B) from Hyman and Malenka (2001).*

like a prediction error. (Note, though, that the same signal received at different areas in the brain could have different effects due to locally varying dopamine release and reuptake properties, distinct effects of dopamine on different receptors, cell types, and networks, or to differences in the other inputs to an area; see, e.g., Schultz, 2007).

What is the function of dopamine in its target regions, particularly the striatum? Over the past several decades, suggestions clustered around two key areas: on the one hand, dopamine has been hypothesized to play a role in movement control, and on the other hand, in motivation and reward. Note, though, that these two hypotheses are not necessarily mutually exclusive. On the motor side, damage to the basal ganglia produces a variety of movement impairments ranging from paralysis to tics. Parkinson's disease results from the progressive degeneration of the dopaminergic input to striatum; its symptoms primarily involve movement impairments and problems with movement initiation. Classically, these effects have been understood in terms of a simplified model of the loop-like circuitry of the basal ganglia. Neurons in the cortex project onto striatal medium spiny neurons and ultimately have those connections reciprocated, through the loops, via a series of further steps through additional basal ganglia nuclei (Alexander and Crutcher, 1990; DeLong, 1990). According to the movement control model, these loops contain different pathways that end up having either, in the net, excitatory or inhibitory effects on cortex and on the performance of movements. The shortage of dopamine in Parkinson's disease leads to an overabundance of activity in the inhibitory pathways that is seen as inhibiting movement production.

At the same time, dopamine is also tightly associated with reward and motivation, so much so that an early and influential article (the "anhedonia hypothesis;" Wise, 1982) argued that it alone was essentially the brain's *reward system*. As we will discuss below, modern accounts tend to refine this hypothesis by distinguishing different aspects of reward; for instance, rather than being involved in feelings of subjective pleasure associated with reward, dopamine is now thought to be involved in effects like reinforcement (the tendency to repeat rewarded actions; see Chapter 20). In any case, among the phenomena supporting these ideas is that essentially all major drugs of abuse act directly or indirectly via the dopamine system (reviewed in Wise, 1996). Nicotine, morphine, and ethanol all either directly or indirectly activate dopamine neurons. Cocaine and amphetamine block the dopamine reuptake mechanism and thus enhance dopamine's action in the natural synapse. In addition, amphetamine causes the release of dopamine from presynaptic terminals into the synapse. As a net effect, all these drugs lead to increased dopamine levels in the ventral striatum and other areas, and this is believed to underlie their addictive action.

How should one reconcile the two not obviously related functions – movement and reward? One key concept originated in the analysis of Mogenson and colleagues (1980) of the ventral striatum. They proposed that this structure is the interface where reward influences action. For instance, if the basal ganglia are involved in the selection of actions (a widespread hypothesis) then rewards may influence which actions are chosen through activity in these areas. In particular, reward-related signals carried by dopamine may influence action selection in the striatum, for instance by affecting plasticity there (Reynolds and Wickens, 2002) so as to reinforce rewarded actions and make them more likely to recur. This is essentially the view taken by the reinforcement learning models discussed in the present chapter.

Of course, it is not necessary that these two (or even further) functions of dopamine be completely reconciled. For instance, some of the effects of dopamine on movement, such as possibly those in Parkinson's disease, appear not to be mediated by learning of the sort described above (Gallistel *et al.*, 1974). Instead, it appears as though the overall tendency toward movement is modulated by the overall background ("tonic") level of dopamine. On this view, phasic dopamine signals would serve reinforcement learning whereas tonic dopamine levels in the striatum would facilitate movement (Schultz, 2007). It thus could be more or less coincidental that the same neurotransmitter accomplishes these two different functions at different timescales; it has also been proposed that there is a deeper explanation relating them both, a point developed in the subsequent chapter (and in Niv *et al.*, 2007).

To better understand dopamine's role in learning, we next turn to its role in basic instrumental and classical conditioning tasks.

RESPONSES OF DOPAMINE NEURONS TO OUTCOMES

Dopamine neurons recorded in behaving animals show a rather slow (about 0.1–7 Hz) baseline firing rate punctuated by phasic excitatory and inhibitory responses to a number of different sorts of events. It has been argued (Houk *et al.*, 1995; Montague *et al.*, 1996; Schultz *et al.*, 1997) that the phasic responses elicited by these events can collectively be understood as a reward prediction error similar to, but more general than (see below), the Rescorla/Wagner prediction error.

In one of the first examinations of dopamine neurons in behaving primates, the animals would perform arm movements for small pieces of food hidden within a box (Schultz, 1986). Whenever they found food in the box, a strong phasic activation occurred in the majority of the cells at around the time when the animals touched the food. This activation did not occur when non-food objects were hidden within the box. Thus, dopamine neurons respond differentially to unpredicted objects of differing reward value, which is consistent with a prediction error signal since the prediction error $\delta_k = r_k - V_k(s_k)$ will be positive when a reward is delivered (e.g., $r_k = 1$) but not expected (e.g., $V_k(s_k) = 0$).

However, to verify that the responses really reflect a prediction error (rather than, for instance, just a report of the reward itself) it is necessary to investigate whether they are systematically modulated by *predictions* as well as rewards. One way to do this is to vary the probability with which the animal expects a reward. In one such study (Fiorillo *et al.*, 2003; also described in Chapter 9), five different visual conditioned stimuli (colored images presented on a screen) predicted delivery (versus nondelivery) of liquid reward with different probabilities, ranging in steps of 0.25 from certain delivery (p = 1) to certain nondelivery (p = 0).

According to the Rescorla–Wagner model, when the animal has learned the task, the prediction $V_k(s_k)$ for each stimulus would track the average reward obtained for that stimulus – e.g., 1 for the certain reward stimulus, 0.5 for the stimulus rewarded 50% of the time, and so on. Thus the prediction error for reward delivery ($r_k = 1$ minus the prediction $V_k(s_k)$) would be zero for the always rewarded stimulus, one for the never-rewarded stimulus, and something in-between for the others. Indeed, phasic dopamine responses to a reward have this property, they increase with the size of the prediction error (or, equivalently, decrease with the degree to which the reward was expected; Figure 15.3).

Moreover, when rewards fail to occur, dopamine neurons show a phasic decrease in firing at the time reward would have been expected, consistent with the coding of negative prediction errors. In this case, the prediction error is $r_k = 0$ minus the predictions, and thus the error is negative. For these negative responses, it is harder to detect modulation of firing rate by the degree of expectation, because the background firing rate is already low. Nevertheless, on a more detailed analysis, longer inhibitions are seen when errors are more strongly negative, in accord with the prediction error model (Bayer *et al.*, 2007).

Taken together, the responses of dopamine neurons at the time of reward or non-reward are well explained by a Rescorla–Wagner prediction error. Note again that the responses are not a simple signal of reward delivery or non-delivery, because they are also modulated by expectancy. For this reason it would be incorrect to say that dopamine neurons simply encode the magnitudes of experienced rewards.

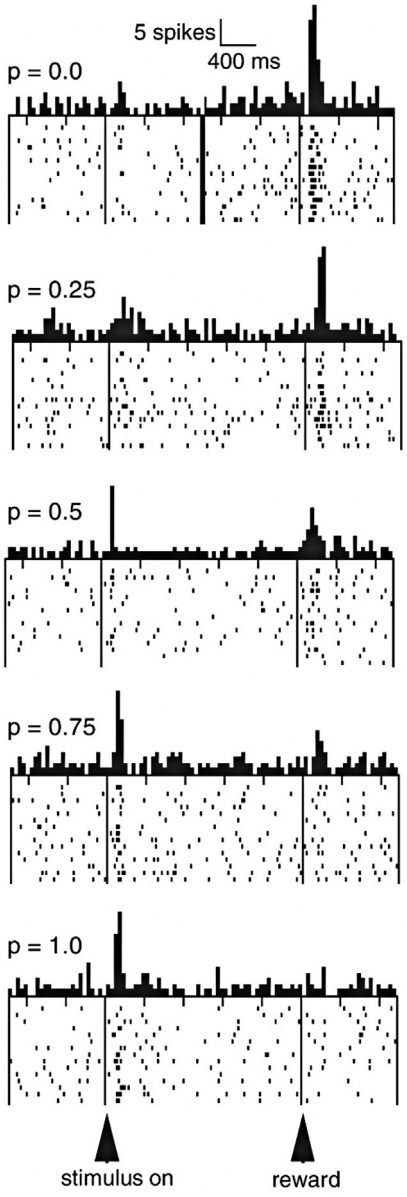

FIGURE 15.3 Peri-stimulus time histograms from a primate dopaminergic neuron in a classical conditioning experiment, reproduced from Fiorillo et al. (2003). The five traces correspond to five cues trained with stochastic reinforcement at different probabilities. Only responses on rewarded trials are shown. Top is from two different situations (separated by thick vertical black bar), with unpredictedly rewarded trials measured in situations without any preceding stimulus. *Adapted with permission from Fiorillo* et al. *(2003).*

Also, although we have so far considered just the response to the outcome at the end of a trial, as can be seen in Figure 15.3, dopamine neurons also respond to the conditioned stimuli that predict reward. These responses can be understood as another reflection of a reward prediction error, but such an understanding requires an extension of the Rescorla–Wagner model to include sequential predictions across time (see below).

The response at the time of the reward or nonreward following training in the blocking experiment described at the beginning of this chapter (Figure 15.4) further corroborates the notion that dopamine neurons code something similar to Rescorla–Wagner prediction errors (Waelti *et al.*, 2001). The absence of reward after a "blocked" stimulus does not reduce dopamine activity, in line with the notion that reward is not expected after a blocked stimulus and that its absence thus results in no prediction error. By contrast, reward delivery after a blocked stimulus elicits dopamine activity together with a positive prediction error (Figure 15.4b).

A great deal of converging evidence for this account has been reported from other recording experiments in monkeys (Bayer and Glimcher, 2005; Bayer *et al.*, 2007; Hollerman and Schultz, 1998; Kawagoe *et al.*, 2004; Matsumoto and Hikosaka, 2009; Mirenowicz and Schultz, 1994; Morris *et al.*, 2006; Nakahara *et al.*, 2004; Satoh *et al.*, 2003; Takikawa *et al.*, 2004; Tobler *et al.*, 2003, 2005), humans (Zaghloul *et al.*, 2009), mice (Cohen *et al.*, 2012) and rats (e.g., Oyama *et al.*, 2010; Roesch *et al.*, 2007). Thus, data from a variety of species suggest that dopamine plays a role that can be captured with learning models based on prediction errors.

For instance, recall that the Rescorla–Wagner model (Figure 15.1a) implies that predictions, V, are derived from the weighted average over previous rewards, with the weights exponentially declining over trials. The prediction error, in turn, is the sum of the current reward, weighted positively, and the negative expected reward (i.e., the sum over previous rewards weighted exponentially, but subtracted). Bayer and Glimcher (2005) used a task in which reward predictions shifted over time in conjunction with a regression analysis to estimate the weights that best explained the elicited, fluctuating dopamine response at the time of the reward (Figure 15.5). The weights estimated to explain dopamine responses bear an uncanny resemblance to that of a Rescorla–Wagner prediction error: they are positive for the current reward, and negative for the preceding rewards, decreasing over trials with a roughly exponential shape.

Data from other measurement techniques also corroborates the notion that dopamine neurons encode a prediction error. For instance, transient changes in dopamine concentration, reflecting dopamine release at target sites such as the striatum, can be recorded in rodents using voltammetry to detect dopamine's chemical signature. These measurements follow many of the same features of the Rescorla–Wagner prediction error (Day *et al.*, 2007).

Moreover, human fMRI experiments have shown prediction-error correlates in the striatal blood oxygen level dependent (BOLD) response resembling those seen in animal dopamine recordings (Figure 15.6), including phasic (event-related) positive and negative prediction error responses (e.g., McClure *et al.*, 2003; O'Doherty *et al.*, 2003) that scale with probability (e.g., Abler *et al.*, 2006; Spicer *et al.*, 2007; Tobler *et al.*, 2007; Chapter 9) and reflect blocking (Tobler *et al.*, 2006) and Rescorla–Wagner-like adjustments to recent rewards (Daw *et al.*, 2011). Going beyond what has been reported for dopamine neurons, the striatal BOLD signal has also been formally shown to comply with the class of reward prediction error theories using an axiomatic definition, as discussed in Chapter 1 (Rutledge *et al.*, 2010). There is evidence that dopamine modulates these hemodynamic correlates of prediction error, particularly in the striatum (e.g., Düzel *et al.*, 2009; Knutson and Gibbs, 2007; Pessiglione *et al.*, 2006; Schonberg *et al.*, 2010). However, it is worth keeping in mind that the BOLD signal is a nonspecific metabolic response and is not an unambiguous report of a particular neural event such as dopamine release.

SEQUENTIAL PREDICTIONS: FROM RESCORLA–WAGNER TO TEMPORAL DIFFERENCE LEARNING

So far, we have reviewed evidence and theory suggesting a role for dopamine in signaling prediction errors to outcomes. The models we have discussed up to here have a number of weaknesses, however. Notably, they treat learning and prediction at the level of the trial. This makes them unable to explain the temporal substructure of predictions and prediction errors during a trial, such as the responses to stimuli as well as outcomes shown in Figure 15.3. It also means that the theories only apply to experimental circumstances with a relatively simple structure: i.e., trials in which subjects observe some stimuli and receive the associated reward, after which the next trial follows independently.

One way to see that such a structure is overly limited is to recall that, although we have not yet drawn out this connection, presumably one of the reasons that the brain predicts rewards is to guide action choice toward more rewarding actions. But many

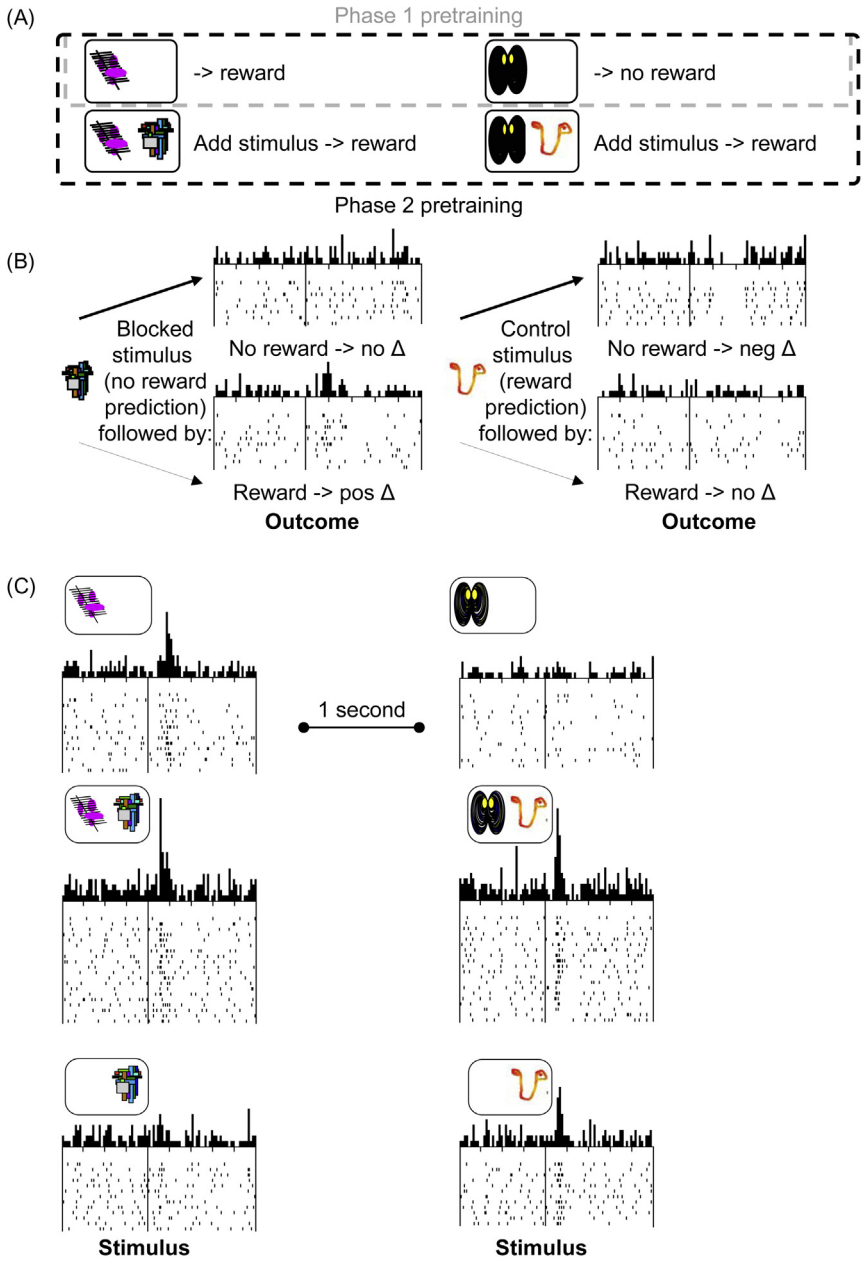

FIGURE 15.4 (A) Schematic of blocking task used with dopamine recordings (Waelti et al., 2001). In a first pretraining phase, a stimulus is paired with reward (top left) whereas a control stimulus is not (top right). Accordingly, the animal forms an association between the left stimulus and reward but not between the right stimulus and reward. In a second pretraining phase, additional stimuli are occasionally presented together with the stimuli previously learned in the first pretraining phase. Both of these compounds are followed by reward. However, according to the Rescorla/Wagner rule, the reward elicits a prediction error in the control compound on the right but not in the experimental compound on the left. This is because the added stimulus is followed by unpredicted reward in the control but not in the experimental case. In consequence, the added stimulus on the left is blocked from learning. The next panels (B, C) are from a third phase during which the added stimuli were occasionally tested on their own (interspersed with the four trial types used during the pretraining phases in order to maintain learning). (B) Outcome tests and outcome-induced responses. On top, the blocked stimulus (left) and its control (right) are both followed by no reward and the responses of a single dopamine neuron at the time of the outcome are shown. The blocked stimulus predicts nothing in particular and according to the Rescorla/Wagner rule no reward elicits no prediction error. This is reflected by the absence of any dopamine response. In contrast, the control stimulus predicts reward and the absence of such reward would elicit a negative prediction error. This is reflected by a phasic depression of the dopamine neuron. On the bottom, the blocked stimulus (left) and its reward-predicting control (right) are followed by reward. According to the rule, this would elicit a positive prediction error for the former but not the latter. Correspondingly, the neuron is activated by reward with the former but not the latter. (C) Stimulus tests and responses. After learning has been established, reward predictive stimuli (top left, middle left and right, bottom right) but not blocked stimuli (bottom left) or stimuli that are not predictive of reward (top right) elicit phasic dopamine activations, in agreement with the presence or absence of prediction errors as suggested by temporal difference learning models. *Adapted with permission from Waelti et al. (2001).*

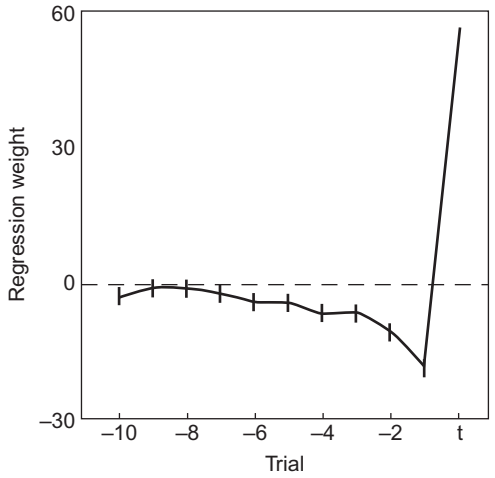

FIGURE 15.5 Average regression weights for a population of dopamine neurons. The weights were estimated to best capture the per-trial firing rate of the neurons as weighted average of rewards received on the current and previous trials. The net function corresponds to the difference between the current reward and an exponentially decaying average over previous trials' rewards, as expected for a reward prediction error. *Adapted with permission from Bayer and Glimcher (2005).*

decisions have longer term consequences than just an immediate outcome within a trial. Consider, for instance, the choice of a play in a game like American football (Romer, 2006). Here, each decision is followed by many others, and rewards (points) are earned in a way that depends on the cumulative combination of many choices in sequence. Other examples of decision tasks with similar sequential structure include driving, mazes, chess, and foraging for food.

In American football, teams must move the ball across the field to the end zone, the goal, in order to win points. But most plays don't immediately score points; instead, they change the field position of the ball, and thereby increase or decrease the chance that the team will win points on subsequent plays. The plays drive changes in the current game situation – called its *state* in reinforcement learning – notably, the field position, how many downs remain and what team has possession of the ball.

In such a situation, if we are to choose actions by predicting their consequences, then considering only the immediate reward (the points scored on a particular play), is clearly a mistake. Players must plan ahead,

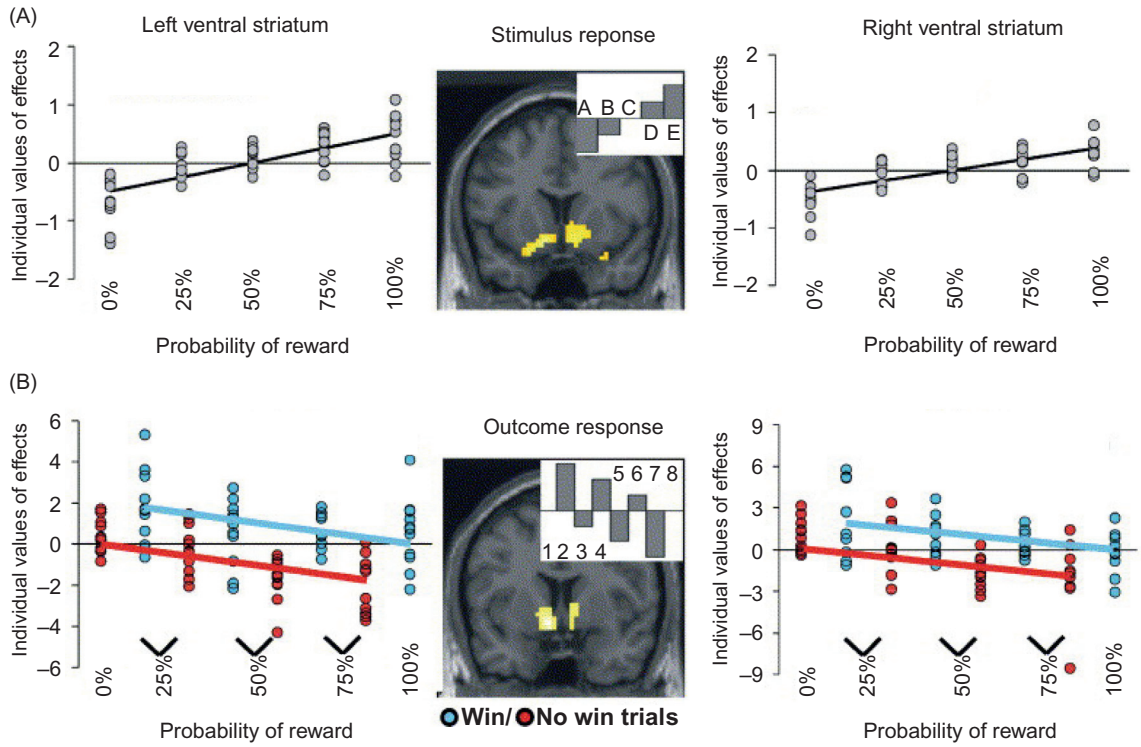

FIGURE 15.6 Graded prediction error responses in human striatum, measured using fMRI, mirror key features of responses seen in dopamine neurons in non-human primates (cf. Figure 15.3). Five different stimuli predicted reward at p = 1, 0.75, 0.5, 0.25 and 0. (A) Stimulus-induced responses. Activations induced by reward predicting stimuli increased with probability, in-line with increasing positive prediction errors. (B) Outcome-induced responses reflecting prediction errors. Rewards (blue) elicited stronger activations the more unpredicted they were. Correspondingly, omitted rewards (red) elicited stronger deactivations the more unpredicted they were. Moreover, when both outcomes were possible, rewards induced more activation than no-rewards. Dopamine neurons show the same response profile. *Adapted with permission from Abler et al. (2006).*

and try to choose actions that will position the team well to earn points on future plays. More technically, they should choose the action that maximizes the aggregate reward accumulated over the long run. Predicting this long term-quantity requires a simple modification to the Rescorla–Wagner model.

As it turns out, the ability to assess the long-run reward consequences of states and actions is tightly linked to the aspects of behavioral and neural data that the Rescorla–Wagner model failed to explain. In particular, both second-order conditioning and the dopaminergic response to reward-predictive stimuli relate to the ability to assess future reward (like points) on the basis of signals (like field position) that bear only a predictive relationship to rewards. As we will see, such a model produces prediction errors to signals – like a conditioned stimulus that predicts reward – that give an organism new information about its future reward prospects, and in this way explains dopamine responses to stimuli as well as rewards.

Let's return to prediction tasks without decisions, as in Pavlovian conditioning, to make these ideas a bit more formal. Imagine that the world proceeds stochastically through a series of states, s_t, with each state producing a (possibly zero) reward r_t, which we assume is a function of the state and so can alternately be written $r(s_t)$. A crucial difference between the Rescorla–Wagner model and the ones we are about to develop is that we are using a new variable, t, to index the progress of the experiment. Previously, we counted trials, k; henceforth, we will divide trials up into small blocks of time and use t to index the progress of time *within each trial*. Similarly, we might subdivide a football game into plays (also each indexed t and associated with a state, like field position, and reward, like points), but a team also plays many football games in a season. Thus we can think of trials, or indeed whole football games, as being encountered repeatedly, but each one made up of many sequential states.

Finally, suppose that, motivated by issues like football strategy, instead of predicting just r_k on the basis of s_k (as we did with Rescorla–Wagner), we wish to predict the sum of all future rewards in some episode, such as a football game or a conditioning trial: $r(s_t) + r(s_{t+1}) + r(s_{t+2}) + \ldots$ (Sometimes delayed rewards are treated as less valuable than immediate ones, a detail we omit here covered in Chapter 10.)

The temporal difference learning rule (Sutton, 1988) offers a way to learn such long run predictions. Define the target of learning, the "value function" $V(s)$, as the cumulative future reward expected following state s:

$$V(s_t) = r(s_t) + E[r(s_{t+1}) + E[r(s_{t+2}) + E[r(s_{t+3}) + \cdots | s_{t+2}] | s_{t+1}] | s_t] \tag{15.3}$$

Although this equation just adds up the rewards in each future state, starting at s_t, it has a rather laborious structure owing to the nested expectations $E[\cdot|s_t]$. This notation refers to the possibility of randomness in the sequence of events: for instance, the same play run at the same situation in football can lead to different outcomes. Thus in defining the expected cumulative reward, we take the expected value (probability-weighted average) over all possible values of s_{t+1}, given s_t, and over all possible s_{t+2} given each s_{t+1}, and so on.[1]

The seeming complexity of Equation 15.3 can be conquered by taking advantage of its repetitive, nested structure. In particular, let us instead write the expected future value from the perspective of the next state, s_{t+1}, as the sum of rewards starting there:

$$V(s_{t+1}) = r(s_{t+1}) + E[r(s_{t+2}) + E[r(s_{t+3}) + \cdots | s_{t+2}] | s_{t+1}] \tag{15.4}$$

But this is just the quantity inside the brackets in Equation 15.3. We can therefore substitute Equation 15.4 into Equation 15.3 to rewrite the definition of the value function in a particularly useful *recursive* form, known as the Bellman Equation (Bellman, 1957):

$$V(s_t) = r(s_t) + E[V(s_{t+1}) | s_t] \tag{15.5}$$

This equation embodies a crucial and practically useful insight. Let us restate in English what this all means. From any starting state, s_t, we are trying to predict the function V, which is the sum of the reward in that state, plus the reward in the next state, plus the reward in the state following that one, and so on. What Equation 15.5 says is just that this unwieldy, long sum over a series of rewards can equally well be thought of as the reward in the starting state, plus all the rest. Crucially, "all the rest" is just the sum over the series of rewards starting in the next state: that is, it is the value function viewed from state s_{t+1}. Equation 15.5, then, expresses the value at any state s_t as the sum of the reward there and the value of the successor state s_{t+1}. The latter value, recursively, accounts for the sum of the rest of the rewards, at s_{t+1}, s_{t+2}, and so on.

[1]Here and below we have assumed that the task belongs to a family known as *Markov processes*; each state's probability depends only on its predecessor state. This assumption is crucial for temporal difference learning because it is ultimately what allows the value function to be decomposed into the recursive form of Equation 15.5. We return to this point in the next chapter.

We can use this definition as the basis of a learning rule for estimating V from trial-and-error experience with states and rewards (Sutton, 1988; Sutton and Barto, 1998). Note that this is a difficult problem, since V at any state is defined as a long sum over future rewards. However if the value function is well learned, then if an organism encounters a state, a reward there, and a successor state, the equality in Equation 15.5 should hold, on average, for the two successive value estimates and the reward. (It is only "on average" since at a particular moment the organism experiences only one of the possible successor states s_{t+1}, whereas the expectation $E[\cdot|s_t]$ Equation 15.5 refers to the probability-weighted average over all possible successors.)

Conversely, the failure of Equation 15.5 to hold (on average) means we have not yet learned the value function. We can subtract the two sides of this equation to define an error signal expressing the extent of this mismatch in much the same way that we did to define Rescorla–Wagner's trial-based prediction error:

$$\delta_t = r_t + V(s_{t+1}) - V(s_t) \tag{15.6}$$

This is called the temporal difference prediction error. $V(s_t)$ and $V(s_{t+1})$ in this equation now refer to the *learner's own predictions* about these values. (Also note that we have switched back to the more compact notation r_t for the reward in the state s_t.) The temporal difference learning rule (Sutton, 1988) differs from the Rescorla–Wagner rule discussed above in that it uses this prediction error to update the prediction $V(s_t)$, rather than the prediction error defined by Equation 15.1 above. (The update rule itself is the same as Rescorla–Wagner's, from Equation 15.2.)

Apart from the change in the granularity of temporal indexing (from trials k to timeslices of trials t), the difference between the new model and Rescorla/Wagner's, then, is just the addition of the term $V(s_{t+1})$ to the prediction error. This reflects the desire to learn not merely the immediate reward, r_t, as in Rescorla/Wagner, but the sum over the series of all the rewards in subsequent states as well. By the recursive decomposition of the value at the current state from Equation 15.5, the subsequent state's value stands in for the sum of rewards in that state and all states thereafter.

We can take Equation 15.6 apart to better examine how temporal difference learning works. First, what does $V(s_{t+1})$ mean in this rule? During learning, the learner is maintaining a set of predictions V, one for each state, and updating them according to the learning rule. $V(s_{t+1})$ is the learner's *own current estimate* of the value of the new state s_{t+1}. That is, having observed that state s_t was followed by state s_{t+1}, the learner uses its estimated value of the new state as a proxy for the rewards remaining in the rest of the episode. Although $V(s_{t+1})$ represents a long run cumulative prediction, this recursive trick allows it to be updated immediately at every step – nudging it toward a new estimate of its true value, $r_t + V(s_{t+1})$, rather than waiting to observe all the remaining rewards in the sequence.

Now consider the interpretation of the expression $V(s_{t+1}) - V(s_t)$. This is the *temporal difference* after which the model is named: the change in predicted value from one step to the next. In many situations (as in most plays in football, where points are not scored), $r_t = 0$, and the prediction error is just the temporal difference. In these cases, if predictions are well learned (and nothing surprising is happening) the expected future value should behave smoothly. Fluctuations in V in the absence of actual reward occur when events produce changes in reward expectation, which should drive learning to update the previous expectations.

In particular, if the temporal difference is positive, this implies that the current reward expectation is better than had been anticipated in the previous state. In football, this might happen if a particularly successful play led to a field position unexpectedly close to scoring. In this case, the previous value was too pessimistic and should be increased. Importantly, in this case the outcome of a play – the new field position – taught the player something about his future reward prospects, *even though points weren't scored*. Conversely, if the temporal difference is negative, future reward expectancy has dropped, which indicates that the previous prediction was too optimistic. For instance, if a restaurant server unexpectedly clears away your wine when a few sips remain, your expected cumulative future reward has dropped by the value of those previously anticipated sips. Had the server instead spilled hot coffee on your skin, this would also produce a negative prediction error, but in this case due to a punishment, instead of a change in your expectations about future reward.

TEMPORAL DIFFERENCE LEARNING AND THE DOPAMINE RESPONSE

The key feature of the temporal difference model is that prediction errors are elicited not just by reward delivery or non-delivery, but also by any new information about future reward expectations. This is because changes in reward expectation correspond to nonzero temporal differences $V(s_{t+1}) - V(s_t)$. Returning to animal conditioning experiments, a conditioned stimulus that predicts reward changes future reward expectations, because the timing and identity of these stimuli

are themselves unpredictable. Their arrival therefore induces changes in the future rewards expected, which induce prediction errors. The arrival of a stimulus predicting reward is like an unexpectedly favorable football play: it implies that future reward prospects are better than had been expected.

This sort of reasoning explains the response pattern emitted by dopamine neurons to stimuli predicting reward with different probabilities even though at the time those stimuli are delivered no actual rewards are obtained (Fiorillo *et al.*, 2003). In Figure 15.3, the stimulus indicating the highest probability of *future* reward elicits the strongest dopamine response. With lower probabilities, responses become smaller. In the model, prediction errors are also increasing with reward probability in this same way. This is because the temporal difference $V(s_{t+1}) - V(s_t)$ on observing the stimulus is larger if the stimulus predicts reward with higher probability. Indeed, if we assume (for simplicity) that the value between trials, $V(s_t)$ is zero, then since the reward r_t is also zero when the stimulus is delivered, the prediction error from Equation 15.6 is just $V(s_{t+1})$, the value of the stimulus. Note that for the same reason, the temporal difference model doesn't change our previous account of the response to the terminal reward in the trial, since here, $V(s_{t+1})$ is the value between trials, i.e., zero, and Equation 15.6 reduces to Equation 15.1. *In all, the temporal difference rule explains not only the reward but also the stimulus responses in Figure 15.3 as reflecting prediction errors.*

Dopamine responses to stimuli in a blocking experiment are similarly consistent with the temporal difference model. A blocked stimulus elicits much less of a response at the time of the conditioned stimulus than a non-blocked, reward-predicting control stimulus (Waelti *et al.*, 2001; Figure 15.4c). This reflects the fact that the blocked stimulus doesn't predict reward, but the control stimulus does. Note however, that if the newly introduced blocked stimulus were instead slightly moved in time so as to precede the stimulus previously paired with reward, then temporal difference learning predicts (and experiments confirm) that it should in this case acquire predictive reward value. Conversely, the Rescorla–Wagner rule is not sensitive to the relative timing of events in a trial, since it is trial-based.

This last observation relates to the fact that in the temporal-difference learning model, stimuli induce prediction errors when, and only when, they cause a change in reward expectations, i.e., when they provide new information. For instance, when one visual stimulus reliably predicts another one, which in turn reliably predicts reward, then only the first but not the second stimulus adds new information about the future. Accordingly, dopamine neurons are activated only by the first but not the second stimulus (Schultz *et al.*, 1993). Conversely, when a second stimulus adds additional information, it does engender prediction error. Thus, when a 25% predictor of reward is followed by either a stimulus predicting reward at 100% (positive prediction error) or another stimulus predicting at 0% (negative prediction error), then the second stimulus activates or depresses dopamine neurons, respectively (Takikawa *et al.*, 2004).

Striatal BOLD correlates of prediction errors in human neuroimaging also appear to report a full temporal difference prediction error, similar to dopamine neurons. Thus striatal BOLD responds to conditioned stimuli according to their reward probability (e.g., Abler *et al.*, 2006; Figure 15.6a; see also Chapter 9). Blocked conditioned stimuli elicit a weaker striatal BOLD response than non-blocked control stimuli (Tobler *et al.*, 2006). Moreover, striatal BOLD responses occur to stimuli predicting points worth money, suggesting higher-order conditioning (e.g., Tobler *et al.*, 2007). Within-trial prediction errors to stimuli providing new value information have even been described in the human striatum (Daw *et al.*, 2011; Seymour *et al.*, 2004).

Finally, the temporal difference model also clears up a behavioral puzzle we noted with the Rescorla/Wagner model: the source of second-order conditioning. As we have just discussed, cues that predict future reward elicit reward prediction errors and activate dopamine (via a positive temporal difference $V(s_{t+1}) - V(s_t)$), in just the same way as unexpected primary rewards do. This error can in turn train positive reward predictions in preceding states, even if primary reward is not subsequently delivered. In this way, the temporal difference algorithm and its proposed dopaminergic implementation explain second-order conditioning – i.e. the transfer of value from one conditioned stimulus to another – as a direct reflection of their recursive learning strategy for training previous reward predictions on the basis of subsequent ones.

FROM ERROR-DRIVEN LEARNING TO CHOICE

We began this chapter with the problem of learning which action to choose, but so far we have talked only about learning to predict rewards. The connection between the two is simple: if a decision maker can predict the reward following a choice – either in one step, like Rescorla–Wagner, or cumulatively over multiple steps, like temporal difference learning – then she can choose the more rewarding action. In other words, in decisions by description ("would you rather have a 50% chance at $100, or $40 for sure") a decision maker

computes a decision variable for each option and chooses between them. In a trial and error ("experiential") learning situation, she must instead learn the decision variable, and this is exactly what the error-driven learning rules we have described can accomplish.

But is there evidence that learned predictions drive choices in the way we have described? And if dopamine carries prediction errors that drive learning about reward predictions, is it causally involved in choice?

Consider an experiment in which a monkey repeatedly chooses between a red and a green target, and receives juice reward stochastically on the basis of its choice. One way to approach this sort of task, drawing on the prediction mechanisms described so far, is to learn a predicted value $Q(a)$ for the reward expected following the choice a of either option. Q here is analogous to V previously, but it is traditional to use different notation to distinguish action- from stimulus-specific values.[2] The Qs can be learned by the Rescorla–Wagner rule (Equations 15.1, 15.2), updating an option's value according to the prediction error received whenever it is chosen.

It is possible to examine whether animals actually learn their decision variables in such a manner by using a regression analysis similar to Bayer and Glimcher's (Figure 15.5) dissection of the dopamine response (Lau and Glimcher, 2005). If animals choose on the basis of value predictions Q for each option, and these are learned as some weighted average of the rewards previously received on that option, then one can estimate what weights best explain the choices. In particular, if these are learned by the same sort of error-driven learning rule associated with Rescorla/Wagner and the phasic dopamine response, the model predicts an exponential function (Figure 15.1). This prediction has been confirmed in two studies of monkey decisions (Sugrue *et al.*, 2004, data in Figure 15.7 here; Lau and Glimcher, 2005). Similar results have been reported for choice experiments with humans (e.g., Seymour *et al.*, 2012) and rodents (Ito and Doya, 2009).

All these considerations suggest that predictive value learning underlying choice is also based on an error driven mechanism, of the sort associated with phasic dopamine responses. Dopamine is also well positioned to drive such learning from an anatomical point of view, because it affects plasticity at synapses from cortical neurons onto the medium spiny neurons of striatum (Reynolds and Wickens, 2002). The corticostriatal connections, as we have already mentioned, are also involved in facilitating and suppressing movements. Combining the two elements, as described in this form by Frank *et al.* (2004), goes some way toward resolving the central puzzle of dopamine with which this section began – its dual roles in motivation and movement – and fleshing out the suggestion of Mogenson and colleagues (1980) a limbic-motor gateway.

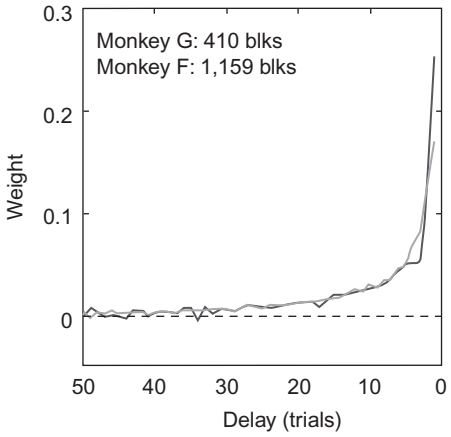

FIGURE 15.7 Functions relating choices and previous rewards in a decision task for two monkeys. These roughly follow an exponentially decaying form, consistent with an error-driven learning model. *Adapted with permission from Sugrue* et al. *(2004).*

To assess the *causal* role of dopamine in reinforcement learning, Frank and colleagues (2004) studied learning to choose or avoid actions in human patients with Parkinson's disease playing a reinforcement learning game similar to the monkey task of Figure 15.7. By training subjects to develop preferences between several different pairs of options before testing these options against one another in novel transfer pairings, the researchers were able to distinguish to what extent the preferences learned were based on

[2]To expand on this notational point, RL distinguishes state values $V(s)$ from state-action values $Q(s, a)$. We previously considered tasks like Pavlovian conditioning, in which stimuli (states) were followed by rewards or other stimuli, and in this case we defined the expected value of the state, $V(s)$ as the expected (cumulative) reward following it. If states and rewards also depend on the agent's decisions – which is not true in Pavlovian conditioning but is true in football – then to choose an action we want to learn $Q(s, a)$, the value of an action (e.g., passing) in a state (first down on the 50-yard line). In simple experiments where a monkey faces the same choice over and over again for immediate reward (e.g., choice between a red and green target) then there is only one state, and we abbreviate the state-action values $Q(a)$. Finally, state values $V(s)$ are still relevant even in decision tasks, because $Q(s, a)$ is reduced back to $V(s)$ once we determine a particular *policy* or choice of action for each state – for instance, if at any state I choose the action for which $Q(s, a)$ is maximal.

learning to favor the better option versus avoiding the worse one. The experimenters reasoned that positive dopamine action (reporting positive prediction errors) would promote learning to choose the better action, and therefore that Parkinson's patients (since the disease degenerates dopamine neurons) would tend to learn the tasks, if at all, primarily via avoidance. Accordingly, when patients with Parkinson's disease were tested off their dopamine-restoring medication, they tended toward learning to avoid the inferior action; but when tested on dopamine replacement therapy, this pattern reversed and medicated patients tended toward learning to choose the better action. Similar effects of Parkinson's disease and its medication have now been reported from a number of labs using different reinforcement learning tasks (Bodi *et al.*, 2009; Cools *et al.*, 2006; Rutledge *et al.*, 2009).

Animal experiments using optogenetic activation of dopamine neurons also provide causal support for a role of dopamine in learning about actions and choosing accordingly. Animals learn to return to locations where their dopamine neurons have been activated (Tsai *et al.*, 2009), prefer a lever that provides both food and stimulation of dopamine neurons to one providing only food (Adamantidis *et al.*, 2011), nose poke for phasic stimulation of dopamine neurons (Kim *et al.*, 2012), and avoid locations where their dopamine neurons have been inhibited (Tan *et al.*, 2012). Also, of course, the reinforcing action of drugs of abuse (which pharmacologically activate, mimic, or otherwise enhance dopaminergic function) is consistent with such a causal role of dopamine in learning mechanisms (Redish, 2004).

The learning mechanisms described also follow on from the classic idea from psychology (Thorndike, 1911; see Chapter 20) that trial-and-error learning occurs by the *reinforcing* action of reward, i.e., that actions followed by reward are more likely to be repeated in the future. In the models described above, reinforcement is not reward per se, but reward prediction error: Actions followed by positive reward prediction error are strengthened, and more likely to be repeated in the future. This idea is also the basis of a version of temporal difference algorithms, called actor-critic methods, which involve separate learning methods for long run state values $V(s)$ (how much reward is expected in the future), and which action to take in each state. Prediction errors computed by the former, called the critic, serve as reinforcers to help the other module, the actor, learn which actions to take. This is particularly useful in the context of sequential decision tasks, like football, in which the ultimate rewarding consequences of an action can be deferred by many steps. The prediction error related to arriving, for instance, at a better-than-expected field position can reinforce a good choice immediately, even if scoring occurs only later.

Finally, then, the choice experiments mentioned thus far involve only a series of independent, isolated choices, each with its own reward. But we have stressed that the temporal difference learning model associated with dopamine is well suited to sequential decision tasks like football, involving multiple, interleaved choices and rewards. In this case, we expect a similar strategy of learning the long run values $Q(s, a)$ of actions in particular states (e.g., the value of running or passing at different field positions) via temporal difference or actor-critic methods, and choosing on this basis. Choices in sequential decision tasks are consistent with such a mechanism (Daw *et al.*, 2011; Fu and Anderson, 2008), though (as discussed in Chapter 20) the same experiments provide evidence that this mechanism is nonexclusive and organisms also pursue additional strategies for solving the sequential decision-making problem.

CONCLUSIONS

Electrophysiological recordings from dopamine neurons suggest that phasic activity changes contribute to reward learning by coding errors in the prediction of reward. In this way, dopamine neurons may provide target neurons in the striatum and cortex with detailed information about the value of the future. Such information could be used to plan and execute profitable behaviors and decisions well in advance of actual reward occurrence and to learn about even earlier reliable predictors of reward. Moreover, the notion that phasic dopamine actions can be described with a reward prediction error model captures empirical findings not only from electrophysiological recordings in monkeys, rats and humans but also from other modalities, such as voltammetry and human neuroimaging. Importantly, decision behavior in learning tasks is consistent with the proposed mechanism, and causal experiments involving manipulation of dopamine support this role. Although prediction error coding is probably not the only function of dopamine neurons, it provides a good approximation to much of its phasic activity. The next chapter extends this core hypothesis to consider more detailed computations and how these mechanisms interact with other brain systems.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., Spitzer, M., 2006. Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage. 31, 790–795.

Adamantidis, A.R., Tsai, H.C., Boutrel, B., Zhang, F., Stuber, G.D., Budygin, E.A., et al., 2011. Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. J. Neurosci. 31, 10829–10835.

Alexander, G.E., Crutcher, M.D., 1990. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. 13, 266–271.

Andén, N.E., Fuxe, K., Hamberger, B., Hökfelt, T., 1966. A quantitative study on the nigro-neostriatal dopamine neurons. Acta Physiol. Scand. 67, 306–312.

Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 47, 129–141.

Bayer, H.M., Lau, B., Glimcher, P.W., 2007. Statistics of midbrain dopamine neuron spike trains in the awake primate. J. Neurophysiol. 98, 1428–1439.

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Bodi, N., Keri, S., Nagi, H., Moustafa, A., Myers, C., Daw, N.D., et al., 2009. Reward learning and the novelty seeking personality: a between and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. Brain. 2132, 2385–2395.

Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. Psychol. Rev. 58, 313–323.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature. 482, 85–88.

Cools, R., Altamirano, L., D'Esposito, M., 2006. Reversal learning in Parkinson's disease depends on medication status and outcome valence. Neuropsychologia. 44, 1663–1673.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron. 69, 1204–1215.

Day, J.J., Roitman, M.F., Wightman, R.M., Carelli, R.M., 2007. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat. Neurosci. 10, 1020–1028.

DeLong, M.R., 1990. Primate models of movement disorders of basal ganglia origin. Trends Neurosci. 13, 281–285.

Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B.H., Tobler, P.N., 2009. Functional imaging of the human dopaminergic midbrain. Trends Neurosci. 32, 321–328.

Fiorillo, C.D., Tobler, P.N., Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 299, 1898–1902.

Felten, D., Shetty, A., 2009. Netter's Atlas of Neuroscience. Philadelphia Saunders-Elsevier.

Frank, M.J., Seeberger, L.C., O'Reilly, R.C., 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 306, 1940–1943.

Freund, T.F., Powell, J.F., Smith, A.D., 1984. Tyrosine hydroxylase-immunoreactive boutons in synaptic contact with identified striatonigral neurons, with particular reference to dendritic spines. Neuroscience. 13, 1189–1215.

Fu, W.T., Anderson, J.R., 2008. Solving the credit assignment problem: explicit and implicit learning of action sequences with probabilistic outcomes. Psychol. Res. 72, 321–330.

Gallistel, C.R., Stellar, J., Bubis, E., 1974. Parametric analysis of brain stimulation reward in the rat: I. The transient process and the memory-containing process. J. Comp. Physiol. Psychol. 87, 848–860.

German, D.C., Dubach, M., Askari, S., Speciale, S.G., Bowden, D.M., 1988. 1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine-induced parkinsonian syndrome in Macaca fascicularis: which midbrain dopaminergic neurons are lost? Neuroscience. 24, 161–174.

Glimcher, P.W., 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. 108 (Suppl. 3), 15647–15654.

Grace, A.A., Bunney, B.S., 1983. Intracellular and extracellular electrophysiology of nigral dopaminergic neurons–3. Evidence for electrotonic coupling. Neuroscience. 10, 333–348.

Groves, P.M., Linder, J.C., Young, S.J., 1994. 5-hydroxydopamine-labeled dopaminergic axons: three-dimensional reconstructions of axons, synapses and postsynaptic targets in rat neostriatum. Neuroscience. 58, 593–604.

Haber, S.N., Knutson, B., 2010. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology. 35, 4–26.

Hollerman, J.R., Schultz, W., 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. 1, 304–309.

Houk, J.C., Adams, J.L., Barto, A.G., 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of information processing in the basal ganglia. MIT Press, Boston, pp. 249–270.

Hyman, S.E., Malenka, R.C., 2001. Addiction and the brain: the neurobiology of compulsion and its persistence. Nat. Rev. Neurosci. 2, 695–703.

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

Kamin, L.J., 1969. Predictability, surprise, attention and conditioning. In: Campbell, B.A., Church, R.M. (Eds.), Punishment and Aversive Behavior. Appleton-Century-Crofts, New York, pp. 279–296.

Kawagoe, R., Takikawa, Y., Hikosaka, O., 2004. Reward-predicting activity of dopamine and caudate neurons–a possible mechanism of motivational control of saccadic eye movement. J. Neurophysiol. 91, 1013–1024.

Kim, K.M., Baratta, M.V., Yang, A., Lee, D., Boyden, E.S., Fiorillo, C.D., 2012. Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS One. 7, e33612.

Knutson, B., Gibbs, S.E.B., 2007. Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology. 191, 813–822.

Lau, B., Glimcher, P.W., 2005. Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anal. Behav. 84, 555–579.

Matsumoto, M., Hikosaka, O., 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 459, 837–841.

McClure, S.M., Berns, G.S., Montague, P.R., 2003. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 38, 339–346.

Mirenowicz, J, Schultz, W., 1994. Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. 72, 1024–1027.

Mogenson, G.J., Jones, D.L., Yim, C.Y., 1980. From motivation to action: functional interface between the limbic system and the motor system. Prog. Neurobiol. 14, 69–97.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., Bergman, H., 2006. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 9, 1057–1063.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., Hikosaka, O., 2004. Dopamine neurons can represent context-dependent prediction error. Neuron. 41, 269–280.

Niv, Y., 2009. Reinforcement learning in the brain. J. Math. Psychol. 53, 139–154.
Niv, Y., Daw, N.D., Joel, D., Dayan, P., 2007. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology. 191, 507–520.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron. 38, 329–337.
Oyama, K., Hernádi, I., Iijima, T., Tsutsui, K., 2010. Reward prediction error coding in dorsal striatal neurons. J. Neurosci. 30, 11447–11457.
Pavlov, I.P., 1927. Conditional Reflexes. Dover Publications, New York (The 1960 edition is an unaltered republication of the 1927 translation by Oxford University Press).
Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., Frith, C.D., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 442, 1042–1045.
Redish, A.D., 2004. Addiction as a computational process gone awry. Science. 306, 1944–1947.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton Century Crofts, New York, pp. 64–99.
Reynolds, J.N., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15, 507–521.
Roesch, M.R., Calu, D.J., Schoenbaum, G., 2007. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. 10, 1615–1624.
Romer, D.H., 2006. Do firms maximize? Evidence from professional football. J. Pol. Econ. 114, 340–365.
Rutledge, RB, Dean, M, Caplin, A, Glimcher, PW., 2010. Testing the reward prediction error hypothesis with an axiomatic model. J. Neurosci. 30, 13525–13536.
Rutledge, R.B., Lazzaro, S.C., Lau, B., Myers, C.E., Gluck, M.A., Glimcher, P.W., 2009. Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J. Neurosci. 29, 15104–15114.
Satoh, T., Nakai, S., Sato, T., Kimura, M., 2003. Correlated coding of motivation and outcome of decision by dopamine neurons. J. Neurosci. 23, 9913–9923.
Schonberg, T., O'Doherty, J.P., Joel, D., Inzelberg, R., Segev, Y., Daw, N.D., 2010. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. Neuroimage. 49, 772–781.
Schultz, W., 1986. Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. J. Neurophysiol. 56, 1439–1461.
Schultz, W., 2007. Multiple dopamine functions at different time courses. Annu. Rev. Neurosci. 30, 259–288.
Schultz, W., Apicella, P., Ljungberg, T., 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J. Neurosci. 13, 900–913.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science. 275, 1593–1599.
Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., et al., 2004. Temporal difference models describe higher-order learning in humans. Nature. 429, 664–667.
Seymour, B., Daw, N.D., Roiser, J.P., Dayan, P., Dolan, R., 2012. Serotonin selectively modulates reward value in human decision-making. J. Neurosci. 32, 5833–5842.
Spicer, J., Galvan, A., Hare, T.A., Voss, H., Glover, G., Casey, B., 2007. Sensitivity of the nucleus accumbens to violations in expectation of reward. Neuroimage. 34, 455–461.
Sugrue, L.P., Corrado, G.S., Newsome, W.T., 2004. Matching behavior and the representation of value in the parietal cortex. Science. 304, 1782–1787.
Sutton, R.S., 1988. Learning to predict by the method of temporal difference. Mach. Learn. 3, 9–44.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Takikawa, Y., Kawagoe, R., Hikosaka, O., 2004. A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. J. Neurophysiol. 92, 2520–2529.
Tan, K.R., Yvon, C., Turiault, M., Mirzabekov, J.J., Doehner, J., Labouèbe, G., et al., 2012. GABA neurons of the VTA drive conditioned place aversion. Neuron. 73, 1173–1183.
Thorndike, E.L., 1911. Animal Intelligence: Experimental Studies. Macmillan, New York.
Tobler, P.N., 2009. Behavioral functions of dopamine neurons. In: Björklund, A., Dunnett, S., Iversen, L., Iversen, S. (Eds.), Dopamine Handbook. Oxford University Press, Oxford, pp. 316–330.
Tobler, P.N., Dickinson, A., Schultz, W., 2003. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. J. Neurosci. 23, 10402–10410.
Tobler, P.N., Fiorillo, C.D., Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. Science. 307, 1642–1645.
Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W., 2006. Human neural learning depends on reward prediction errors in the blocking paradigm. J. Neurophysiol. 95, 301–310.
Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W., 2007. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J. Neurophysiol. 97, 1621–1632.
Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., et al., 2009. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science. 324, 1080–1084.
Vandecasteele, M., Glowinski, J., Venance, L., 2005. Electrical synapses between dopaminergic neurons of the substantia nigra pars compacta. J. Neurosci. 25, 291–298.
Waelti, P., Dickinson, A., Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. Nature. 412, 43–48.
Wise, R.A., 1982. Neuroleptics and operant behavior: the anhedonia hypothesis. Behav. Brain Sci. 5, 39–53.
Wise, R.A., 1996. Neurobiology of addiction. Curr. Opin. Neurobiol. 6, 243–251.
Zaghloul, K.A., Blanco, J.A., Weidemann, C.T., McGill, K., Jaggi, J.L., Baltuch, G.H., et al., 2009. Human substantia nigra neurons encode unexpected financial rewards. Science. 323, 1496–1499.