CHAPTER

22

c00022

Theoretical and Empirical Studies of Learning

Yael Niv and P. Read Montague

OUTLINE

INTRODUCTION

s0010

p0010 One fundamental goal of behavioral neuroscience is to understand the decision-making processes that animals and humans use in order to select actions in the face of reward and punishment. Recent research has pursued this goal from a variety of perspectives. In behavioral psychology, this question has been investigated through the paradigms of Pavlovian (classical) and instrumental (operant) conditioning, and much evidence has accumulated regarding the learned associations that control simple behaviors. Simple conditioning paradigms also form the backbone of neurobiological approaches to learning, where investigators seek underlying neural mechanisms.

p0020 From a computational perspective, Pavlovian (passive) conditioning is modeled as *prediction learning* – learning the predictive (and sometimes causal) relationships between different events in the environment, such as the fact that the smell of food usually predicts that a tasty meal is forthcoming. Instrumental conditioning, on the other hand, involves learning to select actions that will bring about rewarding events and avoid aversive events. Computationally, such decision-making is treated as attempting to *optimize* the consequences of actions. Thus, from an economic perspective, the study of instrumental conditioning is an inquiry into perhaps the most fundamental form of rational decision-making.

Computational accounts of conditioned behavior have, in recent years, drawn heavily from a class of

models called *reinforcement learning* (RL) models. These models, now commonly used in neurobiology, psychology, and machine learning, all share in common the use of a *scalar reinforcement signal* to direct learning. This is categorically different from, and lies in between, learning from an explicit teaching signal (as in "supervised learning" models, common in artificial intelligence applications), and learning of input statistics without any supervisory signal (as in "unsupervised learning" models, for instance of early visual processing; Lewicki and Olshausen, 1999). Thus algorithms and theory have been developed specifically for the case of RL.

p0030 Importantly, reinforcement learning provides a *normative framework* within which conditioning can be analyzed. That is, it suggests a means by which optimal prediction and action selection can be achieved, and exposes explicitly the computations that must be realized in the service of these. In contrast to descriptive models that describe behavior as it is, normative models study behavior from the point of view of its hypothesized *function* – that is, they study behavior *as it should be* if it were to accomplish specific goals in an optimal way. The appeal of normative models derives from several sources. First, because throughout evolution animal behavior has been shaped and constrained by its influence on fitness, one reasonable starting point is to view a particular behavior as an optimal or near-optimal adaptation to some set of problems (Kacelnik, 1997). Treating behavior as optimal allows for the generation of computationally explicit hypotheses that are directly testable. Second, discrepancies between observed behavior and the predictions of normative models are often illuminating: these can shed light on the neural and/or informational constraints under which animals make decisions, or suggest that animals are, in fact, optimizing something other than what the model has assumed. Such approaches are familiar to economists. In economics, it is frequently assumed that a decision-maker is rational (even though we know people are not), and deviations from rationality are used to devise progressively more accurate theories of human decision-making.

p0040 Finally, as has been recently the case for economic constructs as well, the relevance of reinforcement learning models to human and animal decision-making has been strengthened by measurements of neural correlates of some of the major constituents of these models. Specifically, extracellular recordings in behaving animals and functional imaging of human decision-making have revealed in the brain the existence of a key reinforcement learning signal, the *temporal difference reward-prediction error*.

p0050 In this chapter, we introduce the formal reinforcement learning framework and give a brief background to the origins and history of reinforcement learning models of decision-making (for a comprehensive textbook account of RL methods, see Sutton and Barto, 1998). In the second section, we review the multiple lines of evidence linking reinforcement learning to the function of dopaminergic neurons in the mammalian midbrain. These data demonstrate the strength of the computational model and normative framework for interpreting and predicting a wide range of (otherwise confusing) neural activity patterns. The third section extends these results to more recent data from human imaging experiments. In these experiments, the combination of reinforcement learning models of choice behavior and online imaging techniques has allowed researchers to detect in the brain the existence of specific "hidden variables" controlling behavior (such as the subjective value of different options). Prior to this work such variables could only be postulated, or worse, were presumed to have descriptive ("as if") status alone. Together, the results from the latter-two sections put on firmer ground concepts central to neuroeconomics – for example, expected utility in units of a common currency and temporal discounting. In the fourth section, we discuss aspects of learning not associated with phasic dopamine signals, such as learning about the vigor (or rate) with which actions should be performed (which seems to be related more to tonic levels of dopamine in the striatum), and adapting learning rates to the natural statistics of the learning problem (which has been associated with the functions of the neuromodulators acetylcholine and norepinephrine). The final section discusses some of the fundamental limitations of the current theories of learning, and highlights questions for future research.

REINFORCEMENT LEARNING: THEORETICAL AND HISTORICAL BACKGROUND

s0020

p0060 Historically, the core ideas in reinforcement learning arose from two separate and parallel lines of research. One axis is mainly associated with Richard Sutton, formerly a psychology major, and his PhD advisor, Andrew Barto, a computer scientist. Interested in artificial intelligence and agent-based learning, Sutton and Barto developed algorithms for reinforcement learning that were inspired by the psychological literature on Pavlovian and instrumental conditioning (Sutton, 1978; Barto *et al.*, 1983; Sutton and Barto, 1990, 1998). Stemming from a different background, the second axis was led by electrical engineers such as Dimitri Bertsekas and John Tsitsiklis. Within the fields

of operations research and optimal control, Bertsekas and Tsitsiklis developed stochastic approximations to dynamic programming (which they termed "neuro-dynamic programming") that led to similar reinforcement learning rules (e.g. Bertsekas and Tsitsiklis, 1996). The fusion of these two lines of research reinterpreted the behaviorally-inspired and somewhat heuristically-derived reinforcement learning algorithms in terms of optimality, and provided tools for analyzing their convergence properties in different situations.

s0030

The Rescorla-Wagner Model

p0070

The early impetus for the artificial intelligence trajectory can be traced back to the behaviorist movement in psychology in the early twentieth century. Behaviorism played an important "rigorizing" role for the kinds of experiments and data that psychologists would come to find acceptable, but was largely a-theoretic. This gave rise to the field of "mathematical psychology" in the 1950s, within which statistical models of learning were considered for the first time. In a seminal paper that helped to establish this field, Bush and Mosteller (1951) developed one of the first detailed formal accounts of learning. Together with Kamin's (1969) idea that learning should occur only when outcomes are "surprising," the Bush and Mosteller "linear operator" model finds its most popular expression in the now-classic Rescorla-Wagner model of Pavlovian conditioning (Rescorla and Wagner, 1972). The Rescorla-Wagner model, arguably the most influential model of animal learning to date, explained the puzzling behavioral phenomena of blocking, overshadowing, and conditioned inhibition by postulating that learning occurs *only when events violate expectations*. For instance, in a conditioning trial in which *conditional stimuli* CS_1 and CS_2 (say, a light and a tone) were presented, as well as an affective stimulus such as food or a tail-pinch (termed the *unconditional stimulus*), Rescorla and Wagner postulated that the associative strength of each of the conditional stimuli $V(CS_i)$ will change according to

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta(CS_i, US)\left[\lambda(US) - \sum_i V_{old}(CS_i)\right] \qquad (22.1)$$

p0080

In this *error-correcting* learning rule, learning is driven by the discrepancy between what was predicted ($\Sigma_i V(CS_i)$ where i indexes all the CSs present in the trial) and what actually happened ($\lambda(US)$, whose magnitude is related to the worth of the unconditional stimulus). $\lambda(US)$ quantifies the maximal associative strength that the unconditional stimulus can support and $\eta(CS_i, US)$ is a learning rate that can depend on the salience properties of both the conditional and the unconditional stimuli being associated.

p0090

From our perspective, the Rescorla-Wagner learning model was based on two the important (and innovative) assumptions or hypotheses: (1) learning happens only when events are not predicted, and (2) the predictions due to different conditional stimuli are summed together to form the total prediction in the trial. These assumptions allowed the model to explain parsimoniously several anomalous features of animal learning: why an already predicted unconditional stimulus will not support conditioning of an additional conditional stimulus (as in blocking; Kamin, 1969); why differently salient conditional stimuli presented together might become differentially associated with an unconditional stimulus (as in overshadowing; Reynolds, 1961); and why a stimulus that predicts the *absence* of an expected unconditional stimulus acquires a negative associative strength (as in inhibitory conditioning; Konorski, 1948; Rescorla and Lolordo, 1968).

p0100

The Rescorla-Wagner model explains quite elegantly a large collection of behavioral data (and, furthermore, predicted previously undetected phenomena such as over-expectation; Rescorla, 1970; Kremer, 1978); however, it suffers from two major shortcomings. First, by treating the conditional and unconditional stimuli as qualitatively different, it does not extend to the important phenomenon of second-order conditioning. Second-order conditioning is a behavioral analog to the idea of transitivity: if stimulus B predicts an affective outcome (say, reward) and stimulus A comes to predict stimulus B, then stimulus A also gains predictive value, i.e., *a predictor of a predictor is a predictor*. To be more concrete, suppose that a tone is repeatedly followed by food delivery so that an association forms between them (tone predicts food), then subsequent pairing of a light with the tone can confer predictive value to the light (for instance, the animal will approach the light and perhaps salivate when it comes on) – this effect is second-order conditioning. This laboratory paradigm is especially important given the prevalence of conditioning of humans to monetary outcomes, which are second-order predictors of a wide range of affectively desirable unconditional stimuli, such as food and shelter.

p0110

The second shortcoming of the Rescorla-Wagner rule is that its basic unit of learning is (the artificial contrivance of) a conditioning *trial* as a discrete temporal object. Not only does this impose an experimenter-oriented parsing of otherwise continuous

events, but it also fails to account for the sensitivity of conditioning to the precise temporal relations between the conditional and the unconditional stimuli within a trial (that is, whether they appeared simultaneously or serially, their order of appearance, and whether there was a time lag between them).

s0040 ## Temporal Difference Learning

p0120 To overcome these two problems, Sutton and Barto (1990) suggested the *temporal-difference learning rule* as a model of Pavlovian conditioning (i.e., prediction learning). Temporal-difference (TD) learning is quite simply a temporally extended version of the Rescorla-Wagner model discussed above. However, although the distinctions between the Rescorla-Wagner model and the TD model will seem subtle, the differences allow the TD model to account for higher-order conditioning and make it sensitive to the (potentially) important temporal relationships within a learning trial (Sutton and Barto, 1990).

p0130 In TD learning, the goal of the learning system (agent) is to estimate the future value of different states. For example, from a learning standpoint, the TD model assumes that the goal of a rat running about in a novel arena containing hidden rewards (e.g. food pellets hidden underneath bedding) is to learn the value of various positions in the arena. One way to evaluate the locations would be to estimate for each location the total amount of reward that the rat could expect to receive in the distant future. However, after a location is visited, there are many paths leading away from it, which yield variable amounts of reward. According to TD learning, a useful value function is the average amount of future reward expected when starting from each location.

p0140 To this end, in TD learning the time within a trial is explicitly represented (t below), and learning occurs at every timepoint within a trial, according to

$$V_{new}(S_i, t) = V_{old}(S_i, t) + \eta\left[r(t) + \gamma \sum_{S_{k@t+1}} V_{old}(S_k, t+1) - \sum_{S_{j@t}} V_{old}(S_j, t)\right] \quad (22.2)$$

p0150 In this specific variant of TD learning, stimuli create long-lasting memory traces (representations), and a separate value $V(S, t)$ is learned for every timepoint of this trace (for instance, a stimulus can predict a reward 5 seconds after its presentation, but not 10 seconds later). As in the Rescorla-Wagner rule, η is a learning rate, and learning is driven by discrepancies between available and expected outcomes; however, the crux of the difference between the rules is in how predictions, or expectations, are construed. In TD learning, the associative strength of the stimuli (and traces) at time t is taken to predict not only the immediately forthcoming reward $r(t)$, but also the future predictions due to those stimuli that will still be available in the next time-step $\Sigma_{Ss@t+1}\ V(S,\ t+1)$, with $\gamma \leq 1$ discounting these future delayed predictions.

p0160 It turns out that the TD learning rule can be derived as a normative prediction learning rule. The formal justification for TD learning as a method for optimal reinforcement learning comes from its relation to dynamic programming methods (Sutton, 1988; Watkins, 1989; Barto *et al.*, 1990). Dynamic programming is a collection of computational methods for solving stochastic sequential decision problems (Bellman, 1957). Departing for the moment from animal conditioning and human decision-making, consider a dynamic process (called a *Markov chain*) in which different states $S \in S$ follow one another according to some predefined probability distribution $P(S_{t+1} \mid S_t)$, and rewards are observed at each state with probability $P_r(S)$. A useful quantity to predict in such a situation is the expected sum of all future rewards, given the current state S_t, which we will call the *value* of state S

$$\begin{aligned} V(S_t) &= E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid S_t] \\ &= E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \,\middle|\, S_t\right] \end{aligned} \quad (22.3)$$

where $\gamma \leq 1$ is a factor discounting the effect of rewards distant in time on the value of the current state (this is necessary to ensure that the sum of future rewards is bounded). The expectation in equation (22.3) is with respect to two sources of stochasticity: (1) the probability of transitioning from one state to the next, and (2) the probability of reward in each state. Note that from this definition of state value it follows directly that

$$V(S_t) = P_r(S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1} \mid S_t) V(S_{t+1}) \quad (22.4)$$

p0170 This recursive relationship or *consistency* between state values lies at the heart of both dynamic programming and TD learning, and can be viewed as a kind of regularization or smoothing through time inasmuch as time indexes transitions from one state to another. The key to learning these values is that the recursive

relationship holds *only* if the values are correct (i.e., they correctly predict the expected discounted sum of future values). If the values are incorrect, there will be a discrepancy between the two sides of the equation

$$\delta(t) = P_r(S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)V(S_{t+1}) - V(S_t). \quad (22.5)$$

p0180 This *temporal-difference prediction error* is a natural "error signal" for improving estimates of the function $V(S_t)$ such that the discrepancy will be reduced

$$V(S_t)_{new} = V(S_t)_{old} + \eta \cdot \delta_t \quad (22.6)$$

p0190 Returning to prediction learning in real-world scenarios, we note that this dynamic programming updating scheme has one problem: it requires knowledge of the dynamics of the environment, that is, $P_r(S)$ and $P(S_{t+1}|S_t)$ (the "world model") must be known in order to compute the prediction error $\delta(t)$ in equation (22.5). This is clearly an unreasonable assumption when considering an animal in a Pavlovian conditioning task, or a human predicting the trends of a stockmarket. Werbos (1977), in his "heuristic dynamic programming methods," and, later, Barto *et al.* (1989) and Bertsekas and Tsitsiklis (1996), suggested that in such a "model-free" case, the environment itself can supply this information stochastically and incrementally. An animal can *sample* the reward probabilities in each state, and the probabilities of transitions from one state to another, as it experiences the task. Updating according to these samples will eventually lead to the correct predictive values. Thus the stochastic prediction error

$$\delta(t) = r_t + \gamma V(S_{t+1}) - V(S_t) \quad (22.7)$$

(where r_t is the reward observed at time t, when in state S_t, and S_{t+1} is the next observed state of the environment) can be used as a Monte Carlo approximation to dynamic programming, in order to learn optimal predictive state values. The resulting learning rule

$$V(S_t)_{new} = V(S_t)_{old} + \eta(r_t + \gamma V(S_{t+1}) - V(S_t)) \quad (22.8)$$

is exactly the same rule as that proposed by Sutton and Barto (1990) in equation (22.2), if we add the Rescorla-Wagner-like assumption that the predictions of the different stimuli comprising the state of the environment are additive (which is not the only way to combine predictions, and is certainly not always the sensible option; see Dayan *et al.*, 2000). This shows that, using TD learning, animals can learn the optimal (true) predictive values of different events in the environment, even when this environment is stochastic and its dynamics unknown.

Optimal Action Selection s0050

p0200 The discussion above holds whenever the probabilities of transitioning between different states or situations in the environment are stationary in time, as in Pavlovian conditioning (in which the animal cannot influence the events by means of its actions) or in situations in which the animal has a fixed behavioral policy (Sutton, 1988). But what about improving action selection in order to obtain more rewards – that is, what about instrumental conditioning? Since the environment rewards us for our actions, not for our predictions (be they correct as they will), the ultimate goal of prediction learning is to aid in action selection.

p0210 The problem of optimal action selection is especially difficult in those (very common) cases in which actions can affect long-term outcomes, or in which an outcome depends on a series of actions. For example, when winning or losing a game of chess, it is not at all simple to infer which were the actions responsible for this outcome, in order to improve the playing policy. This is true in the animal domain as well: when reaching a dead-end in a maze, how will a rat know which of its previous actions was the erroneous one? And, conversely, when it finds the cheese in the maze, how will it know which actions should be credited with the success? This is the (in)famous *credit assignment problem* (Sutton, 1978; Barto *et al.*, 1983; Sutton and Barto, 1998). RL methods solve the credit assignment problem by basing action selection not only on immediate outcomes but also on value predictions, such as those we discussed above, which embody long-term predictions of future outcomes.

p0220 First, note that, given predictive state values such as those learned by TD learning, a person could select the best long-term action at each state if only he knew what state that action would lead to (e.g., McClure *et al.*, 2003a). That is, given the transitions between states, the best action to choose is the one that leads to the state with the highest value. In fact, TD learning was first used in this way to select actions in Samuel's (1959) checker player. But what if this information is not available? For example, imagine deciding whether to buy or to sell a stock on the stock market – clearly, if you knew whether its price would increase or decrease as a result of your (and the rest of the market's) actions, this would be a trivial decision. But what can a human or a rat do in the completely model-free

case – i.e., when it is not explicitly known how different actions will influence the state of the environment?

s0060 ***Actor/critic Methods***

p0230 In one of the first RL papers (which was inspired by neural network models of learning), Barto *et al*. (1983) showed that the credit assignment problem can be solved by a learning system comprised of two neuron-like elements. One unit, the "adaptive critic element (ACE)," constructs an evaluation of different states of the environment, using a temporal-difference like learning rule from which the TD rule above was later developed. This evaluation is used to augment the external reinforcement signal and train a second unit, the "associative search element (ASE)," to select the correct action at each state through a trial-and-error process. These two elements were the precursors of the modern-day actor/critic framework for model-free action selection.

f0010

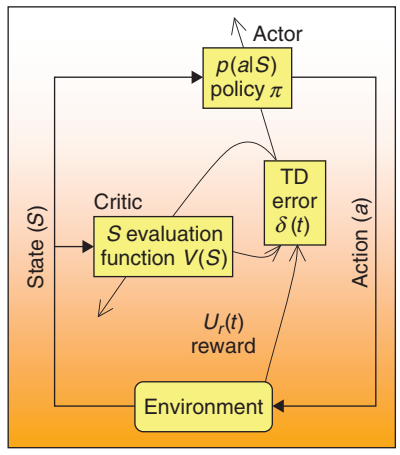


FIGURE 22.1 Actor/critic architecture. The environment provides a state S and a reinforcement signal $r(t)$ to the critic, who uses these to compute a temporal difference prediction error (equation 21.7). This prediction error is used to train the state value predictions $V(S)$ in the critic as well as the policy $\pi(S, a)$ in the actor. Note that the actor does not receive information regarding the actual outcomes of its actions. Rather, the prediction-error signal is a proxy to these, telling the actor whether the outcomes are better or worse than expected.

p0240 The insight in the ASE-ACE model, first due to Sutton (1978), is that even when the external reinforcement for a task is delayed (as in the game of chess), a temporal-difference prediction error can convey, at every time-step, a "reinforcement" signal to the action just chosen. In the absence of external reinforcement ($r_t = 0$), the prediction error $\delta(t)$ in equation (21.7) is equal to $\gamma V(S_{t+1}) - V(S_t)$; that is, it compares the values of two consecutive states. If the action has led to a state with a higher value than the previous state, this prediction error will be positive; if the situation has worsened due to the action taken, it will be negative. In the former case, the tendency to perform this action at this state should be increased (as the action has led to higher expected future rewards), while in the latter it should be decreased. Thus the agent can learn an explicit *policy* – a probability distribution over all available actions at each state $\pi(S, a) = p(a \mid S)$, and an adequate learning rule for this policy is

$$\pi(S, a)_{new} = \pi(S, a)_{old} + \eta_\pi \delta(t) \tag{22.9}$$

where $\eta\pi$ is the policy learning rate and $\delta(t)$ is the prediction error from equation (22.7).

p0250 Thus, in actor/critic models, a critic module estimates state values $V(S)$ based TD learning from experience with the environment, and the same temporal-difference prediction error that is used to train the critic's values is also conveyed to the actor module, which maintains and learns a policy π (Figure 22.1). This method is closely related to policy improvement methods in dynamic programming (Sutton, 1988), and Williams (1992) has shown that in some cases the actor/critic can be construed as a gradient-climbing algorithm for learning a parameterized policy, which converges to a local minimum (see also Dayan and Abbott, 2001). However, in the general case actor/critic methods are not guaranteed to converge (*cf*. Baird, 1995; Konda and Tsitsiklis, 2003). Nevertheless, some of the strongest links between reinforcement learning methods and neurobiological data regarding animal and human decision-making have been through the actor/critic framework. Specifically, actor/critic methods have been convincingly linked to action selection and prediction learning in the basal ganglia (e.g., Barto, 1995; Houk *et al*., 1995; Joel *et al*., 2002), as will be detailed in the next section. (More recent work (Morris *et al*., 2006; Roesch *et al*., 2007) suggests that the learned values may be more sophisticated and actually represent a value function defined over state–action pairs (rather than simply over states independent of actions), for which the learning algorithm is slightly different – see *Q*-learning or SARSA, below.)

s0070 ***State–action Values***

p0260 An alternative to actor/critic methods for model-free RL is to explicitly learn the predictive value (in terms of future expected rewards) of taking a specific action at a certain state. Building on dynamic programming methods of "policy iteration" (Howard, 1960), Watkins (1989) suggested *Q-learning*, a modified temporal-difference method in which the agent learns the value $Q(S, a)$ of each state-action pair (S, a) rather

than value $V(S)$ of each state S. The learning rule itself is quite similar to the state value learning rule above

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta\delta(t) \tag{22.10}$$

however, the temporal-difference prediction error term which drives Q-learning is slightly different

$$\delta(t) = r_t + \max_a \gamma Q(S_{t+1}, a) - Q(S_t, a_t) \tag{22.11}$$

where the $\max_a$ operator means that the temporal difference is computed with respect to what is believed to be the best available action at the subsequent state S_{t+1}. This method is considered "off-policy," as it takes into account the best future action, even if this will not be the actual action taken at S_{t+1}. In an alternative "on-policy" variant called SARSA (state–action–reward–state–action), the prediction error takes into account the actual chosen action, which leads to:

$$\delta(t) = r_t + \gamma Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t) \tag{22.12}$$

p0270 In both cases, given such Q-values, action selection is easy, as the best action at each state S is that which has the highest $Q(S, a)$ value. Furthermore, dynamic programming results regarding the soundness and convergence of "policy iteration" methods (in which a policy is iteratively improved through bootstrapping of the values derived given each policy; Howard, 1960; Bertsekas and Tsitsiklis, 1996) ensure that if the proper conditions on the learning rate are met, these methods will indeed converge to the true optimal (in case of Q-learning) or policy-dependent (in the case of SARSA) state–action values.

APPLICATION OF REINFORCEMENT LEARNING MODELS TO NEURAL DATA

p0280 In recent years, RL models like those highlighted above have been applied to a wide range of neurobiological and behavioral data. In particular, the computational functions of neuromodulators such as dopamine, acetylcholine, and serotonin have been addressed using a growing family of RL models. Among these neuromodulatory systems, the dopamine system has long attracted the most attention, perhaps due to its well-known connection with disease processes like drug addiction, Parkinson's disease, and schizophrenia, as well as its role in reward learning and working memory. It is in elucidating the role of dopamine signals in the brain that computational models of learning in general, and TD learning in particular, have had their most profound and sustained impact on neuroscience.

p0290 Continuing in the spirit of a historical narrative, let us turn back two decades to the 1980s and early 1990s of the previous century, when it became clear that antipsychotic medication (i.e., dopamine receptor blockers), while mitigating many of the dramatically troubling symptoms of schizophrenia (auditory hallucinations, paranoia, etc.), also caused hedonic blunting. That is, patients receiving this type of medication appeared to not derive pleasure from stimuli and behavioral acts formerly known to cause pleasure. Dopamine receptor blockers were also shown to have detrimental effects on reward learning in laboratory animals. In light of these observations, Roy Wise proposed the "anhedonia hypothesis" of dopamine function (Wise *et al.*, 1978a, 1978b). According to this proposal, the role of dopamine is to mediate the rewarding or primary motivational characteristics of natural stimuli such as food, water, and sex, as well as those of drugs of abuse (Wise, 1982; for a recent review see Wise, 2004). Specifically, this suggested that dopamine equals reward; that is, that there is an equivalence between the level of dopamine in the brain and "reward value." Wise's hypothesis initiated a surge of studies into the effects of neuroleptics on reward-mediated learning, and the results indicated that blocking dopamine is like removing the reward contingent on an animal's actions (i.e., it causes extinction of responding, as if the reward is absent; see, for example, Franklin and McCoy, 1979; Willner *et al.*, 1987). The puzzle of the role of dopamine in the brain seemed close to being solved.

p0300 At that time, the lab of Wolfram Schultz, who pioneered single-unit recordings from the midbrain of awake and behaving monkeys, began recording the activity of dopaminergic neurons while monkeys underwent simple instrumental or Pavlovian conditioning (Romo and Schultz, 1990; Ljungberg *et al.*, 1992; Schultz *et al.*, 1993). As expected, these cells showed phasic bursts of activity when the monkey was given a rewarding sip of juice or a morsel of apple. Surprisingly, however, if food delivery was consistently preceded by a tone or a light, after a number of trials the dopaminergic response to reward disappeared. Contrary to the anhedonia hypothesis, the lack of measurable dopaminergic response to reward delivery did not accompany extinction, but rather acquisition – the monkeys began showing conditioned responses of anticipatory licking and arm movements to the reward-predictive stimulus. Indeed, not only the monkeys' responses to the tone, but also their dopaminergic neurons began responding

to the tone, showing distinct phasic bursts of activity whenever the tone came on.

p0310 This pattern of results was also true for the difference between self-initiated reaching for reward (in which case dopamine neurons responded phasically to touching the reward) versus cue-initiated movements (in which case the neurons responded to the cue and not the reward). Rather than mediating the effects of affectively rewarding stimuli, it seemed that (quoting the conclusion sentences from papers of the time) "dopamine neurons exert a predominantly enabling effect on neurons more directly involved in the internal generation of movement" (Romo and Schultz, 1990: 592); "during acquisition of a simple behavioral task, dopamine neurons respond to unconditioned and conditioned salient stimuli that attract the attention of the animal, induce behavioral activation, and are associated with reward" (Ljungberg *et al*., 1992: 145); and "dopamine neurons respond phasically to alerting external stimuli with behavioral significance whose detection is crucial for learning and performing delayed response tasks" (Schultz *et al*., 1993: 900); all this "while not conveying specific information about the physical characteristics of stimuli nor the emergent behavioral reaction" (Romo and Schultz, 1990: 592).

p0320 A resolution of this conundrum was suggested in the mid 1990s, when Read Montague, Peter Dayan, Terry Sejnowski, and colleagues noticed that this pattern of dopaminergic responding throughout the course of learning conforms exactly to the characteristics of a reward prediction error (Montague *et al*., 1993, 1994, 1995, 1996). Indeed, the hallmark of temporal-difference prediction errors is that they occur only when events are not predicted. For instance, in a simulated Pavlovian conditioning scenario in which a tone CS is followed 2 seconds later by a food US, prediction errors arise as a result of the unexpected US early in training when the relationship between the CS and the US is not yet known (Figure 22.2a), but not later in training when the CS comes to predict the US (Figure 21.2b). Providing that the CSs occur randomly and thus can not be predicted, at late stages of training they themselves generate a prediction error (similar to the one that had previously accompanied the US delivery) which can support second-order conditioning. Moreover, in trials in which the US is not delivered, a negative prediction error occurs at the precise time of the expected US delivery (Figure 21.2c; such precise timing also necessitates a stimulus representation that can track elapsed time, as detailed in the figure caption).

p0330 The close correspondence between the findings of Schultz and colleagues regarding phasic dopaminergic firing patterns and these characteristics of a temporal-difference prediction error (Figure 22.2d–f) led Montague *et al*. (1996) to suggest the *reward-prediction error theory of dopamine* (see also Schultz *et al*., 1997). (Interestingly, dopaminergic neurons do not seem to be involved in the signaling or prediction of aversive stimuli (Mirenowicz and Schultz, 1996; Tobler *et al*., 2003; Ungless *et al*., 2004), in which the neuromodulator serotonin has been implicated instead (Daw *et al*., 2002).) Within this theoretical framework, it was immediately clear why dopaminergic neurons fire to unexpected rewards but not to those that are predicted by previous stimuli, and why dopamine is necessary for reward-mediated learning in the basal ganglia. Indeed, the shift in dopaminergic activity from the time of reward to the time of the predictor (Takikawa *et al*., 2004) resembles the shift of behavioral responses from the time of the US to that of the CS in Pavlovian conditioning experiments (Schultz *et al*., 1997; Hollerman and Schultz, 1998). Moreover, the model explained why, after training, dopamine neurons did not fire above baseline in the time period between a predictive cue and the reward delivery – in the absence of new information, there are no prediction errors at these intermediate times. From the point of view of downstream neural structures, continuous baseline firing of dopaminergic neurons can be interpreted as "things are just as expected."

p0340 The basic characteristics of phasic dopaminergic responding have since been replicated in many variants (Hollerman and Schultz, 1998; Schultz, 1998; Tobler *et al*., 2003; Takikawa *et al*., 2004; Bayer and Glimcher, 2005). In fact, recent work aimed at putting the prediction-error hypothesis to quantitative test has demonstrated that the correspondence between phasic dopaminergic firing and temporal-difference prediction errors goes far beyond the three basic characteristics depicted in Figure 21.2. For instance, using a general regression model that does not assume temporal-difference learning *a priori*, Bayer and colleagues (Bayer and Glimcher, 2005; Bayer *et al*., 2007) have rigorously shown that the contribution of previously experienced rewards to the dopaminergic response to the reward in the current trial is exactly according to an exponentially weighted average of past experience (as is implicit in the temporal-difference learning rule; see Figure 22.3). Moreover, conditioned stimuli predicting probabilistic rewards or rewards of different magnitudes have been shown to elicit a phasic dopaminergic response that is indeed proportional to the magnitude and/or probability of the expected reward (Fiorillo *et al*., 2003; Morris *et al*., 2004; Tobler *et al*., 2005; Figure 22.4a, b), and firing patterns in tasks involving probabilistic rewards are in accord with a constantly back-propagating error signal (Niv *et al*., 2005b; Figure 22.4b, c). Even in sophisticated

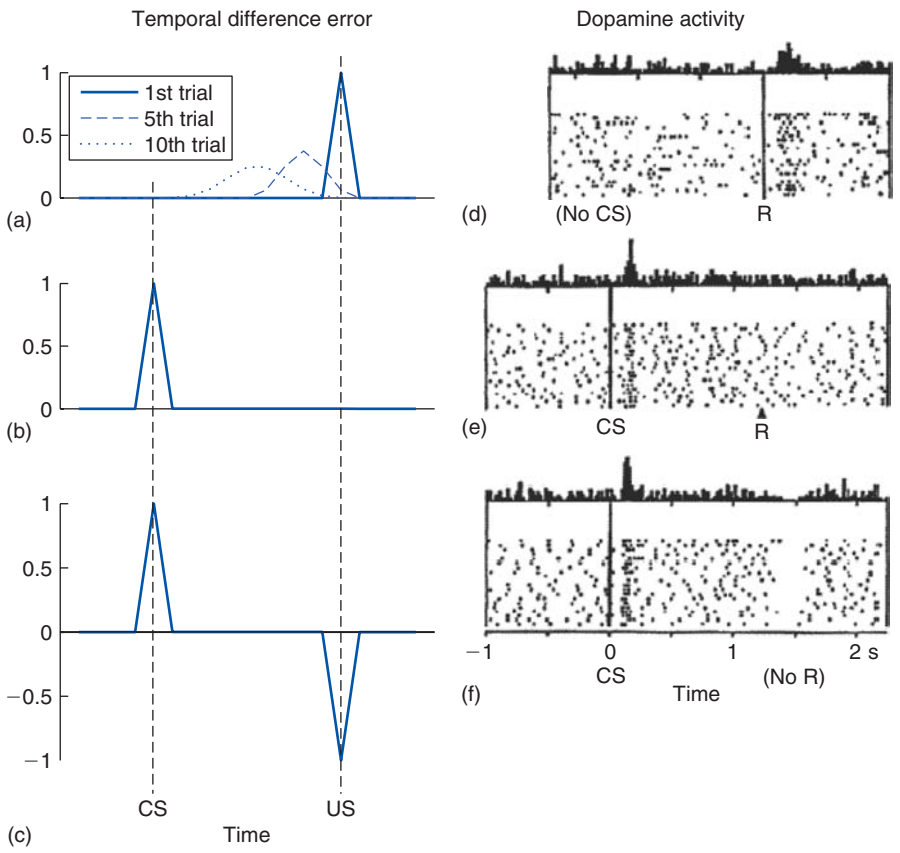
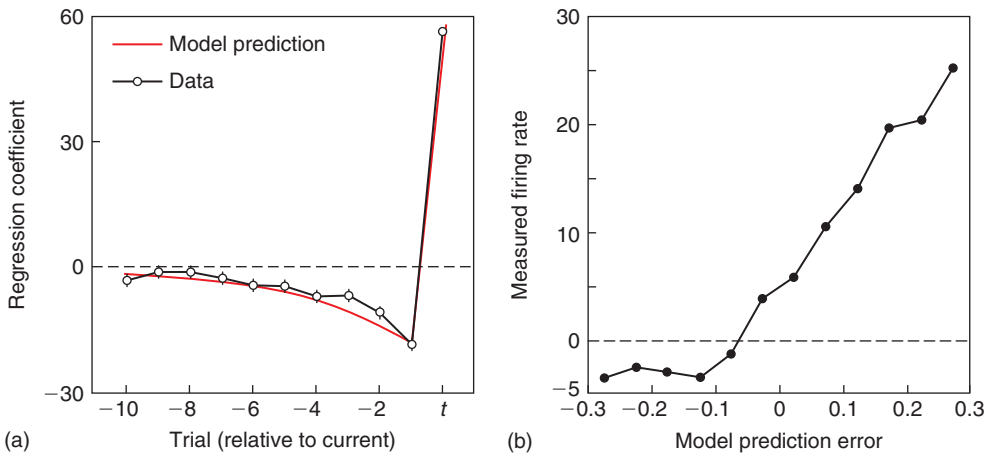
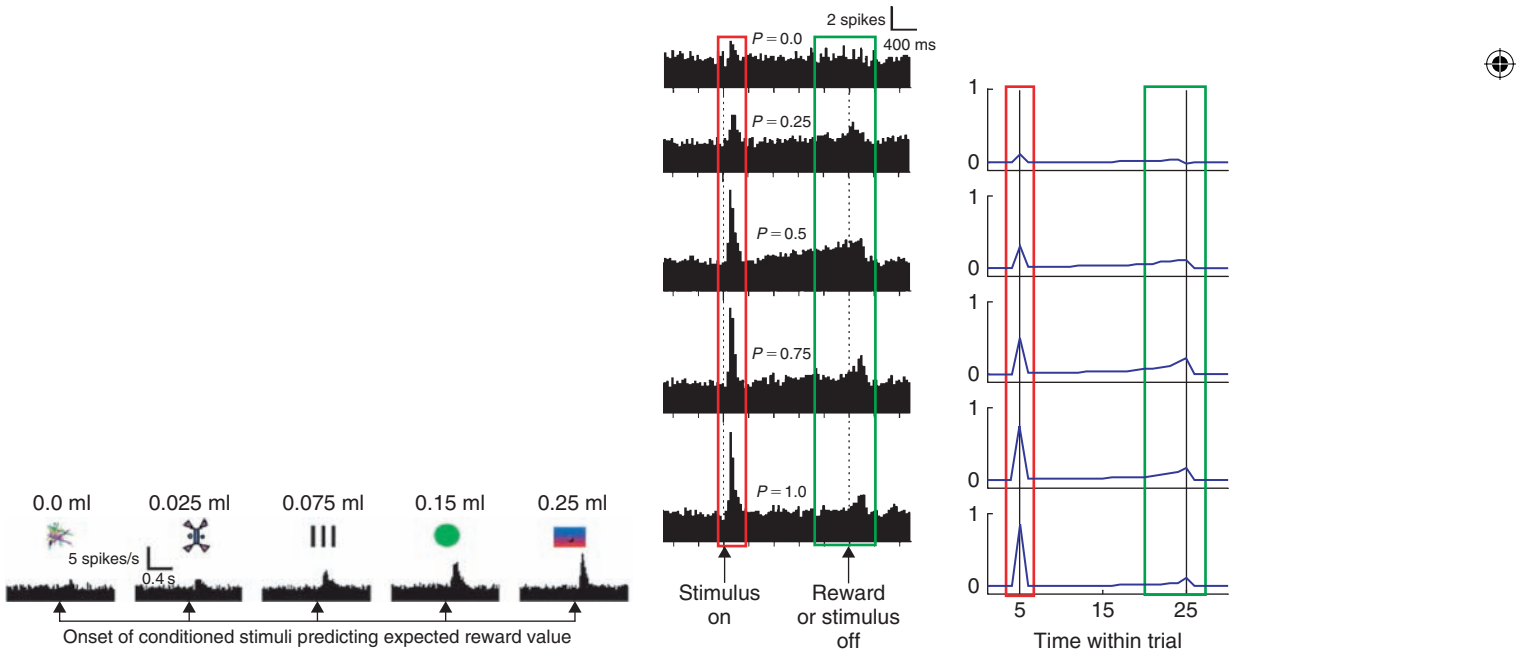


FIGURE 22.2 (a–c) Temporal-difference prediction errors in a simple Pavlovian conditioning task. A tone CS is presented at random times, followed 2 seconds later with a food US. At the beginning of training (a), the affectively significant US is not predicted, resulting in prediction errors. As learning progresses (trials 5 and 10 are plotted as examples), the prediction error propagates back (Niv *et al*., 2005a) as values of preceding timesteps are updated (equation 21.8). When the predictive relationships are completely learned (b), the now-predicted US no longer generates a prediction error, rather, the unpredicted occurrence of the CS is accompanied by a prediction error. If the US is unexpectedly omitted (c), a negative prediction error is seen at the time in which the US was expected, signaling that expectations were higher than reality. In these simulations, the CS was represented over time with the commonly used serial compound state representation (Kehoe, 1977, Sutton and Barto, 1990), and there was no discounting ($\gamma = 1$). (d–f) Firing patterns of dopaminergic neurons in the ventral tegmental areas of monkeys performing an analogous instrumental conditioning task. Each raster plot shows action potentials (dots) with each row representing a trial, aligned to the time of the cue (or the reward). Bar histograms show the summed activity over the trials plotted below. When a reward is given unexpectedly, dopaminergic neurons respond with a phasic burst of firing (d). However, after conditioning with a predictive visual cue (which, in this task, predicted a food reward if the animal quickly performed the correct reaching response), the predicted reward no longer elicits a burst of activity, and the phasic burst now accompanies the presentation of the predictive cue (e). In "catch" trials, in which the food reward was unexpectedly omitted, dopaminergic neurons showed a precisely-timed pause in firing, below their standard background firing rate (f). Subplots 22.2 (d–f) adapted from Schultz *et al*., (1997). Note that the discrepancies between the simulation and the dopamine neuron firing patterns in terms of the magnitude and spread of the prediction errors at the time of the reward likely result from the temporal noise in reward delivery in the instrumental task, and the asymmetric representation of negative and positive prediction errors around the baseline firing rate of these neursons (Niv *et al*., 2005a). f0020

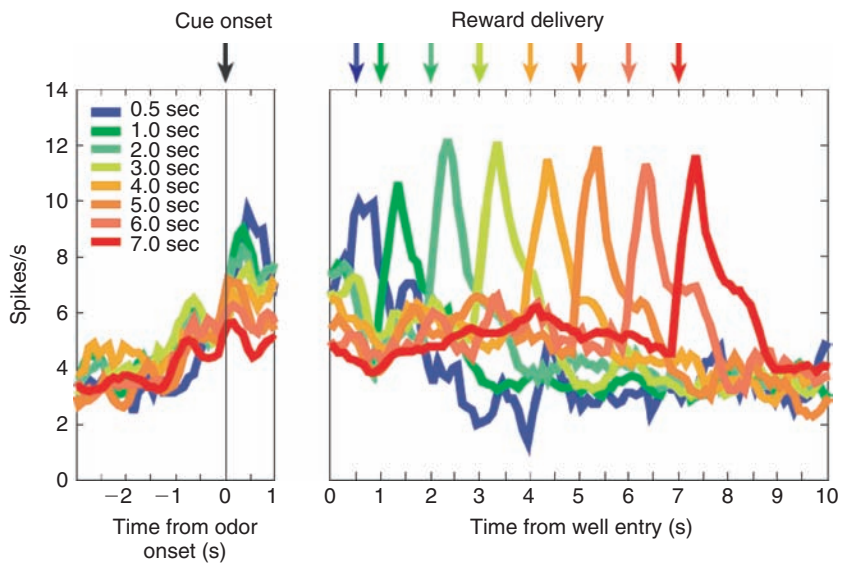
conditioning tasks such as blocking and appetitive conditioned inhibition, Waelti *et al*. (2001) and Tobler *et al*. (2003, 2005) have shown that the dopaminergic response is in line with the predictions of temporal-difference learning. And in the timing domain, recent results show that the dopaminergic activity to a cue predicting a delayed reward is attenuated in proportion to the delay (Figure 22.5), as is expected from a signal predicting the expected sum of *discounted* future rewards (Roesch *et al*., 2007). Finally, direct measurements of extracellular dopamine in the nucleus accumbens (a brain area also referred to as ventral striatum; a major target for dopaminergic projections from the ventral tegmental area) using fast-scan cyclic voltammetry (which has subsecond temporal resolution) have confirmed that the phasic changes in levels of dopamine in target structures indeed conform quantitatively to a prediction error signal (Paul Phillips, personal communication; see also Day *et al*., 2007; Chapter 24 of this volume), despite the non-linearities between dopamine neuron firing and actual synaptic discharge of the transmitter (Montague *et al*., 2004).



f0030 **FIGURE 22.3** Dopaminergic responses depend on past rewards, as is predicted by temporal-difference learning. Here, single-unit recordings of dopaminergic neurons were conducted while monkeys performed a rewarded saccade task (Bayer and Glimcher, 2005; Bayer *et al.*, 2007). In this task, a visual cue signaled the start of a trial, after which the monkey could choose when to perform a saccade toward a target. The magnitude of the reward (volume of fruit juice) depended on the time of the saccade, such that a longer wait was rewarded with more juice, up to an unsignaled deadline, after which saccades earned no juice. This allowed for a range of prediction errors to be sampled. The monkey's task was to learn to optimal time-to-look in order to receive the maximum reward. (a) Prediction errors at the time of the reward were regressed against the rewards in the past 10 trials, which should, according to TD learning, determine the amount of reward predicted (and thus the magnitude of the prediction error). In solid red is the dependency of the prediction error on past rewards as derived from the theoretical model (with a learning rate of $\eta = 0.7$; see Bayer and Glimcher, 2005 for details), in black circles are the regression coefficients obtained from the data. The data are in close fit with the model prediction. (b) the measured prediction error is linear with the model-derived prediction error, at least in the domain of positive prediction errors (see Bayer *et al.*, 2007, for data indicating that negative prediction errors are encoded by the length of the pause in firing, rather than the magnitude of the pause below the baseline firing rate). Figure adapted from Bayer and Glimcher, 2005.



f0040 **FIGURE 22.4** Dopaminergic responses comply with the predictions of temporal-difference learning. (a) Phasic prediction errors at the time of a cue predicting reward are proportional to the magnitude of the predicted reward (adapted from Tobler *et al.*, 2005). (b, c) When different cues predict the same reward but with different probabilities, the prediction error at the time of the cue is proportional to the predicted probability of reward (red rectangles; compare panel (b) (data) to panel (c) (TD simulation)). However, due to the low baseline firing rate of midbrain dopaminergic neurons, negative prediction errors cannot be encoded with as deep a "dip" in firing rate as is the height of the "peak" by which positive prediction errors are encoded. As a result, when rewards are probabilistic, averaging over rewarded and unrewarded trials will create an apparent ramp leading up to the time of the reward (green rectangles; compare panel (b) (data) to panel (c) (TD simulation)). Panel (b) adapted from Fiorillo *et al.*, 2003; Panel (c) adapted from Niv *et al.*, 2005a.

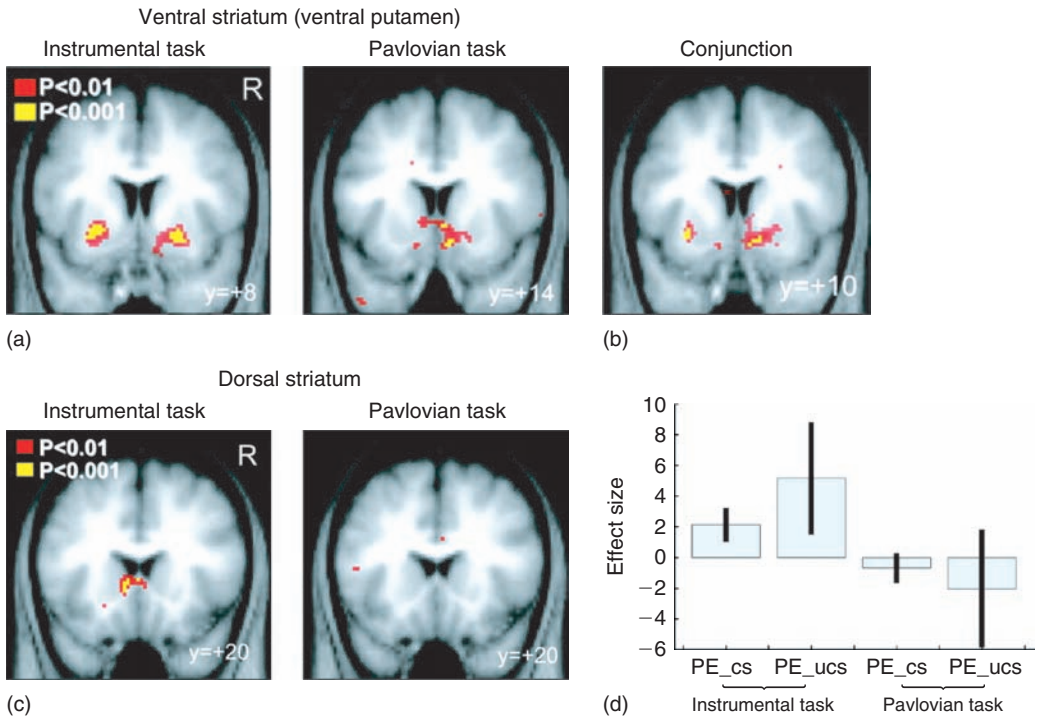


f0050 **FIGURE 22.5** Average firing rate of 19 dopaminergic neurons, recorded in rats performing an odor-discrimination task in which one of the odors predicted a delayed reward that will be delivered in a food-well. Color indicates the length of the delay preceding reward delivery from 0.5 to 7 seconds. Activity is aligned on odor onset (left) and food-well entry (right). Rewards were given in the well providing that the rat entered the correct well (as indicated by the odor cue), and remained there until the time of the reward. Adapted from Roesch et al. (2007).

p0350 Note that the prediction-error theory of dopamine is a *computationally precise* theory of the *generation* of phasic dopaminergic firing patterns. It suggests that dopaminergic neurons combine their diverse afferents (which include inputs from the medial prefrontal cortex, the nucleus accumbens shell, the ventral pallidum, the central nucleus of the amygdala, the lateral hypothalamus, the habenula, the cholinergic pedunculopontine nucleus, the serotoninergic raphe, and the noradrenergic locus coeruleus; Christoph *et al*., 1986; Floresco *et al*., 2003; Geisler and Zahm, 2005; Matsumoto and Hikosaka, 2007; Kobayashi and Okada, 2007) to compute a temporal-difference reward-prediction error. Moreover, it suggests that dopamine provides target areas with a neural signal that is theoretically appropriate for controlling learning of both predictions and reward-optimizing actions. Following the analogy between the dopamine signal and the temporal-difference prediction-error signal in actor/critic models (Joel *et al*., 2002), it has been suggested that the signal reported by dopaminergic neurons in the ventral tegmental area to ventral striatal and frontal target areas, is used to train predictions (as in the critic; Barto, 1995; Waelti *et al*., 2001), while a similar signal reported by dopaminergic neurons in the substantia nigra pars compacta to dorsal striatal target areas is used to learn an action-selection policy (as in the actor; Miller and Wickens, 1991; Wickens and Kötter, 1995; Houk *et al*., 1995; Joel and Weiner, 1999).

p0360 Recently, the role of phasic dopamine in action selection was assessed by combining non-trivial decision-making tasks in rodents or monkeys with single-cell recordings of dopaminergic activity. This is especially interesting, as temporal-difference methods do not fully prescribe the form of the reward-prediction error in tasks involving action selection. As mentioned above, different computational algorithms, namely actor/critic, *Q*-learning and SARSA, make different predictions regarding the nature of the cue-related prediction error (Niv *et al*., 2006b), making electrophysiological evidence critical in constraining the algorithm actually used by the brain. In a recent study, Morris *et al*. (2006) trained monkeys with cues predicting reward with different probabilities. Interestingly, in "catch" trials, in which the monkeys were given a choice between two cues, the cue-elicited prediction errors matched best the errors corresponding to the cue that would subsequently be chosen. This is contrary to the straightforward predictions of an actor/critic mechanism, and more in line with SARSA learning. However, data from rats performing a more dynamic odor-discrimination task (Roesch *et al*., 2007) and from monkeys engaged in a difficult random-dot motion discrimination task (Nomoto *et al*., 2007) suggest that predictions (and thus prediction errors) can be sensitive to the information available at every time-point, representing stimuli before a choice is made, and representing the chosen cue only later. These different results can be incorporated into one learning scheme using appropriate representation of the states in the task, an issue that we shall return to in the last section of this chapter.

p0370 To end this section, we should mention that there are alternative psychological theories regarding the role of dopamine in conditioned behavior (for a recent debate-style review, see Berridge, 2007). These include Redgrave and colleagues' "incentive salience" (see, for example, Redgrave *et al*., 1999; Horvitz, 2000; Redgrave and Gurney, 2006), Berridge and



f0060 **FIGURE 22.6** BOLD correlates of reward-prediction error in a Pavlovian and an instrumental task can be anatomically dissociated. (a) Correlates of a reward-prediction error signal are seen in the ventral striatum (specifically, in the ventral putamen) in both an instrumental task involving action selection in order to maximize rewards, and a yoked Pavlovian task in which the subject observes choices made by the computer. (b) The conjunction of the activations verifies a similar involvement of the ventral striatum in both task variants. (c) However, correlates of reward-prediction errors are seen in the dorsal striatum only in the instrumental task. (d) Coefficient estimates for the prediction error at the time of the stimulus (PE_cs) and the reward (PE_ucs) for each task, from the peak voxel for the contrast: instrumental prediction error > Pavlovian prediction error. Figure adapted from O'Doherty *et al.*, 2004.

Robinson's "wanting" versus "liking" (e.g., Berridge and Robinson, 1998; Berridge, 2007), and ideas about dopamine signaling uncertainty (Fiorillo *et al.*, 2003). A discussion of the merits and pitfalls of the different theories is beyond the scope of this chapter, and in many cases would involve the less-than-satisfactory comparison of qualitative suggestions to quantitative predictions of an RL model. Nevertheless, in as far as these theories are indeed fundamentally different from the prediction-error theory (which is not always clear), it is our opinion that, to date, no alternative has mustered as convincing and multidirectional experimental support as the prediction-error theory of dopamine.

s0090 EVIDENCE FROM IMAGING OF HUMAN DECISION-MAKING

p0380 Although animal conditioning can display complex phenomena that are still well beyond our current understanding of the brain, ultimately we are interested in understanding human decision-making, the computations that underlie it, and the relationship of these computations to neural mechanisms. While the characteristics of human conditioning are similar to those of animal conditioning, the possibility of instructing subjects verbally allows for much more elaborate paradigms in human experiments. Of course, there are severe limitations on our ability to measure neural processes in humans. One technique that has recently been brought to the forefront is functional magnetic resonance imaging (fMRI), in which metabolic correlates of neural activity can be measured non-invasively, albeit at low temporal and spatial resolution (seconds and millimeters, respectively), and subject to noisy measurements (typically necessitating averaging over trials or subjects).

p0390 One great advantage of fMRI is that it allows imaging of activity throughout the entire brain, rather than in only a small population of neurons. Using fMRI as the neural probe of choice places a premium on using precise computational models of decision-making. A model-driven analysis also affords a special insight into brain function: the models identify *hidden variables* that control behavior, such as state values or prediction errors, for which we can search in the brain. This is different from the more straightforward

identification of the neural correlate of perception or motor behavior, and even goes beyond a search for the correlates of abstract entities (such as "reward" or "preference;" e.g., O'Doherty *et al*., 2002) because the computational model can quantitatively specify the dynamics of a hidden variable within a non-stationary learning and decision-making task. Identifying a neural correlate of such a signal advances our understanding of the brain in a way that would not be possible without the model. Of course, this can also lend powerful support for the model that gave rise to the specific values of the hidden variable: models that suggest different relationships between external events (stimuli and rewards) and internal representations (values and preferences) can be compared by testing how well activity in different brain areas correlates to the specific predictions of each model. With these general issues in mind, we turn below to the specific use of RL models in identifying learning signals (e.g. reward-prediction errors) and value-dependent responses in the human brain.

p0400 The first fMRI studies to search for prediction errors in humans implicated the nucleus accumbens and the orbitofrontal cortex (Berns *et al*., 2001; Knutson, *et al*., 2001a; Pagnoni *et al*., 2002), both major dopaminergic targets. O'Doherty *et al*. (2003) and McClure *et al*. (2003b) then used a hidden-variable analysis technique to identify the neural correlates of model-derived temporal-difference prediction errors. These studies again implicated the nucleus accumbens (the ventral striatum) and the putamen (part of the dorsal striatum). Later, functional imaging was used to distinguish between potential sites of Pavlovian versus instrumental learning: O'Doherty *et al*. (2004) showed that fMRI correlates of prediction-error signals can be dissociated in the dorsal and ventral striatum according to whether an action is required in order to obtain reward. For passive prediction-learning tasks the reward-prediction error was evident only in the ventral striatum, while in active tasks it was evident in both the ventral and the dorsal striatum (Figure 22.6). These findings and the model-based analysis that uncovered them suggest that stimulus–response learning typical of actor/critic circuits in humans may be associated with activation in the dorsal striatum.

p0410 Indeed, correlates of prediction errors in the dorsal and ventral striatum have now been seen in multiple studies (see, for example, Li *et al*., 2006; Preuschoff *et al*., 2006; Schönberg *et al*., 2007). Note, however, that the fMRI results cannot isolate dopaminergic activity from other activity in the brain. Furthermore, the measured blood oxygen level dependent (BOLD) signal in a brain area has been suggested to be correlated with local field potentials implying a correlation with the afferent inputs to and the local processing in a brain region (Logothetis, 2003), rather than the spiking activity of neurons within the region. (Local field potentials are electrophysiological signals that are related to the sum of all dendritic synaptic activity within a volume of tissue, thus they are dominated by the arrival of action potential along axons that terminate in the area, rather than the firing of neurons in that area. The local field potential is believed to represent the synchronized input into the observed area, as opposed to the spike data, which represent the output from the area.) Thus, prediction-error correlates in areas of the striatum and the prefrontal cortex are fully in line with the fact that these are the major targets for dopaminergic influence. Furthermore, dopaminergic manipulations (e.g., administration of dopamine enhancers (agonists) or dopamine receptor blockers (antagonists)) in such tasks have been shown to influence both the BOLD measurement of prediction-error activity and learning and action selection (Pessiglione *et al*., 2006), and recent results show that better learners show a higher correlation of striatal BOLD with a reward-prediction error (Schönberg *et al*., 2007). Of course, these areas are also targeted by other afferents, most notably the neuromodulator *serotonin*, which has been suggested as the counterpart to dopamine in the domain of punishment (Daw *et al*., 2002), and might explain why BOLD correlates of *positive* prediction errors for pain and punishment have also been found in the striatum (Seymour *et al*., 2004; Jensen *et al*., 2007; Menon *et al*., 2007). The relative contribution of many neural signals to the measured BOLD responses in these regions awaits more precise pharmacological manipulations and perhaps a serious technological advance.

p0420 Note also that, without temporal specificity and an analysis that specifically aims to tease apart different components of the reinforcement learning model, it is not easy to distinguish between prediction errors and state values at the time of a stimulus. This is because the prediction error at the time of an unpredicted stimulus is $\delta(t) = V(stimulus) - V(baseline)$, which, if we take V (*baseline*) to be 0, is exactly equal to V(*stimulus*). Indeed, many studies have implicated the striatum in representing the anticipated value of outcomes (e.g., Knutson *et al*., 2001a; Delgado *et al*., 2003; Knutson *et al*., 2003), and it is not always clear whether the measured activation is distinct from that attributable to a prediction error. In any case, electrophysiological data show that the striatum is definitely a viable candidate for representing state values (e.g., Schultz *et al*., 1992; Samejima *et al*., 2005). Studies in which outcomes involved both gains and losses have further implicated the striatum in the anticipation of losses, not only

gains, with a decrease in activity correlated with the anticipated loss. Moreover, the degree of deactivation to losses compared to activation to gains ("neural loss aversion") in the nucleus accumbens and the prefrontal cortex was predictive of individual differences in behavioral loss aversion (Tom *et al*., 2007). Finally, outcome values themselves (as well as subjective preferences) have been associated with activations in areas such as the ventromedial prefrontal cortex and the orbitofrontal cortex (e.g., Knutson *et al*., 2001b; O'Doherty *et al*., 2002; Knutson *et al*., 2003; McClure *et al*., 2004).

p0430 The promise of model-driven analysis of imaging data has yet to be fully realized, and the link between computational models of learning and the brain does not end with the identification of the reward-prediction error signal. Recent work has used such a hidden-variable analysis coupled with a reinforcement learning model to investigate the neural substrates of exploration (Daw *et al*., 2006). In "market-like" tasks, model-based approaches have identified learning signals related to so-called fictive outcomes (what might have happened but didn't, also called countefactuals; Lohrenz *et al*., 2007) and a hierarchical RL model has been used to demonstrate that the brain tracks the volatility (or rate of change) of the environment (Behrens *et al*., 2007). Furthermore, contrasting model-free temporal-difference learning with model-based learning algorithms that exploit the higher order structure of the learning task, Hampton *et al*. (2006) have begun to reveal the neural mechanisms of more model-based forms of human learning. One approach that is becoming increasingly common is the use of functional imaging in combination with pharmacological challenges (e.g., Pessiglione *et al*., 2006) or with radioligand-labeled positron emission tomography (e.g., Zald *et al*., 2004) to test more directly the *causal predictions* and *pharmacological hypotheses* of reinforcement learning models in human subjects (Knutson and Gibbs, 2007), respectively.

s0100

BEYOND PREDICTION ERRORS AND PHASIC DOPAMINE

p0440 The theoretical importance of prediction errors and prediction learning is undeniable. However, other computationally important signals have been posited and associated with neural function. In this section, we briefly discuss other components of learning beyond those of prediction errors, and how these might be related to different aspects of neuromodulation in the brain. We begin by applying the reinforcement learning framework to decisions about how fast (or with what vigor) to behave. This highlights the role played by the passage of time (specifically, the opportunity cost of time in terms of devoting time to one action rather than another) in ongoing decision-making. We focus on a recent model that shows that the net rate of rewards quantifies the opportunity cost of time, and discuss the proposal that this quantity is represented by tonic levels of dopamine in the striatum (Niv *et al*., 2007; Niv, 2007a). Dopamine has been argued to convey (at least) two separate signals to efferent (downstream) structures (e.g., Grace, 1991; Floresco *et al*., 2003): On the one hand, firing of dopaminergic neurons induces large but short-lived changes in the amount of dopamine in a synapse (a "phasic" signal, predominantly affecting D1-type low-affinity dopamine receptors). On the other hand, extrasynaptic levels of dopamine change on a much slower timescale (a "tonic" signal which affects high affinity D2-type dopamine receptors). The computational and theoretical differences between phasic and tonic aspects of prediction highlight the importance of carefully treating different timescales of dopamine signaling, and suggest that dopamine can simultaneously fulfill several roles in decision-making, without contradiction.

p0450 In the second part of this section, we discuss normative adaptations of the rate of learning to the decision-making task and the statistics of the environment. Bayesian inference models such as the Kalman filter show how different forms of uncertainty about the environment should affect the rate of learning, and the degree of reliance on previous experience. These effects have recently been associated with both acetylcholine and norepinephrine and their effects on learning and inference (Yu and Dayan, 2002, 2005), considerably enhancing our understanding of the neural basis of learning.

Tonic Dopamine and the Choice of Response Vigor

s0110

p0460 It is somewhat of a curiosity that although the tradition in animal experimentation is to investigate the determinants of *rates* of responding (as in Skinner's investigations of key-pecking in pigeons or lever-pressing in rats, so called "free-operant" experiments because the animal is free to choose when to respond and no trial structure is imposed on behavior), reinforcement learning models of conditioning have concentrated exclusively on the choice of discrete actions at pre-specified timepoints (as in a discrete trial two-alternative choice experiment). However, real-life

decisions most often take place in continuous time. In fact, every choice of action, even that in a discrete trial setting, is accompanied by a choice of the *speed* or *vigor* with which that action will be performed. This decision gives rise to response rates in free operant behavior, to running times in mazes, and to reaction-time data in discrete settings. It also interacts with the influences of motivation on behavior – a hungry rat running down a maze in search of food will run faster than a sated rat.

p0470 That reinforcement learning has, until recently, wholly ignored this aspect of decision making may be due to the historical origins of reinforcement learning theory in computer science. In simulations, or in robot decision-making, decisions can only occur in synchrony with a discrete CPU clock. However, theory does exist that deals with continuous time: this is *average reward* reinforcement learning in a *semi*-Markov decision process (Schwartz, 1993; Doya, 2000; Daw and Touretzky, 2002). Building on this theoretical framework, Niv *et al.* (2005a) recently proposed a reinforcement learning model of optimal rates of responding. In this model of instrumental conditioning, every choice of action is accompanied by a choice of a *latency* with which to perform that action, such that the net overall rate of rewards is maximized. The model successfully replicates and explains the fundamental characteristics of free operant response rates (Niv, 2007b), and explains how motivational states should affect decision-making (Niv *et al.*, 2006a).

p0480 Importantly, the average reward framework highlights an important factor that determines optimal responding: the net rate of rewards, that acts as the opportunity cost of time. To illustrate this, imagine a rat pressing a lever in order to obtain food. Suppose that through previous behavior food had been accrued at a rate of four pellets per minute. When contemplating devoting 5 seconds to executing the next lever-press, the potential benefit of this action (i.e., the probability of its generating reward, and the magnitude of this reward) should thus be weighed against both the (motor and other) costs of performing the action at this speed, and the opportunity cost of time, i.e., the potential loss of (on average) one in three reward pellets due to devoting time to this action rather than continuing to behave according to the previous policy. Because of this cost/benefit tradeoff, the model predicts that when the net rate of rewards is higher all actions should optimally be performed faster, as a result of the elevated opportunity cost (Niv *et al.*, 2007).

p0490 How does this relate to decision-making in the brain? Note that the prediction-error theory of dopamine concentrates on only one aspect of dopaminergic activity and influence: the effect of *phasic* dopaminergic signaling on learning and plasticity. However, dopamine neurons operate in both a phasic and a tonic mode (Grace, 1991; Weiner and Joel, 2002; Bergstrom and Garris, 2003; Floresco *et al.*, 2003; Goto and Grace, 2005), and affect not only synaptic plasticity, but also membrane potentials and neural excitability, which may be particularly sensitive to tonic levels of dopamine (Nicola *et al.*, 2000; Schultz, 2002). Furthermore, the effects of dopaminergic manipulations such as lesions, antagonism, or agonism, are first and foremost seen in the vigor of ongoing behavior, rather than in learning processes. For instance, a multitude of studies has shown that injections of 6-hydroxydopamine into the nucleus accumbens, which causes the death of dopaminergic neurons projecting to that area, profoundly reduce the rate of instrumental responding (for a review, see Salamone and Correa, 2002). As a result, dopamine in the striatum has been linked to invigorating Pavlovian and instrumental responding (Ikemoto and Panksepp, 1999; Salamone and Correa, 2002).

p0500 Combining these lines of evidence, Niv and colleagues have suggested that tonic levels of striatal dopamine represent the net rate of rewards. (In fact, if the tonic level of dopamine reflects spillover from phasic prediction error signals averaged over a longer time-frame due to slow reuptake, it follows computationally that it would, by default, equal the net rate of obtained rewards.) This hypothesis, dovetailing neatly with both computational theories regarding phasic dopamine signals and appetitive prediction errors, and psychological theories about dopamine's role in energizing responses, provides the first normative explanation for the critical role that tonic levels of dopamine play in determining the vigor of responding. It also suggests a route by which dopamine could mediate the effects of motivation on response vigor.

Acetylcholine and Norepinephrine and the Optimal Rate of Learning

s0120

p0510 One issue that we have not yet discussed relates to the assumption, in both the Rescorla-Wagner model and the temporal-difference model, that the predictions of multiple stimuli are simply added up to form the total prediction of reward. At this point, we can treat this as but a simplification: even the Rescorla-Wagner model allowed for different learning rates for different stimuli (for instance, based on their salience, as in overshadowing), implying that the prediction error should not affect all stimuli equally. A natural extension of this idea is to allow for stimuli to contribute

differentially to the overall prediction itself. Here again, control theory can inform us regarding the optimal combination of predictions and allocation of learning to different predictors. Simply put, unreliable stimuli, those with which we have had less experience and thus know less about, should contribute less to the prediction. As in the more general statistical problem of combining multiple sources of evidence, this implies a competitive interaction between predictors, with the most reliable predictor weighted most heavily in an overall weighted average. Conversely, when a prediction error occurs, more learning should be devoted to the stimuli about which there is most uncertainty that is, they should take responsibility for most of the prediction error (Dayan and Kakade, 2000; Dayan *et al*., 2000).

p0520 These ideas are formally couched in statistically optimal Bayesian inference in the Kalman filter model. The Kalman filter assumes an underlying generative model in which each stimulus gives rise to observations (rewards) distributed according to a Gaussian distribution. Furthermore, it assumes that the process by which the rewards are observed (or measured) is prone to Gaussian noise. Different from Rescorla-Wagner or temporal difference learning, the learning process in this case must infer the mean reward associated with each stimulus while taking into account these two sources of variability. Optimal Bayesian inference dictates tracking not only the mean predicted reward, but also the uncertainty in this mean as observations accumulate (thus the learning process includes two learning or update rules). However, the optimal learning rule for the mean is rather similar to the Rescorla-Wagner learning rule with an additional adjustment of the learning rate based on the different sources of variance and on the tracked uncertainty (Dayan *et al*., 2000).

p0530 Yu and Dayan (2005) have further analyzed the effects of uncertainty on learning and inference in a noisy and constantly changing environment. Their model accounts for two types of uncertainty: *expected* uncertainty that arises from known variability in predictors in the environment, and *unexpected* uncertainty due to sudden unforeseen changes in the environment. Building on physiological, pharmacological, and behavioral data, they hypothesize that the first source of uncertainty is signaled by acetylcholine and the second by norepinephrine. Inference in their model is considerably more complex than in the case of the Kalman filter, and approximations are needed to render it feasible for a neural mechanism. In terms of neuromodulation, while acetylcholine and norepinephrine are synergistic in the learning process, their relationship in inference is antagonistic – inconsistencies between prediction and observation must be attributed either to expected or to unexpected uncertainty. If the former is low, the inference will be of an unexpected change in the environment; conversely, when the environment is deemed very stable, prediction errors imply higher (expected) variability of the predictors.

WHAT'S MISSING? CHALLENGES AND FUTURE DIRECTIONS

s0130

p0540 RL models are now used routinely to design and interpret a wide range of reward learning and decision-making experiments; however, we view this success only as an important starting point. One of the reasons that RL models have been successful is that they have been made extremely simple and have been tested in very simple experimental settings. In this last section, we point out some of the experimental data that challenge RL models and separate what we consider to be real challenges from confusion arising from a lack of clarity about the nature of the claims made by the modeling efforts.

p0550 The first challenge emerges from a range of responses of dopamine neurons to stimuli not clearly related to reward prediction. For example, novel stimuli have been shown to cause phasic bursts in dopamine neurons (Schultz, 1998) including even nociceptive stimuli (Coizet *et al*., 2006). By virtue of being novel, such stimuli should not be predictive of any outcome, aversive or appetitive. However, learning and perception are not done on the background of a blank slate. It is reasonable to think that generalization to previously encountered stimuli would play a critical role in the initial appraisal of a novel stimulus. If the experimental (or the general ecological) scenario is such that animals have learned to expect that stimuli predict rewards (as is the case in many experimental situations), it is not surprising that new stimuli will be treated optimistically. Kakade and Dayan (2002) directly addressed this possibility, and furthermore suggested that the novelty responses can function as novelty bonuses that enhance exploration of novel stimuli. In this account, novelty itself acts as a reward and combines with current reward information $r(t)$ to form a reward signal sensitive to novelty, that is, $r_{new}(t) = r(t) + novelty(S_t)$. Kakade and Dayan show how this simple maneuver accounts in detail for the reported novelty responses of dopamine neurons (for instance, for the observation that the novelty burst is frequently followed immediately by a dip of the firing rate below baseline) yet still explains how they also communicate a reward prediction error.

p0560 One would hope that this rather innocuous-looking change in the model did not change appreciably what it learns. In fact, as shown by Ng *et al*. (1999), it does not, and Kakade and Dayan demonstrate clearly how to apply these results to the anomalous dopaminergic data. One issue brought to the forefront by this work is whether the dopamine system responds to aversive stimuli, and whether there should be an opponent system interacting with the information encoded in dopaminergic activity. This possibility is currently under development (Daw and Touretzky, 2000; Daw *et al*., 2002; Doya, 2002).

p0570 A second challenge for RL models in explaining dopaminergic function has arisen primarily due to the complaint that the putative reward-prediction error responses are too short-lived to account for the kinds of learning that they direct (Redgrave *et al*., 1999; Redgrave and Gurney, 2006). The basis for this challenge appears to relate to the issue of how a short burst of dopaminergic activity, and its putative translation into dopamine release in target areas, could account for physiological changes (like no dopaminergic response to a future expected reward) well into the future. To our understanding this is not a relevant challenge to RL models of dopamine function: in the models there is a clear distinction between the carriers of state value (which bridge the temporal gaps) and the reward-prediction errors (which are phasic). While the latter function is ascribed to dopamine, the former never was. Rather, it is presumed to be carried by cortical and/or striatal neurons that show sustained firing.

p0580 A related issue has emerged due to comparisons across vastly different experimental methods and time-scales. Longer-term changes in dopamine, as measured by a technique known as microdialysis, are marshaled as evidence for dopamine's role in what has been called *incentive salience*. Much has been written on this subject, but the argument is well summarized in Berridge (2007). Here, we would like to clarify what we think is limiting about the nature of the debate. In his summary of the issue, Kent Berridge states

> Debate continues over the precise causal contribution made by mesolimbic dopamine systems to reward. There are three competing explanatory categories: "liking", learning, and "wanting". Does dopamine mostly mediate the hedonic impact of reward ("liking")? Does it instead mediate learned predictions of future reward, prediction error teaching signals and stamp in associative links (learning)? Or does dopamine motivate the pursuit of rewards by attributing incentive salience to reward-related stimuli ("wanting")?
>
> (Berridge, 2007: 391).

p0590 In our view, the confusion here derives from at least three clear sources: (1) setting up the problem as though these separate questions are mutually exclusive; (2) comparing qualitative explanations in psychological terms like "wanting" and "liking" to quantitative models that match dopamine spike data to differential equations; and (3) a comparison of dopamine changes at vastly different time-scales. In Berridge's review, microdialysis measurements of dopamine levels under various behavioral challenges are compared to the reward-prediction error models that account for spike data. To be clear about the scope of the RL model of dopaminergic function, it applies strictly to rapid transients in spike rates in the 50–250 millisecond range and does not apply to other time-scales of dopaminergic modulation that may well carry other information important for cognitive processing and behavioral control. The examples above analyzed by Kakade and Dayan (2002) illustrate this issue. For example, the temporal-difference model is agnostic with regard to baseline dopamine levels or even fluctuations on slightly slower time-scales (e.g., minutes to hours). Consequently, the model would not account for microdialysis results whose measurements lie in these temporal regimes.

p0600 The last challenge is really an area of opportunity. Creatures ranging from sea slugs (e.g., *Aplysia*) to humans appear to be equipped with reinforcement learning systems – systems that broadcast some kind of evaluation signal to widespread regions of the nervous system and influence learning and decision-making. A large body of neurobiological and behavioral data support such a position. In general, RL systems can be quite fast and efficient at learning, provided that the creature is pre-equipped with representations appropriate to the RL problems that it will face (see Dayan and Abbott, 2001). In the absence of appropriate representations, RL systems often perform miserably. Not much is known about how RL problems are represented in the human brain or how such representations should be modified by experience. This is an open future challenge for RL models; that is, to design experimental probes that reveal the structure of underlying representations.

CONCLUSION s0140

p0610 To summarize, computational models of learning have done much to advance our understanding of decision making in the last couple of decades. Temporal-difference reinforcement learning models have provided a framework for optimal online model-free learning, which can be used by animals and humans interacting with the environment in order to learn to

predict events in the future and to choose actions such as to bring about those events that are more desirable. Investigations into the decision-making behavior of both animals and humans support the existence of such a mechanism, controlling at least some types of rational behavior. The prediction-error hypothesis of dopamine has further linked these algorithmic ideas to possible underlying neural substrates, specifically, to learning and action selection in the basal ganglia modulated by phasic dopaminergic signals. Converging evidence from a wide variety of recording and imaging methods supports this hypothesis.

p0620 It seems that reinforcement learning has been most powerful (unfortunately for neuroscience, almost unique) in tying together the three levels – computation, algorithm, and implementation (Marr, 1982) – into one coherent framework that is used not only for gleaning understanding, but also for shaping the next generation of experimental investigations. Whether the theory can be elaborated to account for results of future experimentation without losing its simplicity and elegance, or whether it is eventually abandoned and replaced by a newer generation of computational learning theories, reinforcement learning has already left its permanent mark on the study of decision-making in the brain.

s0150 ## Acknowledgments

p0630 This work was supported by the Human Frontiers Science Program (YN), The Kane Family Foundation (PRM), The Dana Foundation and Autism Speaks (PRM), National Institute on Drug Abuse (R01 DA11723 to PRM), National Institute of Neurological Disorders and Stroke (R01 NS045790 to PRM), and The Angel Williamson Imaging Center(PRM).

References

Baird, L.C. (1995). Residual algorithms: reinforcement learning with function approximation. In: A. Prieditis and S. Russell (eds), *Proceedings of the 12th International Conference on Machine Learning* (IMLL 95). San Mateo, CA: Morgan Kaufman, pp. 30–37.

Barto, A.G. (1995). Adaptive critic and the basal ganglia. In: J.C. Houk, J.L. Davis, and D.G. Beiser (eds), *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press, pp. 215–232.

Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Systems Man Cyber.* 13, 834–846.

Barto, A.G., Sutton, R.S., and Watkins, C.J.C.H. (1989). Sequential decision problems and neural networks. In: D.S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*. Cambride, MA: MIT Press, pp. 686–693.

Barto, A.G., Sutton, R.S., and Watkins, C.J.C.H. (1990). Learning and sequential decision making. In: M. Gabriel and J. Moore (eds), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge, MA: MIT Press, pp. 539–602.

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

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

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221.

Bellman, R.E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bergstrom, B.P. and Garris, P.A. (2003). "Passive stabilization" of striatal extracellular dopamine across the lesion spectrum encompassing the presymptomatic phase of Parkinson's disease: a voltammetric study in the 6-OHDA lesioned rat. *J. Neurochem.* 87, 1224–1236.

Berns, G.S., McClure, S.M., Pagnoni, G., and Montague, P.R. (2001). Predictability modulates human brain response to reward. *J. Neurosci.* 21, 2793–2798.

Berridge, K.C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacol. (Berl.)* 191, 391–431.

Berridge, K.C. and Robinson, T.E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Rev.* 28, 309–369.

Bertsekas, D.P. and Tsitsiklis, J.N. (1996). *Neuro-dynamic Programming*. London: Athena.

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

Christoph, G.R., Leonzio, R.J., and Wilcox, K.S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *J. Neurosci.* 6, 613–619.

Coizet, V., Dommett, E.J., Redgrave, P., and Overton, P.G. (2006). Nociceptive responses of midbrain dopaminergic neurones are modulated by the superior colliculus in the rat. *Neuroscience* 139, 1479–1493.

Daw, N.D. and Touretzky, D.S. (2000). Behavioral results suggest an average reward TD model of dopamine function. *Neurocomputing* 32, 679–684.

Daw, N.D. and Touretzky, D.S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation* 14, 2567–2583.

Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks* 15, 603–616.

Daw, N.D., O'Doherty, J.P., Dayan, P. *et al*. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879.

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

Dayan, P. and Abbott, L.F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

Dayan, P. and Kakade, S. (2000). Explaining away in weight space. In: T. Leen, T. Dietterich, and V. Tresp (eds), *Advances in Neural Information Processing Systems*, Vol. 12. Cambridge, MA: MIT Press, pp. 24–30.

Dayan, P., Kakade, S., and Montague, P.R. (2000). Learning and selective attention. *Nat. Neurosci.* 3, 1218–1223.

Delgado, M.R., Locke, H.M., Stenger, V.A., and Fiez, J.A. (2003). Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. *Cogn. Affect. Behav. Neurosci.* 3, 27–38.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation* 12, 219–245.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks* 15, 495–506.

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

Floresco, S.B., West, A.R., Ash, B. *et al*. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nat. Neurosci.* 6, 968–973.

Franklin, K.B.J. and McCoy, S.N. (1979). Pimozide-induced extinction in rats: stimulus control of responding rules out motor deficit. *Pharmacol. Biochem. Behav.* 11, 71–75.

Geisler, S. and Zahm, D.S. (2005). Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *J. Comp. Neurol.* 490, 270–294.

Goto, Y. and Grace, A.A. (2005). Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nat. Neurosci.* 8, 805–812.

Grace, A.A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience* 41, 1–24.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* 26, 8360–8367.

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

Horvitz, J.C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656.

Houk, J.C., Adams, J.L., and Barto, A.G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: J.C. Houk, J.L. Davis, and D.G. Beiser (eds), *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press, pp. 249–270.

Howard, R.A. (1960). *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press.

Ikemoto, S. and Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Res. Rev.* 31, 6–41.

Jensen, J., Smith, A.J., Willeit, M. *et al*. (2007). Separate brain regions code for salience vs valence during reward prediction in humans. *Hum. Brain Mapp.* 28, 294–302.

Joel, D. and Weiner, I. (1999). Striatal contention scheduling and the split circuit scheme of basal ganglia-thalamocortical circuitry: from anatomy to behaviour. In: R. Miller and J. Wickens (eds), *Conceptual Advances in Brain Research: Brain Dynamics and the Striatal Complex*. New York, NY: Harwood Academic Publishers, pp. 209–236.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks* 15, 535–547.

Kacelnik, A. (1997). Normative and descriptive models of decision making: time discounting and risk sensitivity. In: G.R. Bock and G. Cardew (eds), *Characterizing Human Psychological Adaptations: Ciba Foundation Symposium 208*. Chichester: Wiley, pp. 51–70.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks* 15, 549–559.

Kamin, L.J. (1969). Predictability, surprise, attention, and conditioning. In: B.A. Campbell and R.M. Church (eds), *Punishment and Aversive Behavior*. New York, NY: Appleton Century Crofts, pp. 242–259.

Kehoe, E.J. (1977). Effects of serial compound stimuli on stimulus selection in classical conditioning of the rabbit nictitating membrane response. PhD thesis, university of Iowa.

Knutson, B. and Gibbs, S.E.B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacol (Berl.)* 191, 813–822.

Knutson, B., Adams, C.M., Fong, G.W., and Hommer, D. (2001a). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* 21, RC159.

Knutson, B., Fong, G.W., Adams, C.M. *et al*. (2001b). Dissociation of reward anticipation and outcome with event-related fmri. *NeuroReport* 12, 3683–3687.

Knutson, B., Fong, G.W., Bennett, S.M. *et al*. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fmri. *NeuroImage* 18, 263–272.

Kobayashi, Y. and Okada, K.-I. (2007). Reward prediction error computation in the pedunculo-pontine tegmental nucleus neurons. *Ann. N.Y. Acad. Sci.* 1104, 310–323.

Konda, V.R. and Tsitsiklis, J.N. (2003). On actor-critic algorithms. *SIAM J. Control Optimization* 42, 1143–1166.

Konorski, J. (1948). *Conditioned Reflexes and Neuron Organization*. New York, NY: Cambridge University Press.

Kremer, E.F. (1978). The Rescorla-Wagner model: losses in associative strength in compound conditioned stimuli. *J. Exp. Psychol. Animal Behav. Proc.* 4, 22–36.

Lewicki, M.S. and Olshausen, B.A. (1999). A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A* 16, 1587–1601.

Li, J., McClure, S.M., King-Casas, B., and Montague, P.R. (2006). Policy adjustment in a dynamic economic game. *PLoS ONE* 1, e103.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopaminergic neurons during learning of behavioral reactions. *J. Neurophysiol.* 67, 145–163.

Logothetis, N.K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *J. Neurosci.* 23, 3963–3971.

Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proc. Nat. Acad. Sci. USA* 104, 9493–9498.

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, CA: Freeman & Co.

Matsumoto, M. and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447, 1111–1115.

McClure, S.M., Daw, N.D., and Montague, P.R. (2003a). A computational substrate for incentive salience. *Trends Neurosci.* 26, 423–428.

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

McClure, S.M., Li, J., Tomlin, D. *et al*. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* 44, 379–387.

Menon, M., Jensen, J., Vitcu, I. *et al*. (2007). Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: effects of dopaminergic modulation. *Biol. Psych.* 62, 765–772.

Miller, R. and Wickens, J.R. (1991). Corticostriatal cell assemblies in selective attention and in representation of predictable and controllable events. *Concepts Neurosci.* 2, 65–95.

Mirenowicz, J. and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* 379, 449–451.

Montague, P.R., Dayan, P., Nowlan, S.J. *et al*. (1993). Using aperiodic reinforcement for directed self-organization. In: C.L. Giles, S.J. Hanson, and J.D. Cowan (eds), *Advances in Neural Information Processing Systems*, Vol. 5. San Mateo, CA: Morgan Kaufmann, pp. 969–976.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1994). Foraging in an uncertain environments using predictive hebbian learning. In: Tesauro and J.D. Cowan (eds), *Advances in Neural Information Processing Systems*, Vol. 6. San Mateo, CA: Morgan Kaufmann, pp. 598–605.

Montague, P.R., Dayan, P., Person, C., and Sejnowski, T.J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature* 377, 725–728.

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

Montague, P.R., McClure, S.M., Baldwin, P.R. *et al*. (2004). Dynamic gain control of dopamine delivery in freely moving animals. *J. Neurosci.* 24, 1754–1759.

Morris, G., Arkadir, D., and Nevet, A. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43, 133–143.

Morris, G., Nevet, A., Arkadir, D. *et al*. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* 9, 1057–1063.

Ng, A.Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, pp. 278–287.

Nicola, S.M., Surmeier, J., and Malenka, R.C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu. Rev. Neurosci.* 23, 185–215.

Niv, Y. (2007a). Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? *Ann. NY Acad. Sci.* 1104, 357–376.

Niv, Y. (2007b). The Effects of Motivation on Habitual Instrumental Behavior. Unpublished doctoral dissertation, The Hebrew University of Jerusalem.

Niv, Y., Daw, N.D., and Dayan, P. (2005a). How fast to work: response vigor, motivation and tonic dopamine. In: Y. Weiss, B. Schölkopf, and J. Platt (eds), *Advances in Neural Information Processing Systems*, Vol. 18. Cambridge, MA: MIT Press, pp. 1019–1026.

Niv, Y., Duff, M.O., and Dayan, P. (2005b). Dopamine, uncertainty and TD learning. *Behav. Brain Func.* 1, 6.

Niv, Y., Joel, D., and Dayan, P. (2006a). A normative perspective on motivation. *Trends Cogn. Science* 10, 375–381.

Niv, Y., Daw, N.D., and Dayan, P. (2006b). Choice values. *Nat. Neurosci.* 9, 987–988.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacol. (Berl.)* 191, 507–520.

Nomoto, K., Watanabe, T., and Sakagami, M. (2007). Dopamine responses to complex reward-predicting stimuli. *Soc. Neurosci. Abst.* 33, 749.5.

O'Doherty, J.P., Deichmann, R., Critchley, H.D., and Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron* 33, 815–826.

O'Doherty, J., Dayan, P., Friston, K. *et al*. (2003). Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron* 38, 329–337.

O'Doherty, J.P., Dayan, P., Schultz, J. *et al*. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454.

Pagnoni, G., Zink, C.F., Montague, P.R., and Berns, G.S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nat. Neurosci.* 5, 97–98.

Pessiglione, M., Seymour, B., Flandin, G. *et al*. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442, 1042–1045.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51, 381–390.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975.

Redgrave, P., Prescott, T.J., and Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.* 22, 146–151.

Rescorla, R.A. (1970). Reduction in effectiveness of reinforcement after prior excitatory conditioning. *Learning Motiv.* 1, 372–381.

Rescorla, R.A. and Lolordo, V.M. (1968). Inhibition of avoidance behavior. *J. Comp. Physiol. Psychol.* 59, 406–412.

Rescorla, R.A. and Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: A.H. Black and W.F. Prokasy (eds), *Classical Conditioning II: Current Research and Theory*. New York, NY: Appleton Century Crofts, pp. 64–99.

Reynolds, G.S. (1961). Attention in the pigeon. *J. Exp. Anal. Behav.* 4, 203–208.

Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neurosci.* 10, 1615–1624.

Romo, R. and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-intiated arm movements. *J. Neurophysiol.* 63, 592–606.

Salamone, J.D. and Correa, M. (2002). Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behav. Brain Res.* 137, 3–25.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3, 210–229.

Schönberg, T., Daw, N.D., Joel, D., and O'Doherty, J.P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.* 27, 12860–12867.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron* 36, 241–263.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.* 12, 4595–4610.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during succesive steps of learning a delayed response task. *J. Neurosci.* 13, 900–913.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.

Schwartz, A. (1993). Thinking locally to act globally: a novel approach to reinforcement learning. In: *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 906–911.

Seymour, B., O'Doherty, J.P., Dayan, P. *et al*. (2004). Temporal difference models describe higher order learning in humans. *Nature* 429, 664–667.

Sutton, R.S. (1978). A Unified Theory of Expectation in Classical and Instrumental Conditioning. Unpublished Bsc thesis, Stanford University.

Sutton, R.S. (1988). Learning to predict by the method of temporal difference. *Machine Learning* 3, 9–44.

Sutton, R.S. and Barto, A.G. (1990). Time-derivative models of Pavlovian reinforcement. In: M. Gabriel and J. Moore (eds), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge, MA: MIT Press, pp. 497–537.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Takikawa, Y., Kawagoe, R., and 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.

Tobler, P.N., Dickinson, A., and 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., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. *Science* 315, 515–518.

Ungless, M.A., Magill, P.J., and Bolam, J.P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science* 303, 2040–2042.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48.

Watkins, C.J.C.H. (1989). Learning with Delayed Rewards. Unpublished doctoral dissertation, Cambridge University, Cambridge.

Weiner, I. and Joel, D. (2002). Dopamine in schizophrenia: dysfunctional information processing in basal ganglia-thalamocortical split circuits. In: G.D. Chiara (ed.), *Handbook of Experimental Pharmacology* Vol. 154/II, *Dopamine in the CNS II*. Berlin: Springer Verlag, pp. 417–472.

Werbos, P.J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook* 22, 25–38.

Wickens, J. and Kötter, R. (1995). Cellular models of reinforcement. In: J.C. Houk, J.L. Davis, and D.G. Beiser (eds), *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press, pp. 187–214.

Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8, 229–256.

Willner, P., Towell, A., and Muscat, R. (1987). Effects of amphetamine and pimozide on reinforcement and motor parameters in variable-interval performance. *J. Psychopharmacol.* 1, 140–153.

Wise, R.A. (1982). Neuroleptics and operant behavior: the anhedonia hypothesis. *Behav. Brain Sci.* 5, 39–53.

Wise, R.A. (2004). Dopamine, learning and motivation. *Nat. Rev. Neurosci.* 5, 483–495.

Wise, R.A., Spindler, J., de Wit, H., and Gerberg, G.J. (1978a). Neuroleptic-induced "anhedonia" in rats: pimozide blocks reward quality of food. *Science* 201, 262–264.

Wise, R.A., Spindler, J., and Legault, L. (1978b). Major attenuation of food reward with performance-sparing doses of pimozide in the rat. *Can. J. Psychol.* 32, 77–85.

Yu, A.J. and Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks* 15, 719–730.

Yu, A.J. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron* 46, 681–692.

Zald, D.H., Boileau, I., El-Dearedy, W. *et al*. (2004). Dopamine transmission in the human striatum during monetary reward tasks. *J. Neurosci.* 24, 4105–4112.