CHAPTER

16

Advanced Reinforcement Learning

Nathaniel D. Daw

Box 16.1: Yael Niv

OUTLINE

Introduction 299

The RL Formalism 300
Markov Decision Processes 300
Values, Policies, and Optimal Policies 300

Learning 301
Learning Rules 301
Learning Rates and Uncertainty 302

Rewards and Punishments 306
The Subjectivity of Reward 306
The Construction of Reward 306
Punishment and Avoidance 307
Dopamine and Punishment 307

States, Stimuli, and Perceptual Uncertainty 308
Theory: Partial Observability and Perceptual Inference 310
From Perception to Decision 311
POMDPs and Neuroscience 312

Actions 313
Vigor, Opportunity Costs, and Tonic Dopamine 313
Action Hierarchies, Action Sequences, and Hierachical RL 314
Multi-Effector Action 316

Conclusion 317

References 317

INTRODUCTION

This chapter takes a deeper look at reinforcement learning (RL) theories and their role in neuroeconomics. The previous chapter described a prominent and well-studied hypothesis about a neural and computational mechanism for learning to choose rewarding actions, centered on the midbrain dopamine system and its targets, particularly in the striatum (Houk *et al.*, 1995; Montague *et al.*, 1996; Schultz *et al.*, 1997). That chapter described how the phasic firing of dopamine neurons appears to report a *prediction error* that would be appropriate for updating expectations about long-term future reward. It described the evidence that this signal may drive learning about action preferences, by affecting plasticity at synapses onto the medium spiny neurons of striatum, which have also long been believed to be involved in movement initiation and execution. Psychologically, as detailed in Chapter 21, this mechanism appears to implement a well-studied category of behavior known as habitual learning. Computationally, the dopamine response closely resembles the prediction error from temporal-difference (TD) learning, an algorithm for learned optimal control from computer science. Because of this correspondence, this theory offers a direct line from the hypothesized neural and psychological mechanisms to *normative* considerations about how an efficient agent *should* choose, a particularly important level of understanding from a neuroeconomic perspective.

Working outward from the relatively secure core of dopamine and TD learning, this chapter considers extensions and areas of current investigation. In particular, after reviewing the RL theories in greater formal detail than the preceding chapter, we consider a number of problems that arise in matching up the theory's abstract elements — learning updates, rewards, states, and actions — to more realistic ones relevant to an experimental or real-world situation faced by a biological organism. We approach these questions starting from the normative, computational perspective — how an optimizing agent should learn — and in each case

Neuroeconomics. DOI: http://dx.doi.org/10.1016/B978-0-12-416008-8.00016-4

review theory from the computational study of RL that can serve as a framework for conceptualizing how the brain might approach the problem. In each of these examples, the resulting theory preserves the TD prediction error mechanism at the heart of a more complex and realistic account.

THE RL FORMALISM

Markov Decision Processes

We begin by detailing, more formally than in the previous chapter, the problem solved by RL algorithms such as temporal-difference learning (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Laying out these formalisms will allow us to expose the correspondence between abstract elements of the theory and aspects of real-world decisions by biological organisms. Making different parts of this mapping between theory and experiment work raises a number of questions and necessitates a number of elaborations, which are ultimately the topics of this chapter and of much current research in neuroeconomics.

RL algorithms have primarily been developed (within computer science) to identify optimal decisions in a formal class of tasks known as *Markov decision processes* (MDPs). MDPs are a class of decision-problem stripped down enough to be amenable to fairly straightforward mathematical analysis, while still covering a broad range of tasks and preserving a number of the elements of nontrivial real world decisions. The core elements of an MDP are a set of situations or *states*, $\mathscr{S}$, and a set of *actions*, $\mathscr{A}$, plus a specification of transitions (how states and actions lead to other states) and rewards.

Within the framework of MDPs, the "world" has what are called *discrete dynamics* (as opposed to continuous time dynamics): The world's state takes on a new value from $\mathscr{S}$ at each timestep, t (we write this as a random variable s_t, s_{t+1}, etc.). At each timestep, the agent also chooses an action a_t from $\mathscr{A}$ and receives a reward r_t, which measures the utility received on that timestep. We take the reward to be a real number and assume that it is a function of the state and (for notational simplicity) deterministic: $r_t = r(s_t)$.

The actions are important because they influence the evolution of the state, and hence the obtained rewards. Specifically, the state at any time $t+1$, s_{t+1}, is a (probabilistic) function of the preceding state, s_t, and action, a_{t}, determined by a transition distribution $P(s_{t+1}|s_t, a_t)$. The most important simplifying assumption of the MDP — the Markov property for which it is named — is that this state transition probability depends only upon the current state and action; conditional on these, the new state is independent of all earlier states and actions. Rewards also obey the Markov property, since they depend only on the current state and, conditional on this, are independent of any earlier history. By constraining the relationships between events across time, the Markov conditional independence property simplifies analysis, learning, and decision making and is key to RL algorithms. (In particular, it allows formulating the Bellman equation, Equations 2 and 3 below.) But as we will discuss later in the chapter, it also raises some difficulties relating RL states to the sensory experiences of organisms, which typically do not obey the Markov property.

Despite these limitations, many decision tasks can be characterized as MDPs, including Tetris, American football, and elevator scheduling (see Sutton and Barto, 1998, for many examples). Modeled as MDPs, different decision problems correspond to different state and action sets, and different transition and reward functions over them.

Values, Policies, and Optimal Policies

We are now in a position to evaluate decisions based on how much reward they obtain. As discussed in the previous chapter, the key complication for decision making in this setting is that each action has long-term consequences for the decision-maker's reward prospects, via its influence on the successor state, s_{t+1}, and thence indirectly on all subsequent states and rewards. Accordingly, just as a particular choice of play in American football may not itself score points but may set up field position that influences subsequent scoring potential, to choose actions in an MDP we must take account of both the immediate and deferred consequences of actions.

Let us define the decision variable — the quantity we wish our choices to optimize — as the expected cumulative, discounted future reward. This quantity, called the *state-action value function*, depends not only on the current state and action, but also on the actions we take in subsequent states. First, let us define a *policy* $\pi(s)$ as any mapping from states to actions. We write the value of taking an action in a state, and then following some policy π thereafter:

$$Q^{\pi}(s_t, a_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} \ldots | s_t, a_t, \pi] \tag{16.1}$$

Here, elaborating the similar equations in the previous chapter, we have discounted future rewards exponentially in their delay by the discounting parameter $\gamma \leq 1$. We have also used the notation $E[\cdot | s_t, a_t, \pi]$ to stand in for the complicated expectation over all possible future sequences of states and rewards given the starting state and action, and the policy π.

We can make that expectation over future states more explicit by rewriting Equation 16.1 in recursive form (the Bellman equation; Bellman, 1957):

$$Q^{\pi}(s_t, a_t) = r(s_t) + \gamma \sum_{s_{t+1} \in \mathscr{S}} P(s_{t+1}|s_t, a_t) Q^{\pi}(s_{t+1}, \pi(s_{t+1})) \tag{16.2}$$

This form of the value function expresses the series of rewards from Equation 16.1 as the first, $r(s_t)$, plus all the rest. The trick is that the cumulative value of "all the rest" is also given by the value function, but evaluated at the successor state s_{t+1}. This value is discounted and averaged over possible successors according to their probabilities. Thus the value function is defined in terms of the recursive relationship between the values at different states. See the previous chapter for a more detailed discussion of such recursions and their relevance to neuroeconomics; our main goal here is to characterize what it means to choose optimally in this setting.

Since the expected future value at each state is a function of the policy π, we need to consider optimality over policies rather than actions individually. As it turns out, for any MDP there exists at least one deterministic *optimal policy* π^* which is globally best in the sense that at every state, its expected future reward $Q^*(s, \pi^*(s))$ is at least as good as that for any other policy. (See, e.g., Puterman, 1994, for details.) We can define the optimal value function, and its associated policy, again recursively by using a form of the Bellman equation that explicitly chooses the best action in each state on the right side of the equation:

$$Q^*(s_t, a_t) = r(s_t) + \gamma \sum_{s_{t+1} \in \mathscr{S}} P(s_{t+1}|s_t, a_t) \max_{a_{t+1} \in \mathscr{A}} Q^*(s_{t+1}, a_{t+1}) \tag{16.3}$$

The optimal policy, $\pi^*(s) = \arg\max_a[Q^*(s, a)]$, is then given by the assignment, to each state, of the highest-valued action. That the optimal value function defines the optimal policy is a formal instantiation of the intuitive strategy of choosing actions by predicting their long-run rewards, which occupied much of the previous chapter and underlies much theorizing about the role of dopamine. After all, the difficulty in selecting actions in an MDP is that they have long-term effects. However, these consequences are exactly what the value function Q^* measures. If you can learn it, then selecting the best action in a state is as simple as comparing its value between candidates and choosing the best.

In the rest of this chapter, we attempt to put elements of this abstract theory back into a biological and psychological context. We begin with mechanisms for prediction learning, and then consider the real world counterparts of the main components of MDP's: rewards, states, and actions.

LEARNING

Learning Rules

The basic strategy assumed by RL theories in neuroscience is that organisms learn state-action values by trial and error, and use these as decision variables to guide choice. Starting from Equation 16.3, two important approaches present themselves.

The first is the one described in the previous chapter, and widely associated with dopamine. This is to maintain internal estimates of $Q(s, a)$ for all states and actions, and update these according to a temporal-difference prediction error (Sutton, 1988). We define the prediction error as the extent Equation 16.3 fails to hold in a particular state-action-state sequence, by taking the difference between the two sides of the equation and replacing the expectation over possible successor states with the s_{t+1} actually observed:

$$\delta_t = r_t + \gamma \max_{a_{t+1} \in \mathscr{A}} Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \tag{16.4}$$

We can use this prediction error with the update rule $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \cdot \delta_t$ to improve our estimates, and ultimately (under various technical conditions) to converge on the true optimal values Q^*. This version of the algorithm is called *Q-learning* (Watkins and Dayan, 1992). A few variations on this theme are sometimes seen in the literature. For instance, a related temporal-difference algorithm called *SARSA* can be derived from the form of the Bellman equation in Equation 16.2, thereby replacing the max over actions in Equation 16.4 with the action actually taken, in order to learn the policy-specific rather than the optimal state-action values (Rummery and Niranjan, 1994). Another variant, called the *actor-critic*, is based on the Bellman equation for the state value V (see Chapter 15), and learns both a state value function $V^{\pi}(s)$ and a corresponding action selection policy $\pi(s)$ (Barto, 1995). Here, the values and the policy are both updated by a temporal-difference prediction error similar to Equation 16.4. Importantly, all these three approaches share the essential strategy of learning their decision variables using a temporal-difference prediction error based on some form of the Bellman equation. (More can be learned about each in Sutton and Barto's classic textbook, Sutton and Barto, 1998.)

Owing to their roots in the Bellman equation, the key feature of these theories is that they use their own estimates of the reward expected from a state ($Q_t(s_{t+1}, a_{t+1})$ in Equation 16.4) to train the reward predictions for the states that preceded it. As discussed extensively in Chapter 15, this feature, called *bootstrapping*, has a number of important echoes in behavioral and neural data, including the anticipatory responses of dopamine neurons. Thus, although there have been some efforts to distinguish which particular version of the temporal-difference prediction error best corresponds to the dopamine response (Morris *et al.*, 2006; Roesch *et al.*, 2007), the larger point is that the sort of empirical and computational considerations discussed in the previous chapter connect the dopaminergic system to this family of RL algorithms. Collectively, these algorithms are known as *model-free RL*.

A distinctly different approach to RL, which does not involve prediction errors similar to dopaminergic responses but may also be relevant in neuroeconomics, is to focus on learning the state transition and reward functions, $P(s_{t+1}|s_t, a_t)$ and $r(s_t)$, which together define an MDP. This approach is called ***model-based RL*** because those two functions constitute an "internal model" or characterization of the task contingencies. Equation 16.3 defines the values Q^* in terms of these quantities (and Q^*, in turn, defines the optimal policy), and so given the model you can use Equation 16.3 to compute the values and derive a policy. The transition and reward distributions are fairly straightforward to learn: it is easy to directly observe an example of a one-step state transition or reward, and to average many such examples to estimate the functions. (In contrast, you can't easily collect samples of long-run state-action values Q since they accumulate rewards that unfold over many steps. This is why learning Q directly via the model-free methods above requires bootstrapping or other tricks.) The flip side of the simplicity in learning an internal model is computational complexity in using it: in order to recover the predicted long-run values, it is necessarily to explicitly evaluate Equation 16.3. To see how this can be done, note that Q^* is on both sides of Equation 16.3; if you substitute the right hand side of the equation into itself for Q^* repeatedly, you "unroll" the recursion into a series of nested sums taking expectations over the series of future states and rewards. The standard algorithm for computing values from the transition and reward models, called *value iteration*, essentially corresponds to evaluating this expectation stepwise (Sutton and Barto, 1998).

Although model-free RL has received the majority of attention in neuroscience, due to its relationship with the dopamine system, there has been an increasing understanding that the brain also uses model-based methods (Daw *et al.*, 2005). This is the primary topic of Chapter 21.

Note a confusing terminological issue: as used in computer science and neuroscience – and in this book – the term "reinforcement learning" refers broadly to learning in the context of decision problems, and comprises many particular sorts of learning including both the model-free and model-based approaches discussed above. Economists, in contrast, sometimes use the term "reinforcement learning" to refer more specifically to one particular approach to learning, essentially the model-free strategy.

Learning Rates and Uncertainty

At the heart of model-free RL as used in theories of dopamine is the concept of value learning by an error-driven update, e.g. $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + Q_t(s_t, a_t) + \alpha \cdot \delta_t$. As discussed in Chapter 15, such a rule seems sensible, at least informally: it fractionally nudges the prediction Q in the direction that reduces the prediction error δ. But can we rationalize this procedure on more formal grounds? And in particular, can we give some principled interpretation to the so-far arbitrary learning rate parameter α?

It is familiar in economics and computer science to frame learning in terms of statistical reasoning. On this view, learning some quantity (here, Q) from a series of noisy observations is really just a statistical estimation problem. That is, if we specify formal assumptions about the statistical structure of the noise, then our estimates given our observations at each step (in effect, the learning rule) follow directly from the rules of probability (Dayan and Long, 1998). In particular, learning at each step requires combining what we previously knew about the value with new evidence from the current observation. The rules of probability determine how optimally to weight such sources of information when combining them, in effect, determining the appropriate step size α (Dayan *et al.*, 2000; Kakade and Dayan, 2002).

To flesh out how these ideas relate to error-driven learning, we sketch a statistical counterpart (Kakade and Dayan, 2002) to the Rescorla–Wagner rule (Rescorla and Wagner, 1972; Equation 16.1 from Chapter 15) for classical conditioning. This is a simpler example than TD in an MDP, but preserves many key elements. In this setting, we encounter a series of trials indexed k, each with a stimulus s_k followed by an associated reward r_k. We attempt to predict only the immediate reward given the state; here there is no accumulation of predicted rewards across trials, which is mainly what makes this example more tractable than the full MDP situation.

To reason statistically about this problem, we must make some assumptions about the noise. Assume that there is some *true* reward contingency, $V_k(s)$, associated with each stimulus, but that this is unknown and not directly observed because the obtained rewards are corrupted by Gaussian noise:

$$r_k = V_k(s_k) + \varepsilon_k \tag{16.5}$$

where $\varepsilon_k \sim N(0, \sigma_r)$.

Set up this way, the problem of reward prediction is just the problem of noisy measurement. If I have previously observed a number of rewards following stimulus, s, then I can estimate $V_k(s)$, but only up to a limited degree of accuracy (conversely, with some *uncertainty*) due to the measurement noise. One way to characterize such uncertainty is as a *distribution*, which assigns a probability (e.g., a degree of belief) to each possible value of V being the correct one conditional on the previously observed stimuli and rewards. The estimation problem here is constructed to ensure that these distributions will take the form of Gaussians, i.e., at each step, for each stimulus, s, we can express $P(V_k(s)|o_1\ldots o_{k-1}, s_1\ldots s_{k-1})$ as $N(\mu_k(s), \sigma_k(s))$.

The distribution (formally, the *posterior distribution*) describing our estimates about V has both a mean value $\mu_k(s)$, and a *variance* $\sigma_k^2(s)$, the latter characterizing uncertainty as the spread of belief away from the mean. This approach therefore generalizes the sorts of learning rules we have considered thus far to account explicitly for uncertainty about the learned estimate.

If we now observe a new trial with a stimulus s_k and reward r_k, then the new posterior distribution over $V_k(s_k)$ is given by the laws of probability, specifically *Bayes' rule*. We are combining two uncertain sources of information about the value, one being the previous estimate, and the other (from Equation 16.5) the new, noisy measurement. Intuitively, the weight given to each of these information sources in determining the updated posterior depends on their relative uncertainty (Figure 16.1). If I was sure already, a single noisy measurement won't move my belief much; conversely, if the measurement is much more precise than my previous beliefs, then it will largely replace them. This is an instance of a principle of Bayesian cue combination that is also prominent in other areas of psychology, such as perception, where it is often used to describe optimal combination between different modalities such as vision and audition (Knill and Pouget, 2004).

Formally, applied to this problem, Bayes' rule (together with identities for manipulating Gaussian distributions) implies update equations for the mean and variance of our posterior distribution given each new observation. Strikingly, the rule for the mean estimate, $\mu_k(s_k)$ takes exactly the familiar form of an error-driven update: $\mu_{k+1}(s_k) = \mu_k(s_k) + \kappa_k \cdot \delta_k$, with prediction error $\delta_k = r_k - \mu_k$.

It is worth stopping to consider what we have accomplished here. We have just produced a rational derivation justifying from first principles the sort of error-driven learning rule that is widely hypothesized in neuroeconomics on more empirical grounds. More importantly, this derivation sheds additional light on the problem because the new variable κ_k, plays the role of the learning rate α, but is no longer an arbitrary, free parameter. Instead, its value can be computed from Bayes' rule as:

$$\kappa_k = \frac{\sigma_k^2(s_k)}{\sigma_k^2(s_k) + \sigma_r^2(s_k)} \tag{16.6}$$

which reflects a comparison between the relative uncertainty of the previous beliefs, $\sigma_k^2(s_k)$, and the "noisiness" (or variability) in the rewards, $\sigma_r^2(s_k)$ (from Equation 16.5).

Equation 16.6 makes explicit the intuition from Figure 16.1: the learning rate is maximal when uncertainty about the value is high relative to small noise in the observed rewards, and the opposite situation

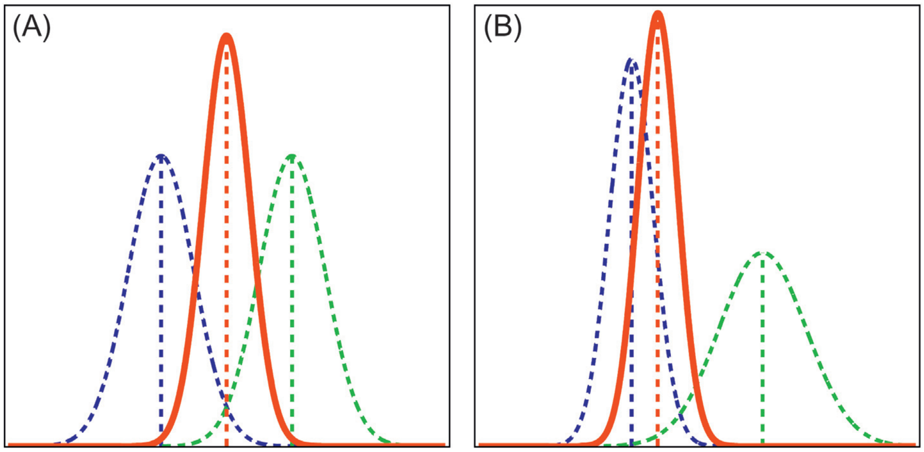

FIGURE 16.1 Bayesian cue combination weights evidence sources by their uncertainty. The blue and green Gaussians represent the distribution over the true value given by the previous evidence and the new observation, respectively. Each distribution has a mean (the vertical dashed line) and uncertainty, expressed as a variance around that mean. The posterior distribution arising from combining the two, in red, has a mean that interpolates between the two source means. These are weighted according to their uncertainty, such that if the two evidence sources are equally reliable (A) the resulting mean is their average, but if one is more uncertain than the other (B) the posterior mean is shifted toward the more reliable estimate.

results in low learning rates. The qualitative relationship illustrated here between uncertainty and learning rates is very general, though quantitatively fleshing out similar probabilistic approaches in other learning problems such as action value estimation in MDPs is quite complex (Dearden *et al.*, 1998; Daw *et al.*, 2005; Behrens *et al.*, 2007). In fact, it is important to stress that many real world problems, even those quite close to the "toy problem" described here can be intractable to this kind of analysis for what are largely technical reasons.

The theoretical dependence of the learning rate on the outcome noise, from Equation 16.6, may have a counterpart in the prediction error-related responses of dopamine neurons. Although dopamine neurons will generally respond more for larger unexpected rewards (i.e., larger prediction errors) this response scales relative to the range of rewards expected (Tobler *et al.*, 2005). In fact, the response to a small reward can be indistinguishable from that for a large reward, so long as both occur at the top of their respective ranges. One way to understand this effect (due to Preuschoff and Bossaerts, 2007) is to assume that the neurons report not the raw prediction error δ_k but instead the prediction error scaled by the learning rate, $\kappa_k \cdot \delta_k$, which is, anyway, the quantity that should ultimately control the amount of plasticity at recipient structures. The range of rewards can be thought of as corresponding to the outcome noise $\sigma_r^2(s_k)$, which scales the learning rate through Equation 16.6, and in this way would normalize the modeled dopaminergic responses.

The recognition of the key role of uncertainty, not just for controlling learning rates but for probabilistic reasoning in many other areas of perception and cognition, has driven a sustained but so far still reasonably speculative interest in possible mechanisms by which the brain may represent and manipulate uncertainties (Yu and Dayan, 2005; Ma *et al.*, 2006).

So far we have considered the update rule only for the mean of the value distribution. In the relatively simple example given here, the corresponding update rule for the uncertainty (variance) of the posterior is not very interesting: uncertainty starts high and declines toward zero as you collect more observations, leading to a decaying learning rate. However, the dynamics of uncertainty in more elaborate inference problems have a number of interesting psychological counterparts.

An important case arises if the true reward contingencies $V_k(s)$ are not stationary, but instead change over time, as is typical in the laboratory and likely often the case in real-world learning — for foraging animals different food patches may deplete and replenish, for animals in a lab experiment block transitions may shift the contingencies sporadically. A simple statistical assumption about such change (but one not justified in most laboratory environments) is that the true values fluctuate from trial-to-trial according to Gaussian random walks. Together with Equation 16.5, this assumption leads to a model called the Kalman filter (Kalman, 1960; Sutton, 1992; Kakade and Dayan, 2002). Here, the possibility that the values have changed between trials contributes additional uncertainty to the posterior distribution at each step, ensuring that the learner never becomes entirely certain about a stimulus' current value. This leads, asymptotically, to uncertainty stabilizing at the level where the information coming in (due to each observation) matches that "lost" due to the ongoing noisy change in the true values, in turn making the learning rate κ_k asymptotically constant over trials. These considerations clarify the constant learning rate often assumed in RL models: it is appropriate (asymptotically) in a non-stationary task of this sort.

A benefit of this kind of theoretical exercise is that it grounds a previously arbitrary parameter — the learning rate — in objective, experimentally manipulable and measurable aspects of the environment. For instance, in the Kalman filter, the *level* at which the learning rate asymptotes depends on the amount of trial—trial change in the true values (their *volatility*: the variance of their random walk), because this controls the asymptotic posterior uncertainty $\sigma_k^2(s_k)$ in Equation 16.6. Intuitively, the faster the value being learned is changing, the less relevant is previous experience (relative to new experience) in inferring its current value, so higher learning rates should be used to place more weight on new observations. This prediction has been tested and confirmed in RL experiments with humans, in which volatility was manipulated (Behrens *et al.*, 2007).

These last results, finally, point to the issue of *metalearning*, or estimating the parameters such as the volatility and outcome noise, which should determine uncertainty and learning rates. The Kalman filter takes these as given (e.g., the outcome noise appears as a constant in Equation 16.6), but subjects clearly have to learn them as well. This is possible by further elaborating statistical models in the spirit of the one described above to include an additional level of inference about the noise parameters. Following this approach, Behrens and colleagues. (2007) constructed trial-by-trial timeseries reflecting subjects hypothesized learning about value volatility and report that these covary with BOLD signals in the anterior cingulate cortex.

The relationship between learning rates and environmental volatility may also have a counterpart in an older literature from psychology on *associability* or the degree to which different stimuli in a conditioning

BOX 16.1

ELIGIBILITY TRACES

By Yael Niv

Eligibility traces (Barto *et al.*, 1981; Sutton and Barto, 1998) are a computational construct that is perhaps easier to justify from a biological, real-world learning perspective than from a theoretical one. The main idea is that every state that is visited by the agent or animal can remain eligible for updating for a certain period of time, so that prediction errors due to several forthcoming rewards and state transitions (and not just the immediate ones) can modify the state's value. In this way, a reward can easily modify the value of states and actions that occurred in the not-immediately recent past, allowing learning even with delayed outcomes.

In practice, rather than committing to a timeframe of eligibility, one can set an eligibility "trace" *e(s)* that decays exponentially over time:

$$e_{k+1}(s) = 1 \quad \text{for } s = s_k \text{ (the current state)}$$

$$e_{k+1}(s) = \lambda e_k(s) \quad \text{for all other states } s \neq s_k$$

(and similarly for state-action pairs). At each timestep, the values of all states are updated according to the prediction error multiplied by the state's eligibility trace, such that some states are "more eligible" for updating than others. The scheme above is what Sutton & Barto termed "replacing traces" as the trace for the visited state becomes 1, replacing its previous value. Another option is to add 1 to the (λ-decayed) previous trace of the visited state, thus generating "accumulating traces" (Sutton and Barto, 1998).

Temporal difference learning with eligibility traces is termed TD(λ), after the parameter that governs the decay of the eligibility trace, which must be between 0 and 1. At one extreme, if $\lambda = 0$ we get standard temporal difference learning (also known as TD(0)), as the current state is the only state eligible for updating at each timestep. At the other extreme of TD(1) (also known as Full Monte Carlo Learning), on every timestep all states that have ever been visited are updated, a scheme that is equivalent to using the full return (all the future rewards) to update the value of a state. These TD variants are all normative, that is, under the right conditions on learning rate and sampling, they can be shown to converge on the correct state values (Sutton and Barto, 1998). In general, the TD(λ) algorithm can be viewed as (exponentially) averaging learning with different-horizon returns, with the shorter returns being weighted more strongly. Thus the "forward view" of eligibility traces sees the traces as a mechanism that allows learning from future rewards. An equivalent "backward view" sees eligibility traces as a mechanism that allows each reward to affect not only the just-visited state, but also those states visited in the recent past (with "recency" decaying exponentially; Killeen, 2011).

In practice, eligibility traces can be seen as a memory mechanism that helps bridge gaps between events. This is extremely convenient for learning in the real world, as eligibility traces allow learning even if the state space is not strictly Markov (Loch and Singh, 1998; Singh *et al.*, 1994): values can be updated correctly even if non-Markov states intervene between an action and its consequences (Todd *et al.*, 2009). Consider, for instance, a classic trace-conditioning experiment: a rat sits in an experimental chamber for a length of time, then on some occasions a light turns on, turns off, and after two seconds a food pellet is delivered to the rat. The state of the world in which the rat is sitting in the box with no light on is thus ambiguous: if this state occurs before any light has turned on, the rat has little reason to expect reward to follow this state. However, if this state occurs after the *light on* state, the rat should expect that reward is forthcoming. Unless the rat represents these two situations as two unique states, the task is not a Markov one. Specifically, the occurrence of the intervening *no light* state will impair the rat's ability to learn to predict reward as a result of the light turning on. The light on state will be followed by a state that has low value (as it frequently leads to nothing) and thus will not acquire high value as befitting a situation that leads to reward with 100% certainty. However, with eligibility traces, it is clear that the value of the light on state will be updated to reflect the upcoming reward, suffering only from a learning rate that is reduced by a factor of λ.

The neuroscience of eligibility traces is rather straightforward: states can remain eligible for updating through prolonged activity of neurons representing a certain state (Seo *et al.*, 2007), and/or through any form of synaptic tagging that marks recently active synapses for future plasticity through LTP or LTD (Izhikevich, 2007), for instance, elevated levels of calcium in the dendritic spines of medium spiny neurons (Wickens and Kötter, 1995). Indeed, eligibility traces have been proposed to be integral to cerebellar learning (Wang *et al.*, 2000; McKinstry *et al.*, 2006). Bogacz and colleagues found behavioral evidence for eligibility traces in data from humans performing a decision-making task (Bogacz *et al.*, 2007), and an analysis of dopaminergic

BOX 16.1 (*cont'd*)

prediction errors in rats undergoing classical conditioning led Pan and colleagues (2005) to conclude that the measured dopamine firing patterns could only result from a system learning with a low learning rate and long-lasting eligibility traces. Interestingly, in both cases the tasks employed had a partially observable state-space which led to non-Markov dynamics. It is in these cases that eligibility traces should be most useful.

experiment are susceptible to slow versus rapid learning (Pearce and Hall, 1980). Since we have already explained why uncertainty about a stimulus' value should control the rate of learning about it, in the terms of this chapter, we would identify associability with uncertainty (Dayan and Long, 1998; Dayan *et al*., 2000; Courville *et al*., 2006). The key findings in this area are the many demonstrations that animals learn faster following surprising events (Pearce and Hall, 1980). These results go beyond the observation that learning is error-driven. Instead, if one experiences large (positive or negative) prediction errors on a particular trial, then one's rate of learning from prediction errors on *subsequent* trials should be elevated. One way to understand these effects (which experiments in rodents and humans trace to a network centering on the amygdala, especially its central nucleus; Holland, 1997; Roesch *et al*., 2010; Li *et al*., 2011) is that unexpected events may increase estimated volatility, which increases uncertainty and therefore also increases learning rates.

REWARDS AND PUNISHMENTS

In the remainder of this chapter, we consider the three formal objects of an MDP – states, actions, and rewards – and ask what they correspond to in the biological setting. As we will see, in each case, these constructs are not static but instead each raises an additional learning problem for the brain. We begin with rewards, which are characterized as scalar values describing the immediate utility of a state.

The Subjectivity of Reward

A central question is: What is the *reward function*? Although biologists and psychologists studying learning typically assume that certain outcomes – such as water for a liquid-deprived animal – are rewarding, in principle, RL theory considered alone has nothing to say about which states are rewarding. To the contrary, the theory applies equally to every possible definition of reward, or reward function, although different reward functions will typically predict different optimal policies. In describing the behavior of an organism using RL, we would also expect the reward function to be, at least to some extent, subjective: specific to that organism's preferences, and not directly accessible to the experimenter.

But if the theory is so general, does it actually have any content? For instance, if we hypothesize that dopaminergic responses carry reward prediction errors, does this claim actually mean anything without making additional assumptions about what constitutes a reward? Is there any behavior of a reward prediction error that is universal, regardless of the reward function? This problem is familiar to economists since the analogous question arises for expected utility theory. In that case, the solution was axiomatization (von Neumann and Morgenstern, 1947; see Chapter 1): the choices of an expected utility maximizer can be shown to satisfy a set of basic axioms, regardless of the utility function. There has been a similar program to recast parts of RL in axiomatic form, as discussed in Chapter 1 (Caplin and Dean, 2008). This strategy has been used to verify that BOLD responses in human striatum (though not yet dopamine neurons in primates) comply with the axioms (Rutledge *et al*., 2010).

The Construction of Reward

Another approach to the question "what is reward" is to ask where rewards come from: how the brain "constructs" them. For instance, a sweet taste is rewarding, but this is presumably in virtue of the fact that it predicts some subsequent biological event that is even more directly related to an organism's fitness, such as an increase in blood glucose levels. Meanwhile, sweet tastes (and blood glucose increases) are themselves predicted by more distal events such as the chimes of an ice cream truck. Assuming that the brain is born with only a minimal set of built-in, evolutionarily programmed rewards, such as changes in blood glucose, can it build on this foundation a richer notion of reward?

Indeed, this is a question to which we already have the answer: This is exactly what TD learning does. As discussed in the previous chapter, by learning a long-run future value function over states, TD learns to treat stimuli that are predictive of future events already specified as rewards in many ways equivalently to the ultimate rewards themselves. TD learns to assign high values to states that predict future reward; when encountered unexpectedly, these drive reward prediction error (and the assignment of value to still more distal states that predict them), just like primary rewards. As stressed in Chapter 15, this device explains how the brain comes to value secondary reinforcers such as money. Exactly the same principle would allow the brain to learn that sweet tastes are rewarding (assuming, for the sake of the example, this is not itself inbuilt) because they reliably predict subsequent, slow changes in blood glucose. Thus, fundamental to these theories is an account how the brain builds up a rich landscape of value given a minimal seed.

Punishment and Avoidance

Perhaps the largest set of open questions in this area concerns how to treat punishments in this framework. In principle, aversive events could just be assigned negative reward values and assessed on a common scale together with rewards. Indeed, traditional economic models begin with this assumption. Psychology suggests that aversive processing may be somewhat more complicated, however. Whereas the dopamine system clearly represents a common pathway for many different sorts of appetitive stimuli, it is less clear to what extent aversive stimuli (or costs) are also integrated into the dopaminergic signal. Early animal conditioning experiments (such as *counter-conditioning*, in a which stimulus is trained to predict both reward and punishment) led to the suggestion that appetitive and aversive predictions are actually maintained in separate, opponent channels rather than stored as a single net value summing over positive and negative (Konorski, 1967; Solomon and Corbit, 1974).

One neural constraint that may motivate this approach is that the firing rates of neurons are bounded below by zero, meaning that they map most naturally onto half of the real line; positive or negative numbers rather than both. This constraint is thought to give rise to opponent representations in other situations such as color vision. For RL, the low background firing rates of dopamine neurons suggest a limited dynamic range for reporting unexpected punishments if they are simply coded as negative rewards, since excursions below the baseline are rectified at zero firing rate (though see Bayer *et al.*, 2007). It has been suggested, albeit on quite indirect evidence, that parts of the ascending serotonin system might serve as an aversive opponent to dopamine, carrying the other half of the signal (Daw *et al.*, 2002).

A related question with a long history in animal conditioning is how responses for avoiding (or escaping) punishment are learned and motivated. Psychological theory suggests that the termination of punishment, or more importantly the cessation of the *expectation* of punishment, can be reinforcing (Maia, 2010; Mowrer, 1951; Moutoussis *et al.*, 2008), enabling avoidance learning. Cues predicting danger, but also successful avoidance, can also come to be associated with anticipated relief or safety (D'amato *et al.*, 1968) as well as danger. For instance, a cue may signal the aversive expectancy that an electric shock is imminent, but if that shock can be avoided (e.g., by a lever press), then the cue additionally signals the opportunity for avoidance, a relative improvement. Relief and its anticipation can only be relatively positive (relative to punishment and its anticipation) – the net value of an avoided punishment is clearly nil – but in the context of opponent systems, the negative aspects of fear and the opposing positive aspects of relief might be coded separately, activating both positive and negative channels.

Dopamine and Punishment

Given all this psychological and computational complexity, it is perhaps not surprising that there have been conflicting reports how dopamine neurons behave in response to punishments and stimuli predicting punishment. Although some dopaminergic units are inhibited by these events, or unresponsive (Matsumoto and Hikosaka, 2009; Mirenowicz and Schultz, 1996; Ungless *et al.*, 2004) – consistent with the expectation that they are coded as negative rewards, or separately – there have been reports of other putatively dopaminergic neurons that are *excited* by aversive stimuli as well as rewarding ones (Joshua *et al.*, 2008; Matsumoto and Hikosaka, 2009).

One recent report argued that there were two classes of putatively dopaminergic neurons: one showing the classic prediction error response, and the other excited both by signals of future punishment and reward (Matsumoto and Hikosaka, 2009). In general, a failure to differentiate good from bad outcomes seems hard to explain from the perspective of net decision variables, and for this reason such responses have tended to be understood in terms of arousal or attention rather than reinforcement (Horvitz, 2000). However, as mentioned, rather than the net over reward and punishment, dopamine might preferentially report the positively motivating aspects of anticipated relief from danger, due to avoidance

(Maia, 2010; Moutoussis *et al.*, 2008) in the context of a system where the accompanying aversive aspects are coded elsewhere. Regarding avoidance, many laboratory experiments in this area have used noxious airpuff stimuli to the face or eye, which typically cannot be avoided entirely but may be mitigated somewhat by blinking. In any case, the suggested heterogeneity of dopamine response types also cuts against the concept of the dopaminergic response as a unitary, scalar prediction error (see Chapter 15) and complicates the problem of interpreting the dopaminergic signal at the recipient structures.

Most confusingly, the interpretation of all these results depends importantly on the methods used to classify recorded neurons as dopaminergic. Typically, in extracellular recordings, this classification is based on characteristics of the extracellular responses such as the spike width and the background firing rate, features which are known (originally from more invasive intracellular recordings with verified histology) to be predictive of dopaminergic status. However, dopaminergic neurons are intermixed with other, nondopaminergic neurons, and at least in some dopaminergic regions these electrophysiological properties do not perfectly discriminate neuronal types (Margolis *et al.*, 2006; Ungless and Grace, 2012; Ungless *et al.*, 2004).

There are, however, more technically elaborate techniques that can be used to identify neurons that synthesize dopamine and these explicitly dopaminergic neurons can be examined in detail. When dopaminergic status is verified, false positive rates (neurons that would be wrongly characterized as dopaminergic using electrophysiological criteria) in the rat ventral tegmental area are typically higher than 10% and in one study nearly 40% (Ungless and Grace, 2012).

This brings us back to the question of punishment responsiveness. In one study that used *juxtacellular* and *immunofluorescent* labeling to verify conclusively the presence of dopamine in recorded neurons, all true dopaminergic neurons were found to be inhibited by punishment and all the neurons that were excited by punishment were found to be nondopaminergic neurons that would have been misclassified as dopaminergic if electrophysiological properties alone had been used, as is almost always the case in primate studies (Ungless *et al.*, 2004). Another study that tagged dopamine neurons optogenetically detected small responses to punishment in verified dopaminergic neurons only at a low rate similar to that expected due to chance (Cohen *et al.*, 2012). A third study reported verified dopamine neurons that were indeed excited by punishment, but these were relatively segregated in a constrained anatomical location (the ventral part of the ventral tegmental area) (Brischoux *et al.*, 2009). Confusingly, the punishment-responsive neurons from the primate study by Matsumoto and Hikosaka (2009) tended to be located toward the opposite end of the midbrain (dorsolateral substantia nigra pars compacta) from the punishment-responsive dopaminergic neurons identified in rodents by Brischoux and colleagues (2009), suggesting the importance of future direct study of these neurons using methods that can achieve positive identification of dopaminergic biochemistry.

STATES, STIMULI, AND PERCEPTUAL UNCERTAINTY

The most stylized and unrealistic aspect of the MDP model is the state. In an MDP, the world has a state s_t at each timestep, which determines its subsequent dynamics. In order to use standard RL approaches to solve the MDP, the agent needs to *know* that state: in the parlance of computer science, the state must be *fully observable*. Standard RL approaches rely on the Markov property that, conditional only on the current state and action, all future states and rewards are independent of everything that happened previously. This means that to solve the RL problem in an MDP, the agent need not “remember” any previous states: only the currently observed one matters. Coming at the same point from the other direction, if an agent is solving a task using standard RL algorithms, then whatever the agent uses as its state must contain all history about previous events that is relevant to predicting subsequent states and rewards.

What could this state correspond to in a biological organism? Clearly, in part, it includes the animal’s ongoing perceptual sensations, but as we will see these are typically insufficient to satisfy the Markov property by themselves: the real world is not an MDP, at least one in which the states correspond directly to percepts. Instead, we must identify the RL state with an internal representation comprising not only immediate perceptual sensations, suitably analyzed and processed, but also interoceptive ones (e.g., motivational states such as hunger that indicate which future outcomes will be rewarding) and working memories (of whatever previous stimuli are relevant to future rewards). It may seem that embedding working memories into the agent’s state is a trick allowing the agent to use the MDP approach even in a world that is not really an MDP. And of course that is exactly what it is.

The important point here is thus that several heterogeneous cognitive functions can be hidden behind the seemingly simple notation s_t. On the other hand, that may be a fairly realistic feature of this class of models. The dopamine neurons and their striatal afferents are interconnected with the frontal lobes and hence, by most obvious routes, many synapses away from the sense organs. That is if s_t is the RL system's input – the thing that it maps to value predictions and prediction errors – then it may well be reasonable to envision this input as a highly processed quantity, arriving by way of the brain's whole cortical machinery for perceptual analysis. (That said, there are also some important shortcuts, notably direct connections from precortical sensory areas of the colliculus to the midbrain dopamine neurons, which are thought to play a role in short-latency dopamine responses to sensory events; Dommett *et al.*, 2005.)

Below, we will consider approaches which separate the problem of constructing a state – perceptual analysis – from the reinforcement learning problem itself, with the former "module" providing the input to the latter. As we will see below, researchers in computer science have, in effect, argued that mathematical features of the problem license this separation of function. More practically, as we will also see, existing work on cortical mechanisms for perceptual inference fits the bill well in terms of complementing the RL system as we have already described it here and in the preceding chapter.

Before that, let us get a bit more concrete about what is wrong with the animal's immediate perceptual sensations, from the perspective of RL. First, they do not contain enough of the *right* information. Consider a trace conditioning experiment (Figure 16.2, from one of the earliest papers on the TD model of the dopamine response). Here, a transient visual stimulus signals that a reward will be delivered following a one-second pause. If the reward is omitted, dopamine neurons pause, signaling a negative prediction error, at the time the reward should have occurred. Note that nothing in the external world signals the time of reward delivery, only the passage of time following the transient stimulus. Thus even in this almost trivial example, the animal's immediate sensations (pause, flash, pause, reward) violate the Markov property, since the two pauses (if we think of them as "pause" states) are ambiguous – are they the pause before the flash or the pause before the reward? This is a significant distinction that a straightforward mapping from percepts to states simply fails to incorporate, despite our intuition that the difference between these two pauses is obvious to the subjects and the empirical observation that they are clearly different to the dopamine neurons (Daw *et al.*, 2006). The state sufficient to satisfy the Markov property for this task, and likewise the input that produces the dopaminergic response, must therefore maintain representations of stimulus history (the preceding flash), and also track the passage of time. This is a significant issue with which the earliest models struggled. To simulate this task, early theorists constructed a state representation by hand that fulfilled these desiderata (Montague *et al.*, 1996), but a deeper question is how the brain can build an appropriate representation for an arbitrary task without this kind of outside help.

It is also important to note that an organism's immediate perceptual sensations also have the opposite problem: they contain too much irrelevant information. Even in a simple laboratory situation, a universe of uncontrolled sensations impinge on the animal at each moment, such that in simply writing "pause, flash, pause, reward" we have filtered out dozens of other coincident sensations that happen to be irrelevant to

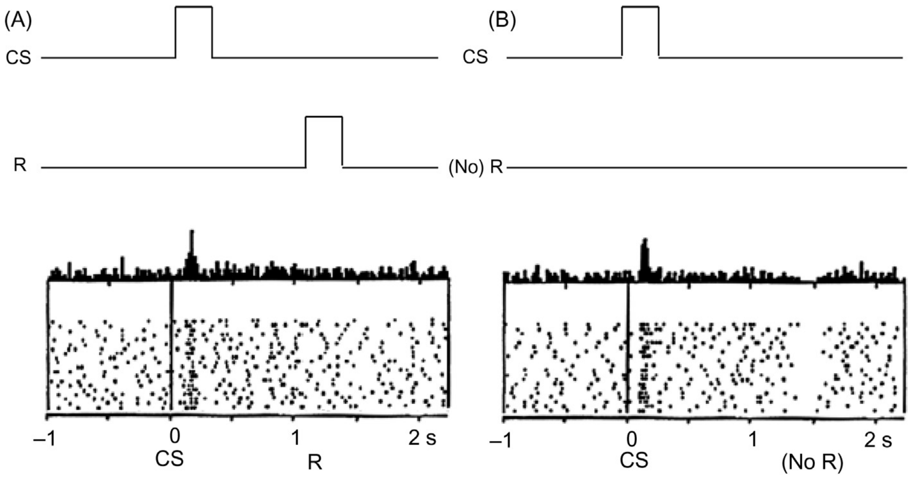

FIGURE 16.2 Events in trace conditioning and dopamine responses. *Adapted with permission from Schultz et al. (1997).* (A) If a transient visual stimulus (CS) signals reward (R), a typical dopamine neuron responds to the stimulus but not the reward. (B) If the reward is omitted, the neuron is briefly inhibited at the time the reward was expected. Note that nothing in the animal's immediate sensory environment predicts the reward timing in this case (only the passage of time following the transient stimulus). The stimuli in this task therefore violate the Markov property.

the task. The problem of learning is the problem of generalization from previous experiences to the current one. If s_t is defined so exhaustively that the animal never encounters the same state twice, then learning can't get off the ground.

A third problem with perceptual sensations is that they are often ambiguous. Many tasks in perception come down to drawing uncertain inferences from noisy measurements – was there a rustle? Where did that sound come from? Is that a tiger in the bushes? Indeed, even the simple example of trace conditioning implicates perceptual uncertainty, because the subjective perception of the passage of time is noisy. Thus, if a stimulus is followed in one second by reward, and the stimulus has been observed but reward has not been received, the subject will face uncertainty as to whether the full second has already elapsed or is still ongoing, and thus uncertainty whether the reward was omitted or is still to arrive (Daw *et al.*, 2006).

In short, to make TD work, the brain needs to construct a state representation that contains adequate history, omits irrelevant information, and somehow copes with perceptual uncertainty. And that is more of a challenge than might be immediately obvious.

Theory: Partial Observability and Perceptual Inference

A standard way of characterizing some of these sorts of problems in computer science is a formalism known as the partially observable MDP (POMDP), which augments the MDP with an explicit characterization of noisy perception (Kaelbling *et al.*, 1998). A POMDP is just an MDP, with the exception that the state s_t is *hidden* or not observable to the agent, who instead receives at each step some noisy observation o_t related to the state. The observation is determined by the hidden state, according to some distribution $P(o_t|s_t)$, but the mapping may be stochastic and the observation may not uniquely identify the state. Note that although the hidden states obey the Markov property by assumption, the observations need not (and generally will not) do so. A violation of the Markov property means that the current observation is insufficient to predict future states and rewards, and thus to solve a POMDP, unlike an MDP, it is necessary to take account of previous observations as well. The distinction between the states and the observations allows POMDPs to characterize problems like trace conditioning, in which the immediate stimuli are too sparse to obey the Markov property, as well as problems in which violations of the Markov property arise due to noisy perception.

As a specific example of such a problem (Kaelbling *et al.*, 1998), consider a task with two hidden states, which we'll call s_L and s_R (Figure 16.3). The world has two doors, behind one of which is a tiger and the other is a pot of gold. In s_L, the tiger is behind the left door, whereas in s_R the tiger is on the right. You have three actions available: a_L and a_R (which open either the left or right door, respectively, and end the game) and a_W, which does nothing for one timestep, after which you can choose again. The reward functions are such that you receive a large negative reward (e.g., -20) for opening the door with the tiger, and a positive reward (e.g., $+10$) for opening the door with the money. Finally, what makes this a POMDP is the observation function. You don't know whether the true state is s_L or s_R (if you did, the problem would be trivial), but each time you wait and listen (a_W), you hear one of two observations, a rustle behind one of the doors: o_L and o_R. These sounds are not particularly reliable, but 60% of the time, the rustle corresponds to the tiger's true position: $P(o_R|s_R) = 0.6$, $P(o_L|s_R) = 0.4$ and likewise for s_L.

Faced with such a problem, what should you do? Informally, in the tiger problem, you should wait and listen repeatedly so as to figure out, from the preponderance of evidence, which side contains the tiger, then choose the other side. How long should you wait? At first, listening improves your expected reward substantially, by giving you evidence about the tiger's location that makes it more likely that you will choose the rewarding door. But as you become more confident about the tiger's location, the marginal improvement from more listening declines. Eventually, the cost of another step's delay in harvesting the reward (due to time discounting, the parameter γ in Equation 16.3 and discussed in more detail in Chapter 10) outweighs the value (in terms of higher

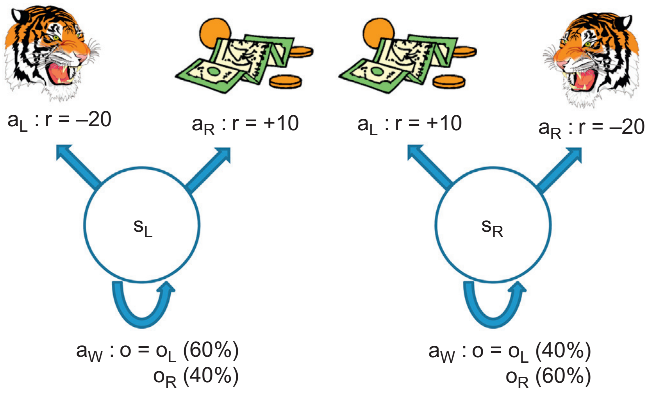

FIGURE 16.3 The tiger POMDP. The subject does not know if she is in state s_L, where the left door a_L is dangerous, or state s_R, where the right door a_R is dangerous. Only by waiting (a_W) and accumulating evidence about which state obtains is it safe to choose a door.

expected reward) of the additional information that would be gained. The precise point that strikes this balance depends on your delay discount preferences, γ, and the relative rewards for mistakes (-20) and successes ($+10$; see Kaelbling *et al*., 1998; Dayan and Daw, 2008, for the full analysis).

More generally however, POMDPs are devilishly hard to solve, and the efficient solution of large POMDPs is an ongoing area of research in computer science. However, the example above suggests that we can make some progress by first focusing on the perceptual subproblem of trying to figure out the hidden state. Inferring the state, in fact, is a straightforward exercise in Bayesian reasoning. In particular, assume that you face a POMDP and you know, or are able to learn, the state transition and observation functions, $P(o_t|s_t)$ and $P(s_{t+1}|s_t, a_t)$; a world model of the task. (For the tiger problem, the state transition function is trivial: if the state is s_L at the start, it stays there, and *vice versa* for s_R, since the tiger doesn't switch rooms.)

Now, conditional on any particular series of observations $o_{1\ldots t}$ and actions $a_{1\ldots t}$, you can infer a probability distribution over the state, called the *belief state* $b_t(s) = P(s_t|o_{1\ldots t}, a_{1\ldots t})$. Given the generative model, tracking b_t is straightforward: in particular from Bayes rule and the various Markov conditional independence properties, we have

$$b_{t+1}(s) = P(s_{t+1}|o_{1\ldots t+1}, a_{1\ldots t+1}) \propto P(o_{t+1}|s_{t+1}) \sum_{s_t} P(s_{t+1}|s_t, a_t) P(s_t|o_{1\ldots t}, a_{1\ldots t}) \tag{16.7}$$

where the first two terms are from the world model and the last is the previous timestep's belief state b_t.

To make this more explicit using the example of the tiger problem, define the initial belief $b_0(s_L)$ as .5 and likewise for s_R, reflecting total ignorance about the state. Then if we observe $o_1 = o_L$ on the first step, the posterior probability $b_1(s_\mathrm{L})$ is $(.6 \cdot b_0(s_L))/(.6 \cdot b_0(s_L) + .4 \cdot b_0(s_R)) = .6$.[1] That is, in this case we are 60% certain that the tiger is behind the left door, with the remaining 40% probability, $b_1(s_R) = 1 - b_1(s_L) = .4$, that the tiger is on the other side. Another such observation, $o_2 = o_L$, will update this value to $b_2(s_L) = (.6 \cdot b_1(s_L))/(.6 \cdot b_1(s_L) + .4 \cdot b_1(s_R))$ or about 69% left, after which an observation $o_3 = o_R$ would reduce the chance that the tiger is on the left to $b_3(s_L) = (.4 \cdot b_2(s_L))/(.4 \cdot b_2(s_L) + .6 \cdot b_2(s_R))$, or back to 60%.

Crucially, this update process at each step was accomplished through considering only the current observation, in light of the current belief state. That is, you can update the belief state recursively, by recomputing the distribution at each step in light of its current value and each new observation. Another way to see this is that in Equation 16.7, the first two terms depend only on the immediate events (o_{t+1} and a_t); the full history of actions and observations only enters through the last term, which is $b_t(s_t)$.

This last point is key to a solution to the one of the problems we started with, which is how to maintain the appropriate stimulus history to satisfy the Markov property. We can update the belief state recursively just by considering each new observation, exactly because at each step the belief state itself summarizes all history about previous actions and observations relevant to determining the future state. Formally, b_t is a sufficient statistic for these long lists of previous events, $o_{1\ldots t}$ and $a_{1\ldots t}$. Thus, other than the distribution b_t, we need not maintain any history information in order to infer subsequent states.

From Perception to Decision

So far, we have only considered the perceptual problem of tracking the state of the world, and not the RL problem of how to choose actions so as to maximize reward. But what we have accomplished above is highly relevant to the latter problem. We could not just apply TD to learn action values in the tiger problem because the immediate observations don't satisfy the Markov property. But we have now defined a quantity, the belief state, that tracks all relevant history. Accordingly, a well-known theorem (Kaelbling *et al*., 1998) states that the belief states themselves satisfy the Markov property, or more specifically that they form the states of an MDP, called the belief state MDP. That is, taking the belief state as input, TD works!

This theorem delivers on our earlier promise to formally license separating the perceptual inference problem – here, tracking the inferred state b_t – from the RL problem. In particular, we can feed b_t as the state input to any RL system for MDPs, such as a TD learner, and use it to learn values and policies as a function of this state (Daw *et al*., 2006; Dayan and Daw, 2008; Rao, 2010). Inference transforms the non-Markovian observations o_t into Markovian states b_t. Since b_t is a distribution reflecting uncertainty about the true state, this also allows us to behave optimally with respect to this uncertainty, choosing, for instance, whether or not to listen again in the tiger problem, depending on how confident we are that we know where the tiger is. The main practical problem here is that b_t (again, being a distribution) is continuous rather than discrete, but this can be dealt with, at least approximately, by learning the value

[1] Here the numerator is from Equation 16.7, which is simplified by the lack of state transitions in this problem, and the denominator is the constant of proportionality from Bayes' rule, which was omitted from Equation 16.7. Bayes' rule is discussed in Chapter 4.

function of b_t using methods for approximating continuous functions.

POMDPs and Neuroscience

The theory of POMDPs helps to clarify what we need out of a state representation for a TD system, and in particular casts the rather ill-defined problem of producing an appropriate state representation in the terms of a well-defined Bayesian inference problem. This is useful for a number of reasons; most importantly it reveals a rather clean correspondence between what the brain needs from the perspective of solving the RL problem, and what theoretical neuroscientists studying perception have already suggested that the brain's sensory systems provide. The idea that the job of the brain's sensory systems is essentially to reconstruct the hidden causes underlying noisy percepts is a longstanding one in neuroscience (often traced back to Helmholtz, 1860). The framework of Bayesian generative modeling and inference underlies prominent modern views of perception (Knill and Pouget, 2004; Yuille and Kersten, 2006), such as models that explain the receptive fields of V1 neurons as detecting latent features common in natural images (Lewicki and Olshausen, 1999; Olshausen, 1996).

An even more direct line exists between the POMDP problem as described here, and recent work on the brain's substrates for judgments about noisy perceptual stimuli (Gold and Shadlen, 2002; Roitman and Shadlen, 2002; Yang and Shadlen, 2007). In neuroscience, this research is considered to fundamentally be about decision making, though this work is largely disjoint from research on RL and other classes of more economic decision making, because the focus is primarily on perceptual uncertainty rather than optimizing utility. For the same reason, however, the two kinds of decision-making studies in neuroscience are quite complementary. The tiger problem described in the previous section is a standard example of a POMDP from computer science, but it is also isomorphic to a well-studied task in perceptual neuroscience, the *dots judgment task* of Newsome and colleagues, discussed in Chapters 8 and 19 (Newsome and Pare, 1988). The task requires judging whether a partially coherent motion stimulus is moving left or right. Here, the tiger task's states s_L and s_R correspond to the possible motion directions, and the observations represent instantaneous morsels of perceived motion energy. A line of research by Shadlen and colleagues characterizes the pre-saccadic responses of neurons in lateral intraparietal area LIP as accumulating evidence about motion direction, captured in their model (Gold and Shadlen, 2002) as a transformed version of the POMDP belief state, $\log(b_t(s_L)/b_t(s_R))$. There is a direct relationship between this belief state's evolution and the dynamics of a drift diffusion process of the kind often used to model this class of tasks, and discussed in detail in Chapter 3.

At least as characterized in these experiments, this system is thus a direct example of a neural representation of exactly the sort of belief state we have argued is necessary for TD to solve the policy optimization part of this task, i.e., learning under what circumstances to respond "left" or "right". (That said, this specific neural population has additional properties that make it an imperfect candidate for a pure representation of belief state, notably that its neurons are also modulated by saccade utility; Platt and Glimcher, 1999.) In any case, the composition of belief tracking and TD models predicts some non-trivial behaviors of the downstream reward prediction error signal as a result of upstream perceptual uncertainty, such as slowly unfolding positive or negative errors reflecting the inference that a particular motion stimulus is easier or harder than expected (Rao, 2010). Such responses have indeed been observed in primate dopamine neurons recorded during the dots judgment task (Nomoto *et al.*, 2010). Using a very different task in rodents, the lesion of orbitofrontal cortex produced a pattern of changes in the responses of dopamine neurons, which modeling suggested was consistent with the elimination of internally generated (i.e., not externally stimulus-bound) aspects of the TD system's state representation (Takahashi *et al.*, 2011). This result led the authors to suggest that the OFC, which is upstream from ventral striatum and dopamine neurons of the ventral tegmental area, might be contributing to the state representation.

Behaviorally, learning and choice behavior in another RL task with a hidden state is well explained as resulting from such a hybrid scheme of combining Bayesian inference of the latent state with TD for learning action values over the inferred state (Gershman *et al.*, 2010b). This task (see also Wilson and Niv, 2011) requires choice between options that are identified by a number of stimulus dimensions (color, shape, etc.). At any particular time, only one dimension is diagnostic as to which option is rewarding, and the other features are distractors. This task thus captures another of the state representation problems we began this section with: the profusion of irrelevant stimuli. This problem of dimensional selective attention admits exactly the same sort of solution, in terms of Bayesian inference about the hidden state, as the other problems we have considered, because in this task the true generative model contains only one relevant stimulus and inference over the hidden state therefore serves to highlight it and suppress the others.

A final point is that, as described in the previous section, inferring the latent state in the task requires a learned generative model of the way the latent states give rise to the observations. (In a POMDP, this includes the observation function $P(o_t|s_t)$ and also the state transition function $P(s_{t+1}|s_t,a_t)$.) It is clear that these functions must also be learned. The neural substrates for this sort of learning about the structure of the latent causes generating experience are an almost completely open problem, but the computational principles underlying such learning are well understood (it is yet another application of Bayesian inference), and this sort of learning has been argued to account for behavior across a number of different settings ranging from animal conditioning to human causal reasoning (Courville *et al*., 2003, 2006; Gershman and Niv, 2010; Gershman *et al*., 2010a; Tenenbaum and Griffiths, 2001). In a real sense, then, the state in RL models is itself a learning problem, much as we have seen for rewards and we will next see for the actions.

ACTIONS

The flip side of the state question is the action question. RL models typically refer to a discrete set of actions $\mathscr{A}$, but of course the movements of biological organisms vary continuously and have complex structure.

Vigor, Opportunity Costs, and Tonic Dopamine

One feature of the actions of biological organisms that is not captured in the traditional MDP framework is *vigor*. Whether it is moving from place to place or pressing a lever, animals produce their actions with varying speed and effort. Indeed, in a great deal of experimental psychology it is the not the choice of action *per se* but the speed by which it is accomplished – for example the rate of pressing on a single lever –that is the primary variable of interest.

That experimental psychologists are interested in something other than decisions might seem of peripheral relevance to students of decisions, except for a puzzling empirical fact: response vigor is also closely tied in with the neural systems we have attributed thus far to RL, and notably with the dopamine system (Lyon and Robbins, 1975; Robbins and Everitt, 2007). If anything, the causal link to vigor is clearer than that to reinforcement learning. For instance, modulations in action vigor are by far the most obvious consequences of manipulations or disorders of dopamine function, easily visible to the naked eye. If you simply inject a rat with a dopamine agonist such as amphetamine, it will become more active, run around, do everything faster. Similarly, human patients with Parkinson's disease (a neurodegenerative disorder affecting dopaminergic neurons) do not typically visit the doctor because they are having problems with reward learning. Their primary symptoms, especially in the early stages of the disease, are difficulties with movement: difficulty initiating movement and a pronounced slowing, called bradykinesia, of those movements that are carried out.

Building on causal evidence like this (rather than the neuronal recording studies of dopamine neurons that launched the RL work), there is a tradition in neuropsychology and behavioral pharmacology of interpreting the function of dopamine in terms of motivation and the modulation of action vigor (Berridge, 2007; Salamone *et al*., 2007). Indeed, some researchers in this area have argued that, contrary to the claims of TD models, dopamine does not drive learning at all, it only modulates action directly (Berridge, 2007). Against this position is evidence (also presented in Chapter 18) that dopamine does affect plasticity at both neural and behavioral levels, in the way that would be anticipated by the RL theories (Frank *et al*., 2004; Reynolds and Wickens, 2002; Rutledge *et al*., 2009; Tsai *et al*., 2009; Wang *et al*., 2011). However, it also seems implausible that these learning effects entirely mediate the substantial and very rapidly appearing vigor-related phenomena described above. Indeed, using a design that compares within- to between-trial effects to isolate changes mediated by learning, Gallistel and colleagues. (1974) argued that the electrical stimulation of dopamine fibers has both direct and learning-mediated effects.

All this raises a question: why does the same neurochemical appear to subserve both reinforcement learning and the control of movement vigor? These two functions appear, at first impression, to be basically orthogonal to one another.

One answer to this question is suggested by modeling work from Niv and colleagues (2006, 2007), who consider a formal account of the problem of vigor control. In their model, which generalizes the MDP framework (technically, it belongs to a class known as *semi-Markov decision problems*), animals must choose not just which discrete action, a, among a set $\mathscr{A}$ to take, but also a continuous vigor, expressed as a latency to completion, τ, with which to effect it. Optimal choice, jointly, over a and τ can be studied in different tasks. The model assumes that the more rapidly an action is taken, the more energetic cost it incurs. The key feature of the problem is then that the choice of vigor τ that optimizes long run expected reward depends on a tradeoff between this energetic cost of behaving more rapidly, and a second sort of cost, the *opportunity cost* of behaving more slowly. The latter arises because while one is, for example, spending thirty seconds walking from one end of the room to another, one is not otherwise earning reward. In fact, during that time one is, in a real sense, delaying all subsequent rewards that might be earned after crossing the room.

In the analysis done by Niv and colleagues (2006, 2007), this opportunity cost takes a simple form: $-\tau \cdot \bar{r}$, where $\bar{r}$ is the average reward per timestep. That is if one takes a monetary example; if on average you could be earning \$10 per minute, then every minute spent walking across the room is \$10 foregone. If you could be earning \$100 per minute, your sloth is all the more costly, and you face increased incentive to walk faster so as to return more rapidly to earning money. Across many tasks, this effect is reflected directly in the optimal choice of τ: you should behave more rapidly in environments where the average reward $\bar{r}$ is higher, since in this case the opportunity costs of inaction weigh more strongly against the energetic costs of vigorous action. This analysis has an interesting counterpart in classic work in optimal foraging theory, where the optimal time to leave a diminishing patch of food also depends on $\bar{r}$ due to a similar opportunity cost argument (Charnov, 1976; Stephens and Krebs, 1987).

The foregoing modeling provides a rational analysis of response vigor, and through the construct of the opportunity cost, grounds it in the reward rate. Experiments have indeed shown that manipulations of reward rate affect response vigor (Guitart-Masip *et al.*, 2011; Haith *et al.*, 2012). The role of the reward rate also suggests an answer to why dopamine is involved in the vigor problem as well as the RL problem, as in the case of bradykinesia in Parkinson's disease (Mazzoni *et al.*, 2007). To understand why, consider the TD error, $\delta_t = r_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$. (Here, compared to Equation 16.4, we have set $\gamma = 1$ and used the SARSA form of the rule, the version that omits the 'max"operation.) This is a standard model, from the RL perspective, of the phasic responses of dopamine neurons. Note that if you sum (or, equivalently, average) the prediction error δ_t over a range of timesteps $t = [1, 2, 3, \ldots]$, the terms involving the predictions Q cancel each other pairwise, since the same prediction is added at one timestep and subtracted at the next. Whatever the predictions, then, in TD learning the sum of prediction errors over time is essentially just the sum of rewards.

The upshot of all this is that implicit in the TD error signal is the average reward $\bar{r}$: it is just the prediction error response viewed in the aggregate over a longer timescale. This observation led Niv and colleagues (2006, 2007) to speculate that whereas phasic dopamine responses (the high frequency bursts and abrupt pauses in action potential rate) have been argued to carry δ_t, the same dopamine signal viewed at a tonic (lower frequency) timescale might report the average reward $\bar{r}$. They further suggested, due to the relationship between opportunity cost and optimal vigor, that this tonic dopamine signal should be expected causally and directly to modulate response vigor, which would tie in and explain the behavioral pharmacology results with which this section began.

The most straightforward mechanistic proposal, given the above considerations, may be to associate the average reward with *extrasynaptic* (see Chapter 5) dopamine concentrations in the striatum. Like other neurotransmitters, dopamine is released into synapses, but dopamine is substantially present in the extracellular fluid as well. This is because rather than releasing their transmitter only at terminal synapses located at the ends of their axons, dopamine neurons contain large numbers of release sites (called *en passant* synapses) throughout their extensive axonal trees. These allow each neuron to release dopamine throughout a large volume of the recipient brain, where it escapes the local synaptic cleft and diffuses through the extracellular fluid (Arbuthnott and Wickens, 2007). This effect is particularly prominent following phasic dopaminergic responses (as for prediction errors) in experiments using voltammetric methods (see Chapter 6), which observe transient increases in extrasynaptic dopamine concentration (Garris *et al.*, 2002; Phillips *et al.*, 2003). The resulting extrasynaptic dopamine can affect striatal medium spiny neurons via high-affinity dopamine receptors that are located away from the synapses. Altogether, it is tempting to hypothesize that these transients, temporally filtered by diffusion and reuptake, give rise, in effect, to an averaging operation by which the baseline extracellular concentration tracks the average reward $\bar{r}$. Of course, reality may be more complicated; for instance, there is evidence that extrasynaptic levels of dopamine in striatum (at least as tracked by microdialysis, a technique for direct chemical sampling at a very slow temporal resolution of one sample per 10 minutes) are regulated independently from phasic firing (Floresco *et al.*, 2003).

Leaving aside the exact mechanism by which the average reward is computed, the Niv *et al.* (2006, 2007) model brings response vigor under the purview of rational RL theories, and poses one possible answer to the question of why the functions of vigor and RL are naturally, almost necessarily, interconnected.

Action Hierarchies, Action Sequences, and Hierachical RL

Another problem with mapping the actions $\mathscr{A}$ in an RL model to the real world is that it's not clear what level of granularity they are supposed to describe. In laboratory experiments an animal might face a choice between looking at a red or a green target, which (appear to) map straightforwardly onto the simple theories discussed above. However, consider the problem of driving a car, which is something most of us have learned to do.

How would we describe driving as an MDP (Sutton, 1995)? At one level, the set of actions might refer to maneuvers like changing lanes, driving straight, or turning left or right. But to execute each of these maneuvers, one needs to carry out different sequences of lower-level actions, operating the pedals and the steering wheel, and in turn to do each of these things by moving muscles. Conversely, at a higher level of analysis one might describe actions like following Interstate 76 to Exit 33 and then following Highway 376 west to the end. These navigational actions each require extended sequences of turning, tracking lanes and so on. In principle, one could describe a navigational problem as an MDP with actions at any of these levels of granularity – steering wheels, right turns, or highway exits. As this example shows, action in realistic environments seems to have a natural hierarchical structure: more abstract actions are composed from more elemental actions. Such abstraction can clearly simplify the choice and learning problems: one cannot (plausibly) plan a road trip from New York to San Francisco in terms of the sequence of operations on the vehicle controls. At the same time, the lower level is essential: one cannot execute such a trip without ultimately operating the steering wheel and pedals.

What action space do the brain's putative RL systems learn over, then? The implausibility of planning road trips in the space of vehicular controls means that if they work only in the "natural" space of movements of the body, we would (literally) never get anywhere. Moreover, analogous interpretational problems occur even in a seemingly simple laboratory problem like choosing between two colored targets, since at a lower level even that abstract choice needs to be expressed through a series of movements with some effector, such as a saccadic eye movement or a button press.

In computer science, these issues have been studied in an area known as *hierarchical RL* (Barto and Mahadevan, 2003; Parr and Russell, 1998; Sutton *et al.*, 1999), and more recently these ideas have been brought to neuroscience by Botvinick and colleagues (2009). Using one popular approach, the options framework, much of the RL machinery we have discussed can be applied over a hierarchical action space, in which higher level actions (known as options) stand in for extended sequences of lower level actions, and this decomposition may be extended hierarchically (Sutton *et al.*, 1999). Thus, at an intermediate level of description, one might choose to make a left turn (an option); at a lower level, a controller for the left turn option has the responsibility to execute the maneuver, by making a series of choices about how to turn the wheel and press the pedals to pursue this goal. Decomposing the action space in this way also decomposes the RL problem into a hierarchy of simpler ones: at the higher levels, the existence of the left turn option simplifies the learning problem by enabling the system to neglect the microstructure of the maneuver, whereas the left turn controller can also, separately, be trained by its own, local, RL process targeted to its own goals.

Hierarchical RL then, seems like a plausible framework for solving the action problem in models of RL in the brain, explaining how the action space can be built up from simple bodily movements to bridge them to the more abstract sorts of relevant to laboratory experiments or real-world decisions (Botvinick *et al.*, 2009). As for how hierarchical RL works, there is really one relevant fact, which is that it still works with TD learning.

To be more precise, in hierarchical RL there are multiple learning problems at different levels, and all of them can be solved by TD learning (or indeed, alternatively by model based RL, or any mixture of these across levels). At the highest level, one still needs to choose what to do so as to maximize reward, and apart from some accounting for options that take time to complete, this works much as before. But making choices over abstract options sloughs off another part of the learning problem, which is that the option controllers (turning left or right) also need to be able to learn to execute their respective maneuvers as efficiently as possible. The key point is that these subsidiary learning problems (known as *intra-option learning*: learning how to complete an option) can each themselves be framed as RL problems and each also solved with TD methods. The main difference is that the option controllers should learn as though they are attempting to optimize a different reward, the *pseudoreward* that they attain by completing their respective *subgoals*, such as completing the turn. Thus they have different different (pseudo) reward prediction errors, giving rise to different value functions, but via the same learning rule (Figure 16.4).

All this suggests that the same sorts of computational and neural mechanisms already described can achieve learning at multiple levels of an action hierarchy, and together produce a more realistic account of what "actions" are (Botvinick *et al.*, 2009). The basal ganglia mechanisms associated with RL seem well suited to subserve such a hierarchy, since the same patterns of connections with cortex and with dopamine are repeated across many different (and apparently reasonably segregated) "loops" through the basal ganglia, connected topographically to different areas of frontal cortex (Alexander and Crutcher, 1990). It has

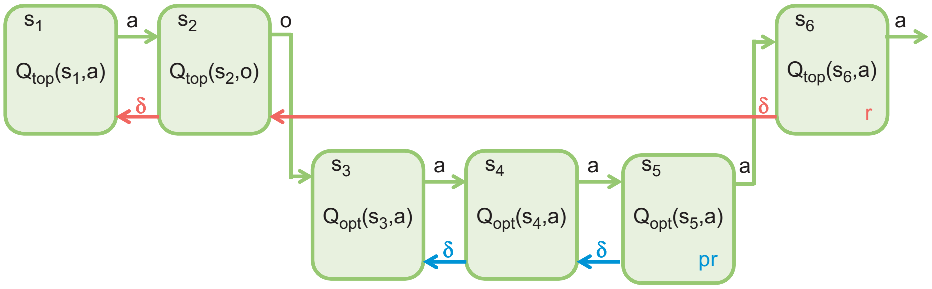

FIGURE 16.4 Prediction errors in hierarchical RL, after Botvinick *et al.* (2009). At the top level, the system may choose either a normal action *a* or an extended option *o*. The latter choice transfers control to the option, which chooses more actions. Each level is learning its own action value estimates, Qtop or Qopt, predicting reward *r* or pseudoreward *pr*, respectively. These learning problems each involve separate reward (or pseudoreward) prediction errors.

been suggested that a hierarchical controller could be laid out with actions at different levels of abstraction represented at different positions along this topography (Bornstein and Daw, 2011; Dezfouli and Balleine, 2012; Frank and Badre, 2012; though see Chapter 21 for a discussion of evidence supporting a different organizational scheme for these loops). One challenging aspect of such a hypothesis is that different levels of the action hierarchy require training from different (reward versus pseudoreward) prediction error signals. If these are all to be carried by dopamine, different dopaminergic responses would have to be heterogeneous and segregated between the levels. This cuts somewhat against the impression (see Chapter 15) of dopamine as a strikingly diffuse and homogeneous broadcast signal. Anatomically, the connections between dopamine neurons and striatum, though diffuse, have at least some topographical organization that might in principle support a coarse hierarchy (Haber *et al.*, 2000), but dopamine neurons have not yet been examined in tasks with hierarchical structure to see if any heterogeneity in their responses emerges.

In any case, more broadly consistent with the hierarchical RL idea, in an RL task involving hierarchical action, BOLD responses in the human striatum covary with the size of positive prediction errors related to changes in expectation about subgoal attainment: that is, with the pseudoreward prediction errors hypothesized to be used to train the option controller (Ribas-Fernandes *et al.*, 2011). (These results, and similar results using EEG, were obtained at times in the task when the top-level prediction error for primary reward was zero.) There is also fMRI evidence suggesting that neural topography respects a related notion of action hierarchy, whereby response rules with different degrees of nested contingent structure recruit distinct frontal and striatal areas along a posterior–anterior axis (Badre *et al.*, 2010; Badre and Frank, 2012). Finally, a line of unit recording studies in rodent dorsolateral striatum has shown that neurons there (but not, for instance, in adjacent dorsomedial striatum) have a sort of "action chunking" property (Jog *et al.*, 1999; Thorn *et al.*, 2010). As an animal learns a task involving a series of actions to navigate a T-maze, different neurons in the population initially respond to different events along the route, but with more training, the neurons come to respond only to the beginning and end of the trial. This may reflect the coding of the entire action chain as a unit, like an option.

Both empirically and theoretically, a major question in this area – the so-called *option discovery* problem – is how the hierarchies get set up in the first place; how a useful set of subgoals can itself be learned from experience. (This is different from the intra-option learning problem, which can be addressed using TD: how best to attain a given subgoal, such as changing lanes.) There is relatively little guidance on this problem from the computational literature, since this is substantially an open problem in computer science as well.

Multi-Effector Action

A related problem of hierarchy in action, also relevant to biology, is the problem of multi-effector action. The body has many effectors – two hands, two legs, and so on. If we treat the elemental actions $\mathscr{A}$ of an RL system as combinations of movements of each of the effectors, then the high dimensionality of the effector space leads to an exponential explosion in the set of candidate actions. This is a classic "curse of dimensionality" and would complicate choice and learning. Another area of hierarchical RL considers learning over high-dimensional action spaces of this sort, as with choosing actions for each member of a team of soccer-playing robots or a fleet of fishing boats (Chang *et al.*, 2003; Rothkopf and Ballard, 2010; Russell and Zimdars, 2003). The main strategy here is to divide and conquer, decomposing the intractable high dimensional problem into a number of smaller ones involving only subsets of effectors considered in isolation. This strategy seems well suited to many natural actions – think of talking on the phone while you walk – basically independent actions involving non-overlapping sets of effectors.

Perhaps more importantly, because the brain's movement systems are organized topographically, they are also well situated to treat choice over different effectors' movements separately from one another. Indeed, this view is implicit in the interpretation of, for instance, activity of neurons in lateral intraparietal cortex as a *value map* over saccade targets and the medial intraparietal cortex as a value map over arm movements (Platt and Glimcher, 1999; see Chapter 20). Such a representation reduces the multi-effector learning problem to multiple, simpler RL problems in each effector separately. If they were learning simultaneously, they would also require multiple prediction errors. Accordingly, in an fMRI study, distinct prediction errors were seen in left and right striatum for movements of the contralateral hand, in a task where these had separable values (Gershman *et al*., 2009; see also Palminteri *et al*., 2009; Wunderlich *et al*., 2009). Finally, of course, if the brain is organized to separate choice between effectors by default, this leads to a new problem of coordination: how to solve tasks like playing the piano or touch typing that require the conjoint action of multiple effectors. In this case the value of (for example) a particular hand movement may depend on what the other hand does, and in such a task these values cannot be represented in separate value maps for each hand. Although there is a classic literature on coordination in movements, especially across the hemispheres (Brinkman, 1984; Laplane *et al*., 1977; Tanji *et al*., 1988), from a neuroeconomic perspective, the valuation and decision problems over coordinated multi-effector actions are so far largely unstudied.

CONCLUSION

Like all models in science, the TD theory of the dopamine response is stylized and simplified. This chapter has examined a number of these simplifications and argued that in each of these cases, the theory's core mechanism for error driven learning can be used to address these additional problems. Thus, for instance, error-driven learning is still called for when the update rule is derived from principles of statistical measurement; the belief states in a POMDP form an MDP that can be solved with TD learning; and learning at multiple levels of an action hierarchy can all take place according to a common TD mechanism.

A related goal of this chapter was to lay the computational foundations for dealing with problems like states and actions in the context of RL theories, in part by examining how these problems have been treated in computer science. From a neuroeconomic perspective, one of the most important aspects of the TD theory is that it connects the hypothesized neural mechanism to precise normative considerations about learned optimal choice, based on the same decision theoretic principles that underlie economic analyses of these problems. The additional computational frameworks presented here – for example POMDPs and hierarchical RL – are not, as yet, so deeply developed in terms of their biological underpinnings. However, they provide a promising theoretical foundation for investigating these issues neuroscientifically, particularly because they dovetail so closely with the relatively well studied TD machinery. In this respect, finally, it is worth noting that due to the rapid turnaround of human neuroimaging experiments, many of the newer ideas discussed here have been investigated using fMRI in humans but not, as yet, more invasive techniques in animals. The time is ripe, guided by the human studies, to examine a number of these issues in animals. Particularly interesting, for example, are theoretical suggestions of vector-valued teaching signals: for example, separate reward prediction error signals for goals and subgoals in hierarchical RL. It remains to be seen whether such a conceptualization might help to uncover or explain any subtle regional variation in dopaminergic signaling properties.

References

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

Arbuthnott, G.W., Wickens, J., 2007. Space, time and dopamine. Trends Neurosci. 30, 62–69.

Badre, D., Frank, M.J., 2012. Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. Cereb. Cortex. 22, 527–536.

Badre, D., Kayser, A.S., D'Esposito, M., 2010. Frontal cortex and the discovery of abstract action rules. Neuron. 66, 315–326.

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

Barto, A.G., Mahadevan, S., 2003. Recent advances in hierarchical reinforcement learning. Discrete Event Dyn. Syst. 13, 341–379.

Barto, A.G., Sutton, R.S., Brouwer, P.S., 1981. Associative search network – a reinforcement learning associative memory. Biol. Cybern. 40 (3), 201–211.

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

Behrens, T., Woolrich, M., Walton, M., Rushworth, M., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.

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

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

Bertsekas, D.P., Tsitsiklis, J.N., 1996. Neuro-Dynamic Programming. Athena Scientific, Belmont, Mass.

Bogacz, R., McClure, S.M., Li, J., Cohen, J.D., Montague, P.R., 2007. Short-term memory traces for action bias in human reinforcement learning. Brain Res. 1153, 111–121.

Bornstein, A.M., Daw, N.D., 2011. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. Curr. Opin. Neurobiol. 21, 374–380.

Botvinick, M.M., Niv, Y., Barto, A.C., 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition. 113, 262–280.

Brinkman, C., 1984. Supplementary motor area of the monkey's cerebral cortex: short-and long-term deficits after unilateral ablation and the effects of subsequent callosal section. J. Neurosci. 4, 918–929.

Brischoux, F., Chakraborty, S., Brierley, D.I., Ungless, M.A., 2009. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc. Natl. Acad. Sci. U.S.A. 106, 4894–4899.

Caplin, A., Dean, M., 2008. Dopamine, reward prediction error, and economics. Q. J. Econ. 123, 663–701.

Chang, Y.H., Ho, T., Kaelbling, L.P., 2003. All learning is local: Multi-agent learning in global reward games. Adv. Neural Inf. Process. Syst. 16, 807–814.

Charnov, E.L., 1976. Optimal foraging, the marginal value theorem. Theor. Popul. Biol. 9, 129–136.

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

Courville, A.C., Daw, N.D., Gordon, G.J., Touretzky, D.S., 2003. Model uncertainty in classical conditioning. Adv. Neural Inf. Process. Syst. 16, 977–984.

Courville, A.C., Daw, N.D., Touretzky, D.S., 2006. Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. 10, 294–300.

D'amato, M., Fazzaro, J., Etkin, M., 1968. Anticipatory responding and avoidance discrimination as factors in avoidance conditioning. J. Exp. Psychol. 77, 41.

Daw, N.D., Courville, A.C., Touretzky, D.S., 2006. Representation and timing in theories of the dopamine system. Neural Comput. 18, 1637–1677.

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

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–1711.

Dayan, P., Daw, N.D., 2008. Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. 8, 429–453.

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

Dayan, P., Long, T., 1998. Statistical models of conditioning. Adv. Neural Inf. Process. Syst.117–123.

Dearden R., Friedman N., Russell S., 1998. Bayesian Q-learning. In: John Wiley & Sons Ltd, pp. 761–768.

Dezfouli, A., Balleine, B.W., 2012. Habits, action sequences and reinforcement learning. Eur. J. Neurosci. 35, 1036–1051.

Dommett, E., Coizet, V., Blaha, C.D., Martindale, J., Lefebvre, V., Walton, N., et al., 2005. How visual stimuli activate dopaminergic neurons at short latency. Sci. Signal. 307, 1476.

Floresco, S.B., West, A.R., Ash, B., Moore, H., Grace, A.A., 2003. Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nat. Neurosci. 6, 968–973.

Frank, M.J., Badre, D., 2012. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cereb. Cortex. 22, 509–526.

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

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

Garris, P.A., Christensen, J.R.C., Rebec, G.V., Wightman, R.M., 2002. Real-time measurement of electrically evoked extracellular dopamine in the striatum of freely moving rats. J. Neurochem. 68, 152–161.

Gershman, S., Pesaran, B., Daw, N., 2009. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J. Neurosci. 29, 13524–13531.

Gershman, S.J., Blei, D.M., Niv, Y., 2010a. Context, learning, and extinction. Psychol. Rev. 117, 197.

Gershman S.J., Cohen J.D., Niv, Y., 2010b. Learning to selectively attend. Proceedings of the 32nd Annual Conference of the Cognitive Science Society, pp. 1270–1275.

Gershman, S.J., Niv, Y., 2010. Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. 20, 251.

Gold, J., Shadlen, M., 2002. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron. 36, 299–308.

Guitart-Masip, M., Beierholm, U.R., Dolan, R., Duzel, E., Dayan, P., 2011. Vigor in the face of fluctuating rates of reward: an experimental examination. J. Cogn. Neurosci. 23, 3933–3938.

Haber, S.N., Fudge, J.L., McFarland, N.R., 2000. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J. Neurosci. 20, 2369–2382.

Haith, A.M., Reppert, T.R., Shadmehr, R., 2012. Evidence for hyperbolic temporal discounting of reward in control of movements. J. Neurosci. 32, 11727–11736.

Helmholtz, H., 1860. Handbuch der Physiologischen Optik. Leopold Voss, Leipzig.

Holland, P.C., 1997. Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. Learn. Behav. 25, 373–399.

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

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

Izhikevich, E.M., 2007. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb. Cortex. 17 (10), 2443–2452.

Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V., Graybiel, A.M., 1999. Building neural representations of habits. Science. 286, 1745–1749.

Joshua, M., Adler, A., Mitelman, R., Vaadia, E., Bergman, H., 2008. Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials. J. Neurosci. 28, 11673–11684.

Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and acting in partially observable stochastic domains. Artif. Intell. 101, 99–134.

Kakade, S., Dayan, P., 2002. Acquisition and extinction in autoshaping. Psychol. Rev; Psychol. Rev. 109, 533.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82, 35–45.

Killeen, P.R., 2011. Models of trace decay, eligibility for reinforcement, and delay of reinforcement gradients, from exponential to hyperboloid. Behav. Process. 87 (1), 57–63.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27, 712–719.

Konorski, J., 1967. Integrative activity of the brain. Leopold Voss, Leipzig.

Laplane, D., Talairach, J., Meininger, V., Bancaud, J., Orgogozo, J., 1977. Clinical consequences of corticectomies involving the supplementary motor area in man. J. Neurol. Sci. 34, 301–314.

Lewicki, M.S., Olshausen, B.A., 1999. Probabilistic framework for the adaptation and comparison of image codes. JOSA A. 16, 1587–1601.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D., 2011. Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. 14, 1250–1252.

Loch J., Singh S., 1998. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In: Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA.

Lyon, M., Robbins, T., 1975. The action of central nervous system stimulant drugs: a general theory concerning amphetamine effects. Curr. Dev. Psychopharmacol. 2, 79–163.

Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A., 2006. Bayesian inference with probabilistic population codes. Nat. Neurosci. 9, 1432–1438.

Maia, T., 2010. Two-factor theory, the actor-critic model, and conditioned avoidance. Learn Behav. 38, 50–67.

Margolis, E.B., Lock, H., Hjelmstad, G.O., Fields, H.L., 2006. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? J. Physiol. 577, 907–924.

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

Mazzoni, P., Hristova, A., Krakauer, J.W., 2007. Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation. J. Neurosci. 27, 7105–7116.

McKinstry, J.L., Edelman, G.M., Krichmar, J.L., 2006. A cerebellar model for predictive motor control tested in a brain-based device. Proc. Natl. Acad. Sci. U.S.A. 103 (9), 3387–3392.

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

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

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

Moutoussis, M., Bentall, R.P., Williams, J., Dayan, P., 2008. A temporal difference account of avoidance learning. Network: Comput. Neural Syst. 19, 137–160.

Mowrer, O.H., 1951. Two-factor learning theory: summary and comment. Psychol. Rev. 58, 350.

Newsome, W.T., Pare, E.B., 1988. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J. Neurosci. 8, 2201–2211.

Niv, Y., Daw, N., Dayan, P., 2006. How fast to work: response vigor, motivation and tonic dopamine. Adv. Neural Inf. Process. Syst. 18, 1019.

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

Nomoto, K., Schultz, W., Watanabe, T., Sakagami, M., 2010. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci. 30, 10692–10702.

Olshausen, B.A., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 381, 607–609.

Palminteri, S., Boraud, T., Lafargue, G., Dubois, B., Pessiglione, M., 2009. Brain hemispheres selectively track the expected value of contralateral options. J. Neurosci. 29, 13465–13472.

Pan, W.-X., Schmidt, R., Wickens, J.R., Hyland, B.I., 2005. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci. 25 (26), 6235–6242.

Parr, R., Russell, S., 1998. Reinforcement learning with hierarchies of machines. Adv. Neural Inf. Process. Syst.1043–1049.

Pearce, J.M., Hall, G., 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. 87, 532.

Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M., Carelli, R.M., 2003. Subsecond dopamine release promotes cocaine seeking. Nature. 422, 614–618.

Platt, M.L., Glimcher, P.W., 1999. Neural correlates of decision variables in parietal cortex. Nature. 400, 233–238.

Preuschoff, K., Bossaerts, P., 2007. Adding prediction risk to the theory of reward learning. Ann. N. Y. Acad. Sci. 1104, 135–146.

Puterman, M.L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, New York.

Rao, R.P.N., 2010. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci. 4, 146.

Rescorla, R., Wagner, A., 1972. Variations in the Effectiveness of Reinforcement and Nonreinforcement. Classical Conditioning II: Current Research and Theory, Appleton-Century-Crofts, New York.

Reynolds, J.N., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15, 507–521.

Ribas-Fernandes, J.J.F., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y., et al., 2011. A neural signature of hierarchical reinforcement learning. Neuron. 71, 370–379.

Robbins, T.W., Everitt, B.J., 2007. A role for mesencephalic dopamine in activation: commentary on Berridge (2006). Psychopharmacology (Berl). 191, 433–437.

Roesch, M.R., Calu, D.J., Esber, G.R., Schoenbaum, G., 2010. Neural correlates of variations in event processing during learning in basolateral amygdala. J. Neurosci. 30, 2464–2471.

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

Roitman, J.D., Shadlen, M.N., 2002. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. 22, 9475–9489.

Rothkopf, C.A., Ballard, D.H., 2010. Credit assignment in multiple goal embodied visuomotor behavior. Front. Psychol. 1, 173.

Rummery, G., Niranjan, M., 1994. On-line Q-learning using connectionist systems, Cambridge University.

Russell, S., Zimdars, A.L., 2003. Q-decomposition for reinforcement learning agents. Proceedings of ICML-03.

Rutledge, R.B., Dean, M., Caplin, A., Glimcher, P.W., 2010. Testing the reward prediction error hypothesis with an axiomatic model. J. Neurosci. 30, 13525–13536.

Rutledge, R.B., Lazzaro, S.C., Lau, B., Myers, C.E., Gluck, M.A., Glimcher, P.W., 2009. Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J. Neurosci. 29, 15104–15114.

Salamone, J.D., Correa, M., Farrar, A., Mingote, S.M., 2007. Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. Psychopharmacology. 191, 461–482.

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

Seo, H., Barraclough, D.J., Lee, D., 2007. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb. Cortex. 17 (Suppl. 1), i110−117.

Singh S.P., Jaakkola T., Jordan M.I., 1994. Learning without state-estimation in partially observable markovian decision processes. In: International Conference on Machine Learning.

Solomon, R.L., Corbit, J.D., 1974. An opponent-process theory of motivation: I. Temporal dynamics of affect. Psychol. Rev. 81, 119.

Stephens, D.W., Krebs, J.R., 1987. Foraging Theory. Princeton University Press, Princeton, NJ.

Sutton, R.S., 1988. Learning to predict by the methods of temporal differences. Mach. Learn. 3, 9−44.

Sutton, R.S., 1992. Gain adaptation beats least squares? In: Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems. Yale University, New Haven, CT, pp. 161−166.

Sutton, R.S., 1995. TD models: modeling the world at a mixture of time scales. In: International Conference on Machine Learning, pp. 531−539.

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

Sutton, R.S., Precup, D., Singh, S., 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artif. Intell. 112, 181−211.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., et al., 2011. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. 14 (12), 1590−1597.

Tanji, J., Okano, K., Sato, K.C., 1988. Neuronal activity in cortical motor areas related to ipsilateral, contralateral, and bilateral digit movements of the monkey. J. Neurophysiol. 60, 325−343.

Tenenbaum, J.B., Griffiths, T.L., 2001. Structure learning in human causal induction. Adv. Neural Inf. Process. Syst. 13, 59−65.

Thorn, C.A., Atallah, H., Howe, M., Graybiel, A.M., 2010. Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. Neuron. 66, 781−795.

Tobler, P.N., Fiorillo, C.D., Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. Science. 307, 1642−1645.

Todd M.T., Niv Y., Cohen J.D., 2009. Learning to use working memory in partially observable environments through dopaminergic reinforcement. In: Advances in Neural Information Processing Systems 21.

Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., et al., 2009. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science. 324, 1080−1084.

Ungless, M.A., Grace, A.A., 2012. Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. Trends Neurosci. 35 (7), 422−430.

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

von Neumann, J.V., Morgenstern, O., 1947. Theory of Games and Economic Behavior, second ed. Princeton University Press, Princeton, NJ.

Wang, S.S., Denk, W., Hausser, M., 2000. Coincidence detection in single dendritic spines mediated by calcium release. Nat. Neurosci. 3 (12), 1266−1273.

Wang, L.P., Li, F., Wang, D., Xie, K., Shen, X., Tsien, J.Z., 2011. NMDA receptors in dopaminergic neurons are crucial for habit learning. Neuron. 72, 1055−1066.

Watkins, C.J.C.H., Dayan, P., 1992. Q-learning. Mach. Lear. 8, 279−292.

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

Wilson, R.C., Niv, Y., 2011. Inferring relevance in a changing world. Front. Human Neurosci. 5, 189.

Wunderlich, K., Rangel, A., O'Doherty, J.P., 2009. Neural computations underlying action-based decision making in the human brain. Proc. Natl. Acad. Sci. U.S.A. 106, 17199−17204.

Yang, T., Shadlen, M.N., 2007. Probabilistic reasoning by neurons. Nature. 447, 1075−1080.

Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron. 46, 681−692.

Yuille, A., Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. 10, 301−308.