Cognitive, Affective, & Behavioral Neuroscience
2009, 9 (4), 343-364
doi:10.3758/CABN.9.4.343

Reinforcement learning, conditioning, and the brain: Successes and challenges

Tiago V. Maia
Columbia University, New York, New York

The field of reinforcement learning has greatly influenced the neuroscientific study of conditioning. This article provides an introduction to reinforcement learning followed by an examination of the successes and challenges using reinforcement learning to understand the neural bases of conditioning. Successes reviewed include (1) the mapping of positive and negative prediction errors to the firing of dopamine neurons and neurons in the lateral habenula, respectively; (2) the mapping of model-based and model-free reinforcement learning to associative and sensorimotor cortico-basal ganglia-thalamo-cortical circuits, respectively; and (3) the mapping of actor and critic to the dorsal and ventral striatum, respectively. Challenges reviewed consist of several behavioral and neural findings that are at odds with standard reinforcement-learning models, including, among others, evidence for hyperbolic discounting and adaptive coding. The article suggests ways of reconciling reinforcement-learning models with many of the challenging findings, and highlights the need for further theoretical developments where necessary. Additional information related to this study may be downloaded from http://cabn.psychonomic-journals.org/content/supplemental.

Conditioning can be divided into two categories: classical and instrumental (see, e.g., Domjan, 2003). The main difference between the two is that in classical conditioning the outcome (e.g., food) does not depend on the animal's actions, whereas in instrumental conditioning it does. The archetypal account of instrumental conditioning is Thorndike's (1898) *law of effect*. According to this law, in instrumental conditioning animals learn stimulus–response (S–R) associations. Given a situation or stimulus S, the animal tries a response R. If the outcome is positive, the connection between S and R is strengthened; if the outcome is negative, the connection is weakened. In this way, the advantageous response or responses for each situation become more likely.

The idea that artificial systems can learn by a similar process of trial and error can be traced to the early days of artificial intelligence (e.g., Michie, 1961). Reinforcement learning has advanced significantly beyond the law of effect, though. For example, the law of effect does not address the crucial *credit-assignment problem* (Minsky, 1963): When a sequence of actions results in an outcome, how do we determine which actions should get credit for the outcome? For example, an animal searching for food in a maze may make several turns before getting to the food. Some of those turns may bring it closer to the food, and others may lead it away. How does the animal learn which turns were instrumental to get to the food? Reinforcement learning has sophisticated solutions for this problem.

Reinforcement learning studies how agents can learn to behave so as to maximize the rewards and minimize the punishments they receive. This is an optimization problem, one that is at the core of animals' ability to learn to obtain the things they need or want and to avoid those that are harmful or undesirable. Reinforcement learning offers formal, mechanistic solutions to this problem. Furthermore, as discussed below, substantial evidence suggests that the brain may implement some of these solutions.

Reinforcement learning essentially studies how artificial systems can solve instrumental conditioning problems. The relation of reinforcement learning to classical conditioning is perhaps less obvious. However, learning to act so as to maximize rewards and minimize punishments requires the ability to *predict* future rewards and punishments. Reinforcement-learning systems therefore typically incorporate this ability. One method that predicts future reinforcements using temporal differences provides a good account of both behavioral (e.g., Sutton & Barto, 1990) and neural (e.g., McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; O'Doherty et al., 2004; Schultz, 1998, 2002; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000; Suri, 2002) findings on classical conditioning.

The law of effect suggests that all instrumental learning consists of the learning of S–R associations. However, animals also learn action–outcome (A–O) or, more generally, situation–action–outcome (S–A–O) contingencies (Dickinson, 1985, 1994; Tolman, 1932). S–R associations are often called *habits* (Dickinson, 1994; Packard & Knowlton, 2002) because, after they are learned, they are autonomous from the outcome. Actions that are guided by

T. V. Maia, tmaia@columbia.edu

knowledge of A–O or S–A–O contingencies, in contrast, are called *goal directed*. The litmus test of whether something is a habit is whether, after it is learned, it is insensitive to manipulations of the value of the outcome (Dickinson, 1985). For example, if, through extensive training, leverpressing for a sucrose reinforcer has become a habit, rats continue to press the lever even after they have undergone aversion conditioning to the sucrose and are no longer interested in consuming it (Adams, 1982). Goal-directed actions, in contrast, are immediately sensitive to reinforcer revaluation procedures. For example, if, due to limited training, leverpressing for a sucrose reinforcer is still under the control of the goal-directed system, rats stop pressing the lever after they have undergone aversion conditioning to the sucrose (Adams, 1982). Instrumental conditioning may produce both habit and goal-directed learning. The distinction between habits and goal-directed actions maps directly onto the distinction between model-free and model-based reinforcement learning (Daw, Niv, & Dayan, 2005, 2006), as discussed in more detail below.

In short, a full appreciation of modern ideas concerning classical conditioning, instrumental conditioning, habits, and goal-directed actions requires an understanding of reinforcement learning. An understanding of related formal tools, such as Markov decision processes, can also be beneficial for the design of experiments in these areas, because these tools can rigorously describe task contingencies that are vastly more complex than those typically devised by psychologists and neuroscientists.

This article consists of two parts. Part I provides a brief but rigorous introduction to reinforcement learning. Excellent introductions to reinforcement learning, aimed primarily at computer scientists and engineers, are already available in the literature (e.g., Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). The material in Part I differs from such general-purpose introductions in that it focuses specifically on the ideas, equations, and techniques that have proven relevant to understanding the neural bases of conditioning and it is aimed primarily at neuroscientists and psychologists. Readers already familiar with reinforcement learning may want to skip this part. Part II provides an up-to-date review of the evidence concerning reinforcement learning in the brain. That review examines not only the many successes using reinforcement learning to understand the neural bases of conditioning, but also several empirical findings that are at odds with standard reinforcement-learning models and therefore are likely to stimulate future theoretical developments.

PART I
Tutorial Introduction to Reinforcement Learning

The Elements of Reinforcement Learning

The fundamental elements of reinforcement-learning problems are *states*, *actions*, and *reinforcements*. States are the equivalent of Thorndike's (1898) "situations" and represent the current state of the environment (e.g., the agent's location, the stimuli currently present). When the agent is in a given state, it selects an action from among those allowable in that state. Partly as a result of that action, the agent may then transition to a new state and may receive some reinforcement. The goal is to learn which actions to select in each state so as to maximize long-term reinforcement.

More formally, the environment is characterized by a set S of states, and for each state $s \in S$, there is a set $A(s)$ of allowable actions. When the agent is in state s, it selects an action $a \in A(s)$. This action leads the agent to transition to a (possibly different) state s' and may also result in some reinforcement r. The agent's goal is to maximize long-term reinforcement, which is usually defined as the discounted sum of future reinforcements,

$$\sum_{t=\tau}^{\infty} \gamma^{t-\tau} r_t,$$

where τ is the current time, r_t is the reinforcement at time t, and γ is a *discount factor* that discounts future reinforcements ($0 < \gamma < 1$).[1]

Markov Decision Processes

The environment in reinforcement-learning problems can often be described as a *Markov decision process* (MDP). An MDP defines how the environment behaves in response to the agent's actions. Formally, an MDP consists not only of the aforementioned sets S and $A(s)$, but also of two functions: a function T that defines the environment's dynamics and a function R that defines the reinforcement given to the agent. Specifically, $T(s, a, s')$, where $s \in S$, $a \in A(s)$, and $s' \in S$, gives the probability of transitioning to state s' when the agent is in state s and performs action a. In other words, T determines the *transition probabilities* that determine the dynamics of the environment. Note that these transitions need not be deterministic: Performing action a in state s may result in a transition to different states. The reinforcement that the agent receives when it is in state s, selects action a, and transitions to state s' is given by $R(s, a, s')$. Sometimes, such reinforcement is not deterministic even given the triplet $<s, a, s'>$; in those cases, $R(s, a, s')$ is the expected value of the distribution of reinforcements when the agent is in state s, selects action a, and transitions to state s'.

This type of decision process is called a Markov process because it obeys the *Markov property*. In systems that obey this property, the future of the system is independent of its past, given the current state. In other words, if we know the current state, knowing additional information about previous states and reinforcements does not improve our ability to predict future states or reinforcements. More formally, let s_t, a_t, and r_t represent the state, action, and reinforcement at time t, respectively.[2] If time starts at 0 and the current time is τ, the history of the system, H, is given by $H = s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_\tau$. Suppose the agent now selects action a_τ. The Markov property tells us that $P(r_\tau = r, s_{\tau+1} = s \,|\, H, a_\tau) = P(r_\tau = r, s_{\tau+1} = s \,|\, s_\tau, a_\tau)$. In other words, knowing the current state is equivalent to knowing the entire history of the system.

MDPs play a central role in the theory of reinforcement learning precisely because the future depends only on the current state, not on the system's history.[3] This makes the problem both easier to formalize (e.g., T and R can be

functions of the current state rather than of the entire history) and computationally more efficient to solve (because we need only to remember and work with the current state, not the entire history).

It is important to realize that it is the *environment*—not the *agent*—that is assumed to have the Markov property. In fact, because the agent learns, the agent's behavior often *does* depend on its history. For example, an agent is likely to act differently the first time that it is on a given state than it does after being on that state several times and therefore having had the opportunity to learn the consequences of its actions on that state.

Most experiments in psychology and neuroscience involve only deterministic contingencies. Hence, instead of having a function T as above, it is usually sufficient to use a function $T(s, a)$ that gives the next state, given that the animal was in state s and performed action a. The resulting MDP is called a *deterministic MDP*. In deterministic MDPs, the function $R(s, a, s')$ can also be simplified to $R(s, a)$, given that s and a univocally determine s'.

MDPs can be represented as directed graphs in which the states are represented by nodes and the actions are represented by arrows. For each state s, there are $|A(s)|$ arrows leaving the corresponding node, one for each action $a \in A(s)$. In the case of deterministic MDPs, each of those arrows connects to state $T(s, a)$. In some cases, the reinforcement $R(s, a)$ is represented next to the corresponding arrow. In other cases, reinforcements can be directly tied to states; for example, being in a state in which the floor grid is electrified is aversive. In those cases, R is a function of only the current state, $R(s)$, and the values of $R(s)$ are written within the node representing state s. Figure 1 gives an example of an MDP for a fictitious animal-learning experiment.

Policies

The agent selects actions according to a *policy*: a mapping from states to actions. In general, this mapping may be nondeterministic, giving more than one action for the same state. Policies are usually denoted by π; the probability of performing action a in state s is denoted by $\pi(s, a)$. If a policy is deterministic, there is a single action for each state; the action for state s can then be represented as $\pi(s)$.

An *optimal policy* gives, for each state, the best action(s) to perform in that state (i.e., one or more of the actions that lead to the largest expected value for the sum of discounted future reinforcements). In general, there can be more than one optimal policy. Nondeterministic policies can be optimal, but there is always an optimal deterministic policy. When the agent has found an optimal policy, it only has to follow that policy to behave optimally. The goal of reinforcement learning can therefore be restated as learning an optimal policy.

Model-Based and Model-Free Reinforcement Learning

Given a known MDP, several techniques can be used to find an optimal policy (Puterman, 2001, 2005). In general, however, the agent does *not* know the MDP. One possible solution is to attempt to learn the MDP (see section "Learning an MDP" in the supplemental materials) and to then use one of the techniques that can find an optimal policy given an MDP. This approach is called *model based* because it involves learning a model of the environment (the MDP). A different, *model-free* approach attempts to learn an optimal policy directly, without first learning the MDP. Each approach has advantages and disadvantages, and no consensus exists as to which is best from a computational perspective (Kaelbling et al., 1996). The brain may implement both approaches and trade off control between them (Daw, Niv, & Dayan, 2005, 2006), as discussed in more detail below.

Value Functions

The credit-assignment problem is a key problem in reinforcement learning. In MDPs, it arises because an action affects not only the immediate reinforcement but also the probability of transitioning to each possible next state. Since different future states lead to different possibilities for further action and reinforcement, the consequences of an action may be spread over time. When deciding on an action, one therefore needs to take into account not only the immediate reinforcement, but also the *value* of the next state: the expected sum of future reinforcements that the agent can get when it starts from that state.

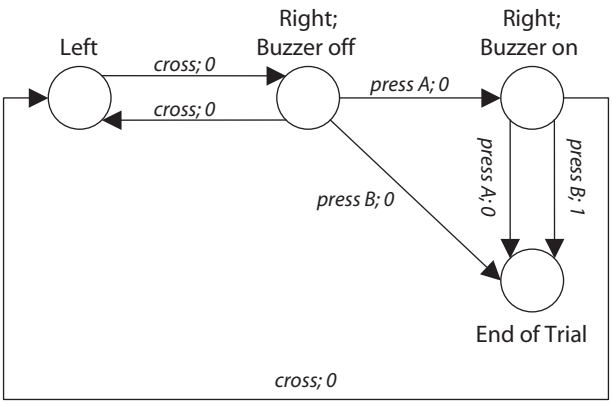

Figure 1. An example Markov decision process for an animal-learning experiment. The apparatus in this fictitious experiment is a shuttlebox with two levers (A and B) on the right compartment. Initially, the animal is put on the left compartment. The animal's task is first to cross to the right compartment and press Lever A. When it has pressed Lever A, a buzzer comes on; the animal then has to press Lever B to receive a food pellet. If the animal presses Lever A after the buzzer is on, or if it presses Lever B before pressing Lever A, the trial is terminated and the animal does not receive food. If, at any point in the trial, the animal crosses back to the left compartment, the cycle is restarted (i.e., the animal has to cross to the right compartment, press Lever A, and then press Lever B). In this experiment, R("Right; Buzzer On", "Press B") = 1 and $R(s, a) = 0$ for all other pairs (s, a). In the figure, the value of $R(s, a)$ is shown next to the corresponding arrow, after the name of the action. Note that not all actions are available in all states. In particular, pressing Lever A and pressing Lever B are not available in the left compartment. Note also that, in every state, the animal has a multitude of other actions available (e.g., sniffing different parts of the apparatus, grooming). These are not shown, because they do not affect state transitions or reinforcements.

Knowing the values of states is extremely useful. When these values are known, the best action can be determined by considering only the available actions' proximate consequences: their likely immediate reinforcements and next states. This circumvents the need to look ahead for an indefinite number of steps to find the delayed consequences of an action. In other words, determining the values of states is a way of solving the credit-assignment problem.

More formally, let the *state-value function*, $V^{\pi}(s)$, be the expected value of the discounted sum of future reinforcements when the agent starts in state s and follows policy π:

$$V^{\pi}(s) = E\left\{\sum_{t=\tau}^{\infty} \gamma^{t-\tau} r_t \mid s_\tau = s\right\}.$$

Both the policy and the transitions between states can be nondeterministic, so the expectation is taken over possible actions (according to π) and state transitions (according to T).

The total expected reinforcement when the agent is in state s, performs action a, and transitions to state s' is $R(s, a, s') + \gamma V^{\pi}(s')$: the sum of the immediate reinforcement $[R(s, a, s')]$ and the reinforcements that will be obtained by starting from s', discounted by γ $[\gamma V^{\pi}(s')]$. If both the policy (π) and the state transitions (T) are deterministic, the value of state s is simply equal to this sum, where a is the action univocally determined by $\pi(s)$ and s' is the successor state univocally determined by $T(s, a)$. In general, however, both the policy and the state transitions are nondeterministic. To get the value of state s, we therefore have to average $R(s, a, s') + \gamma V^{\pi}(s')$ over possible actions a and successor states s', according to the probabilities given by π and T, respectively:

$$V^{\pi}(s) = \sum_{a \in A(s)} \pi(s,a) \sum_{s'} T(s,a,s') \cdot \left[R(s,a,s') + \gamma V^{\pi}(s')\right]. \quad (1)$$

This is known as the Bellman equation.

Temporal-Difference Learning

If the MDP is known, finding the value function $V^{\pi}(s)$ for all $s \in S$ is straightforward (see section "Determining the Value Function for a Known MDP" in the supplemental materials). Often, however, the agent does not know the MDP. Temporal-difference learning (Sutton, 1988) estimates $V^{\pi}(s)$ without knowing or attempting to learn the MDP.

The Bellman equation shows that the value of a state s, $V^{\pi}(s)$, is equal to $R(s, a, s') + \gamma V^{\pi}(s')$ averaged across actions a and successor states s', according to policy π and transition probabilities T, respectively. If the MDP is unknown, however, R and T are unknown. The key insight behind model-free learning is that, each time the agent is in state s, performs some action a, and transitions to some state s', we can take the observed value of $R(s, a, s') + \gamma V^{\pi}(s')$ to be a *sample*. After many visits to state s, the agent likely selected each action a approximately with probability $\pi(s, a)$. Similarly, after the agent selected action a on state s many times, the environment likely transitioned to each possible state s' approximately with probability $T(s, a, s')$. Thus, averaging all the samples of $R(s, a, s') + \gamma V^{\pi}(s')$ provides an estimate of the expected value of $R(s, a, s') + \gamma V^{\pi}(s')$ across all actions a and successor states s', according to policy π and transition probabilities T, respectively—that is, an estimate of $V^{\pi}(s)$. This estimate is commonly denoted $\hat{V}^{\pi}(s)$.

To calculate each sample, $R(s, a, s') + \gamma V^{\pi}(s')$, it would seem that we would need to know $V^{\pi}(s')$. However, we are trying to learn $V^{\pi}(s')$ just as we are trying to learn $V^{\pi}(s)$. We therefore use $\hat{V}^{\pi}(s')$, the estimated value of $V^{\pi}(s')$, instead of the real but unknown $V^{\pi}(s')$. Our sample therefore becomes $R(s, a, s') + \gamma \hat{V}^{\pi}(s')$.

One problem with using a regular average to estimate $V^{\pi}(s)$ is that the Bellman equation assumes that the policy (π), the transition probabilities (T), and the reward function (R) are all fixed. Often, however, while the agent is learning the value function, it is also learning how to behave and is therefore changing its policy π. Furthermore, the environment itself may be nonstationary, so T and R may also be changing. To keep up with these potential changes, recent samples should be weighted more heavily than older samples. Thus, the estimate of the value function, $\hat{V}^{\pi}(s)$, typically uses an *exponential*, *recency-weighted average*—that is, a weighted average in which the weight decays exponentially as one goes toward earlier samples.

Fortunately, there is a simple way of calculating an exponential, recency-weighted average iteratively. Let $x_1, x_2, \ldots, x_n$ be a sequence of numbers, and let $\bar{x}_n$ be an exponential, recency-weighted average of those numbers. Now, suppose we get a subsequent number, x_{n+1}. Then,

$$\bar{x}_{n+1} = \bar{x}_n + \alpha\left[x_{n+1} - \bar{x}_n\right] \quad (2)$$

is an exponential, recency-weighted average of $x_1, x_2, \ldots, x_n, x_{n+1}$ (see, e.g., Sutton & Barto, 1998).

We can use Equation 2 to average the samples of $R(s, a, s') + \gamma \hat{V}^{\pi}(s')$ to get $\hat{V}^{\pi}(s)$. We start by initializing $\hat{V}^{\pi}(s)$ with some value.[4] Then, whenever we get a new sample $R(s, a, s') + \gamma \hat{V}^{\pi}(s')$, we update $\hat{V}^{\pi}(s)$ as follows:

$$\hat{V}^{\pi}(s) \leftarrow \hat{V}^{\pi}(s) + \alpha\left[R(s,a,s') + \gamma \hat{V}^{\pi}(s') - \hat{V}^{\pi}(s)\right]. \quad (3)$$

This is the standard temporal-difference learning equation. Often, $R(s, a, s') + \gamma \hat{V}^{\pi}(s') - \hat{V}^{\pi}(s)$ is denoted by δ:

$$\delta = R(s,a,s') + \gamma \hat{V}^{\pi}(s') - \hat{V}^{\pi}(s). \quad (4)$$

Equation 3 can then also be written as

$$\hat{V}^{\pi}(s) \leftarrow \hat{V}^{\pi}(s) + \alpha\delta. \quad (5)$$

Note that δ represents the difference between the sample of the discounted sum of future reinforcement that we actually got $[R(s, a, s') + \gamma \hat{V}^{\pi}(s')]$ and the one that we predicted $[\hat{V}^{\pi}(s)]$. For this reason, δ is called the *prediction error*. The prediction error represents how things turned out to be relative to what was predicted: Positive prediction errors mean that things turned out better than predicted, and negative prediction errors mean that things turned out worse than predicted.

The prediction error can be seen as the sum of two components: primary reinforcement, $R(s, a, s')$, and the difference in value between the states, $\gamma\hat{V}^{\pi}(s') - \hat{V}^{\pi}(s)$. Thus, for example, an unexpected reward produces a positive prediction error because $R(s, a, s')$ is positive and both $\hat{V}^{\pi}(s)$ and $\hat{V}^{\pi}(s')$ are zero. A positive prediction error can also occur in the absence of reward, if $\gamma\hat{V}^{\pi}(s') > \hat{V}^{\pi}(s)$—that is, if one transitions to a better state. This occurs, for example, when a CS that predicts reward is presented, which induces a change from a state s in which no reward is predicted $[\hat{V}^{\pi}(s) = 0]$ to a state s' in which reward is predicted $[\hat{V}^{\pi}(s') > 0]$. A fully predicted reward in a deterministic environment, on the other hand, does not produce a prediction error because, after learning, $\hat{V}^{\pi}(s) = R(s, a, s') + \gamma\hat{V}^{\pi}(s')$, and therefore $R(s, a, s') + \gamma\hat{V}^{\pi}(s') - \hat{V}^{\pi}(s) = 0$.

Further intuition into temporal-difference learning can be obtained by considering it from two additional perspectives. First, Equation 3 can easily be seen to be equivalent to

$$\hat{V}^{\pi}(s) \leftarrow (1-\alpha)\hat{V}^{\pi}(s) + \alpha\Big[R(s,a,s') + \gamma\hat{V}^{\pi}(s')\Big]. \quad (6)$$

This form shows that the new estimate of $\hat{V}^{\pi}(s)$ is a weighted average of the old estimate and the estimate provided by the current sample $[R(s, a, s') + \gamma\hat{V}^{\pi}(s')]$. For example, if $\alpha = .05$, then 95% of the estimate is determined by the previous estimate and 5% is determined by the current sample. Second, Equation 5 shows that $\hat{V}^{\pi}(s)$ is updated on the basis of the prediction error: If the prediction error is positive, $\hat{V}^{\pi}(s)$ is increased; if the prediction error is negative, $\hat{V}^{\pi}(s)$ is decreased. Thus, $\hat{V}^{\pi}(s)$ is changed in the direction of eliminating the prediction error—that is, in the direction of an accurate prediction.

Finding Optimal Policies With a Model

The previous section showed how value functions can be estimated in a model-free manner, by using temporal-difference learning. As explained below, the actor–critic uses temporal-difference learning in this way; in addition, it has a component devoted to finding optimal policies. Before delving into how the actor–critic tries to find an optimal policy, though, it is useful to see how that is done in the model-based case. As explained below, the actor–critic implements an iterative approximation to this policy-finding procedure (much as temporal-difference learning, which the actor–critic also uses, is also an iterative method to finding the value function).

In the model-based case, once we have a state-value function V^{π} for a policy π, it is easy to find another policy π' that is better than π (unless π is already optimal). This process is called *policy improvement*. We will suppose, as is customary, that all policies are deterministic, so $\pi(s)$ represents the action for state s. Restricting the discussion to deterministic policies does not result in loss of generality, because there is always an optimal policy that is deterministic (Bellman, 1957). The basic idea in policy improvement is that, for each state $s \in S$, we determine whether selecting an action different from $\pi(s)$ and thereafter following policy π is better than following π also at s; if so, we change π to select that other action in s. This procedure improves π, unless π is already optimal.

Let $Q^{\pi}(s, a)$ represent the expected sum of discounted future reinforcements when we are in state s and perform action a. We have already seen that, when we are in state s, perform action a, and transition to state s', the expected sum of discounted future reinforcements is $R(s, a, s') + \gamma V^{\pi}(s')$. To get $Q^{\pi}(s, a)$, we need only to average over possible next states s', weighted by the transition probability $T(s, a, s')$:

$$Q^{\pi}(s,a) = \sum_{s'} T(s,a,s')\Big[R(s,a,s') + \gamma V^{\pi}(s')\Big]. \quad (7)$$

The best action a' in state s is the action that gives the best value of $Q^{\pi}(s, a)$:

$$a' = \arg\max_a Q^{\pi}(s,a). \quad (8)$$

To obtain a new policy π' that improves policy π, we go through all states $s \in S$ and make

$$\pi'(s) = \arg\max_a Q^{\pi}(s,a).$$

If $\pi' = \pi$, then π is already optimal. Otherwise, we can improve π' by following the same two steps used to improve π: (1) finding $V^{\pi'}$ and (2) applying the policy improvement procedure to π'. Repeated application of these steps produces π'', π''', and so on, converging to an optimal policy; this procedure is called *policy iteration*.

The policy improvement procedure presented in this section requires knowledge of the MDP, because R and T are necessary to calculate $Q^{\pi}(s, a)$ (see Equation 7). The next section tackles the problem of finding an optimal policy without a model.

Finding Optimal Policies Without a Model

$Q^{\pi}(s, a)$ represents the value of taking action a in state s and following policy π thereafter, whereas $V^{\pi}(s)$ represents the value of always following π, including in s. The difference $Q^{\pi}(s, a) - V^{\pi}(s)$ therefore represents the relative value of performing a rather than following π in s. If $Q^{\pi}(s, a) - V^{\pi}(s)$ is positive, performing a in s is better than following π; if it is negative, performing a in s is worse than following π.

As noted above, in a model-free approach we cannot calculate $Q^{\pi}(s, a)$, because doing so requires knowledge of R and T. We faced a similar problem when trying to calculate $V^{\pi}(s)$ in a model-free manner, because doing so also required knowledge of R and T. Temporal-difference learning got around this problem by noting that each time one was in state s, performed action a, and transitioned to state s', $R(s, a, s') + \gamma V^{\pi}(s')$ provided a *sample* of $V^{\pi}(s)$. Now, $R(s, a, s') + \gamma V^{\pi}(s')$ also provides a sample of $Q^{\pi}(s, a)$. We can therefore use this sample to estimate the relative value of performing action a rather than following π in s:

$$Q^{\pi}(s,a) - \hat{V}^{\pi}(s) = R(s,a,s') + \gamma\hat{V}^{\pi}(s') - \hat{V}^{\pi}(s) = \delta, \quad (9)$$

where $Q^{\pi}(s, a)$ was replaced by the value of the sample and $\hat{V}$ was used instead of V because in model-free learning we do not know V. This equation demonstrates that the prediction error, δ, shows how much better or worse the action just taken is than following π would have been.

To represent the policy, instead of storing $\pi(s, a)$, let us store a preference $p(s, a)$ for action a in state s, as is done

in *reinforcement comparison* methods (Sutton & Barto, 1998). The policy $\pi(s, a)$ can then be obtained from $p(s, a)$ using a softmax rule, such as

$$\pi(s,a) = \frac{e^{p(s,a)/\tau}}{\sum_{b \in A(s)} e^{p(s,b)/\tau}}, \quad (10)$$

where τ is a temperature parameter that determines the degree of randomness in the action selection.

We can use the prediction error to update the preferences $p(s, a)$. Suppose the agent was in state s, performed action a, and transitioned to state s'. As explained above, the resulting prediction error, δ, is a sample of the relative value of performing action a rather than following π in s. We can therefore use δ to update our preference $p(s, a)$. Specifically, if δ is positive, we should increase $p(s, a)$; if δ is negative, we should decrease $p(s, a)$. Furthermore, we should increase or decrease $p(s, a)$ more when the absolute value of δ is larger. These goals can be accomplished by the following equation:

$$p(s,a) \leftarrow p(s,a) + \beta\delta, \quad (11)$$

where β is a learning rate.

The Actor–Critic

One difficulty with the approach presented in the previous section is that, in order to calculate δ, we need to know $\hat{V}^{\pi}(s)$ (see Equation 9). Each time we update a preference $p(s, a)$, though, we change π, so it would seem that we would need to recalculate $\hat{V}^{\pi}(s)$. The problem is that calculating $\hat{V}^{\pi}$ in a model-free manner is itself an iterative process that requires multiple visits to each state.

The key idea to solve this problem is to try to simultaneously find the best policy and estimate the state-value function for the current policy. Thus, each time the agent is in state s, performs action a, and transitions to state s', it updates $p(s, a)$ using Equation 11, thereby improving the policy, and it also updates $\hat{V}^{\pi}(s)$ using Equation 5, thereby improving its estimate of the state-value function. Naturally, the estimate of the state-value function is always slightly outdated with respect to the current policy, because the policy keeps changing. Nevertheless, these two steps tend to converge toward a good policy.

The actor–critic (Barto, 1995; Barto, Sutton, & Anderson, 1983; see also Witten, 1977) implements these ideas. It consists of two components: the actor and the critic. The actor stores and learns the preferences $p(s, a)$; the critic stores and learns the values $\hat{V}^{\pi}(s)$. When the agent is in state s, it selects an action a according to the probabilities given by Equation 10. The agent then transitions to some state s' and receives reinforcement $R(s, a, s')$, where s' and $R(s, a, s')$ are determined by the environment. Equation 4 is then used to calculate the prediction error, δ. Note that Equation 4 depends on $\hat{V}^{\pi}(s)$ and $\hat{V}^{\pi}(s')$; the values used are those stored by the critic. The critic then updates $\hat{V}^{\pi}(s)$ using Equation 5, and the actor updates $p(s, a)$ using Equation 11. The agent is then ready to select the next action, restarting the entire cycle.

The actor–critic can also be implemented as a connectionist architecture, as shown in Figure 2. The architecture consists of three components: state, actor, and critic. States are represented in the state layer using either distributed (Hinton, McClelland, & Rumelhart, 1986) or localist representations. Distributed representations have the advantage of allowing for generalization over the state space. Localist representations have the advantage of allowing a straightforward interpretation of the architecture, and they will therefore be our main focus. With localist representations, the state layer contains one unit per state. When the system is in a given state, the corresponding unit has a value of 1 and all other units have a value of 0. For simplicity, I will refer to the unit that represents state s as *state-unit s*.

The critic estimates the state-value function and calculates the prediction error. It consists of two units: the value-prediction unit (represented by a circle in Figure 2) and the prediction-error-calculation unit (represented by a square in Figure 2). The value-prediction unit is a linear

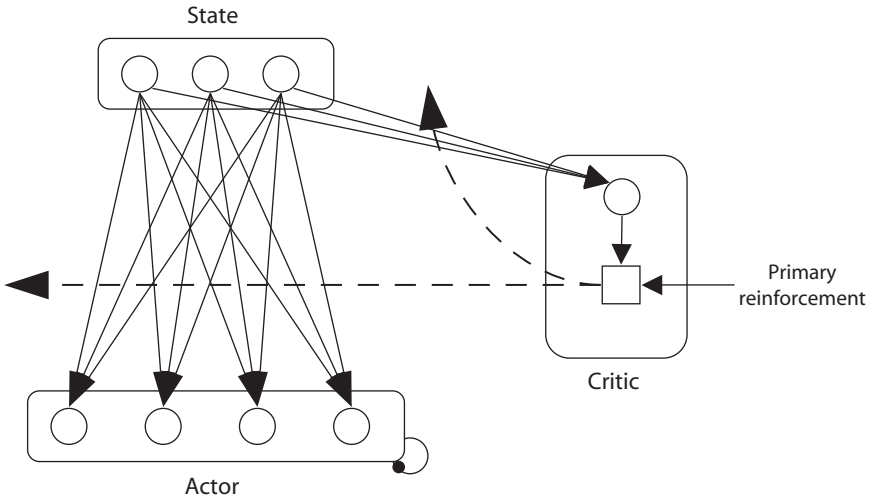

Figure 2. Connectionist implementation of the actor–critic. See the text for details.

unit that estimates the value of the current state. If we represent the weight from state-unit s to the value-prediction unit as w_s, the activation $\hat{V}$ of the value prediction unit is given by

$$\hat{V} = \sum_s w_s x_s, \tag{12}$$

where x_s is the activation of state-unit s and the sum ranges over all state units. With localist state representations, Equation 12 can be simplified to $\hat{V} = w_s$, where s is the current state. In other words, the weight from state-unit s to the value-prediction unit directly represents $\hat{V}(s)$. Therefore, for simplicity, I will refer to that weight as $\hat{V}(s)$.

The prediction-error-calculation unit calculates the prediction error using Equation 4. Let us recall that equation here: $\delta = R(s, a, s') + \gamma\hat{V}(s') - \hat{V}(s)$, where $R(s, a, s')$ is the external reinforcement (labeled "Primary reinforcement" in Figure 2) and $\hat{V}(s)$ and $\hat{V}(s')$ are calculated by the value-prediction unit. Since the value-prediction unit can only calculate the value of the current state, the prediction-error-calculation unit must have a simple memory for the value of the previous state. The prediction error (represented by dashed arrows in Figure 2) is used to update the value of the previous state according to Equation 5;[5] it is also used to update the preference for the action just performed, as explained below.

The actor learns and implements the policy. The actor layer contains one unit per action. For simplicity, I will refer to the unit that represents action a simply as *action-unit a*. Each state-unit s is connected to each action-unit $a \in A(s)$. With localist state representations, the weight w_{sa} connecting state-unit s to action-unit a directly represents the preference $p(s, a)$. I will therefore refer to that weight simply as $p(s, a)$.

The action units are often linear. The activation x_a of action-unit a is therefore given by

$$x_a = \sum_s p(s,a)\, x_s, \tag{13}$$

where x_s is the activation of state-unit s and $p(s, a)$ is the weight connecting state-unit s to action-unit a. With localist state representations, Equation 13 can be simplified to $x_a = p(s, a)$, where s is the current state. The activation of each action unit therefore represents directly the preference for that action in the current state. Equation 10 is then used to obtain a probability distribution over the actions, and an action is chosen according to that distribution. This simulates a competitive process that could also be implemented via lateral inhibition.[6]

When an action a is selected, the agent transitions to a new state s' and receives reinforcement $R(s, a, s')$. The critic then computes the prediction error δ, as explained above, and Equation 11 is applied to update the weight $p(s, a)$ from state-unit s to action-unit a.[7]

To summarize, the critic learns and stores the value function and calculates prediction errors. The actor learns and stores the preferences for each action in each state and selects which action to perform in the current state. Prediction errors are used to update both the critic's value estimates and the actor's state–action preferences. Positive prediction errors lead to increases both in the value estimate for the previous state and in the state–action preference for the action that was just performed in that state. Negative prediction errors have the opposite effect.

PART II
Reinforcement Learning and the Brain

Prediction Errors and Dopamine Neurons

The pioneering work of Wolfram Schultz and collaborators has demonstrated that the phasic activation of midbrain dopamine neurons strongly resembles the prediction error, δ (Mirenowicz & Schultz, 1994; Montague, Dayan, & Sejnowski, 1996; Schultz, 1998, 2002; Schultz et al., 1997; Schultz & Dickinson, 2000; Suri, 2002). Dopamine neurons burst when an animal receives an unexpected reward. However, if the animal has learned to associate a conditioned stimulus (CS) with a subsequent reward, dopamine neurons burst to the CS but not to the reward. Furthermore, if the CS is presented and then the predicted reward is omitted, dopamine neurons fire below baseline approximately at the time at which the reward should have been delivered. All of these findings are consistent with the idea that dopamine neurons report prediction errors: Unpredicted rewards and CSs that predict future rewards produce positive prediction errors; rewards that are fully predicted do not produce any prediction error; and omitted rewards produce negative prediction errors. Phasic dopamine release in target structures, such as the nucleus accumbens, has also been shown to reflect prediction errors (Day, Roitman, Wightman, & Carelli, 2007).

Most fMRI findings in humans have reported activation reflecting prediction errors in areas richly innervated by dopaminergic afferents, such as the striatum and the orbitofrontal cortex (Bray & O'Doherty, 2007; McClure et al., 2003; O'Doherty et al., 2003; Pagnoni, Zink, Montague, & Berns, 2002; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). Furthermore, the presence or absence of activity related to prediction errors in the striatum distinguishes participants who learn to perform optimally from those who do not (Schönberg, Daw, Joel, & O'Doherty, 2007). Because fMRI's BOLD signal correlates better with local field potentials than with neuronal spiking, activity measured using fMRI in a given area may reflect the area's inputs rather than its neuronal spiking (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001). The findings of prediction error activity in areas such as the striatum and the orbitofrontal cortex are therefore believed to reflect dopaminergic input (see, e.g., Knutson & Gibbs, 2007; Niv & Schoenbaum, 2008). This idea is supported by the finding that haloperidol, a dopamine antagonist, and L-DOPA, a dopamine precursor, modulate prediction-error-related activity measured with fMRI in the striatum (Pessiglione et al., 2006). The relative lack of studies reporting activity in the midbrain dopamine areas in humans reflects the difficulties with imaging these small areas using fMRI. A study that used a combination of recently developed fMRI techniques, however, was able to detect prediction-error-related BOLD activity in

the ventral tegmental area (VTA) in humans (D'Ardenne, McClure, Nystrom, & Cohen, 2008).

Habits and Goal-Directed Actions

As noted above, psychology distinguishes between habits and goal-directed actions. Habits correspond to S–R associations (Dickinson, 1985, 1994) and function almost like reflexes: A given stimulus or situation tends to automatically elicit or trigger the response most strongly associated with it. Using the terminology of reinforcement learning, an S–R association is an association between a state and an action. The strength of an S–R association therefore corresponds to the preference for a given action in a given state, $p(s, a)$. As explained above, in the actor–critic, such preferences are stored in the synaptic weights between state units and action units. The actor in the actor–critic therefore learns and stores S–R associations. The process of lateral inhibition (often approximated by selecting an action from the distribution given by Equation 10) selects actions on the basis of the strengths of those S–R associations. Simply put, the actor learns habits and implements them.

The actor–critic, like other model-free reinforcement learning approaches, does not contain a representation of the reward function, $R(s, a, s')$, or of the transition probabilities, $T(s, a, s')$. This type of model therefore does not "know" how its actions relate to reinforcements or state transitions. Instead, the quantities stored by the actor–critic—the values of states, $V(s)$, and the preferences for actions, $p(s, a)$—are "cached" estimates (Daw, Niv, & Dayan, 2005, 2006). Because they are cached, such estimates do not immediately reflect changes in state or action value when the value of the outcome is manipulated. This results in the insensitivity to goal revaluation that characterizes habits in animals (Daw, Niv, & Dayan, 2005, 2006).

Model-based approaches, in contrast, do construct a model that explicitly represents the reward function, $R(s, a, s')$, and the transition probabilities, $T(s, a, s')$. The estimates of the values of states and actions can then be computed online, on the basis of the world model, rather than being cached. Changes such as goal revaluation therefore immediately affect the estimated values of states and actions, which, in turn, immediately affects behavior. In other words, whereas model-free approaches, with their cached estimates, implement habits, model-based approaches, with their online computations, implement goal-directed actions (Daw, Niv, & Dayan, 2005, 2006). Figure 3 presents an example that illustrates the difference between these approaches.

Instrumental conditioning in animals can result in either habits or goal-directed actions, depending on certain parameters of the training procedure, such as the number of training trials (Dickinson, 1985, 1994). Much evidence suggests that habits and goal-directed actions rely on distinct neural systems (Daw, Niv, & Dayan, 2005, 2006; Johnson, van der Meer, & Redish, 2007; Killcross & Coutureau, 2003; Redish & Johnson, 2007; Wickens, Budd, Hyland, & Arbuthnott, 2007; Yin & Knowlton, 2006). Habits (S–R or state–action associations) depend on the dorsolateral striatum (Daw et al., 2005; Johnson et al., 2007; Packard & Knowlton, 2002; Wickens et al., 2007; Yin & Knowlton, 2006). In contrast, goal-directed actions depend on the dorsomedial striatum and prefrontal cortex (Daw et al., 2005; Johnson et al., 2007; Wickens et al., 2007; Yin & Knowlton, 2006). The dorsolateral striatum, which in primates corresponds to a substantial portion of the putamen, participates in the cortico-basal ganglia-thalamo-cortical loop that involves sensorimotor cortices (Alexander, DeLong, & Strick, 1986; Haber, 2003; Yin & Knowlton, 2006). The dorsomedial striatum, which in primates corresponds to a substantial portion of the caudate, participates in the cortico-basal ganglia-thalamo-cortical loop that involves prefrontal associative cortices, which in primates most prominently include the dorsolateral prefrontal cortex (Alexander et al., 1986; Haber, 2003; Yin & Knowlton, 2006). The sensorimotor and associative cortico-basal ganglia-thalamo-cortical loops may therefore be the neural substrates of habits and goal-directed actions, respectively (Yin & Knowlton, 2006).

A key question in the presence of two systems that can guide action is how their control over behavior should be arbitrated. Ideally, each system should control behavior when its predictions are likely to be more accurate than those of the other system. One way of estimating the likely accuracy of a system's predictions is to keep track of its uncertainty. The goal-directed and habit systems may therefore trade off control on the basis of their relative uncertainties (Daw et al., 2005).

The Actor–Critic and the Brain

As explained above, the actor learns and implements habits (S–R associations), which, in the brain, depend on the dorsolateral striatum (Daw et al., 2005; Johnson et al., 2007; Packard & Knowlton, 2002; Wickens et al., 2007; Yin & Knowlton, 2006). It is therefore natural to associate the actor with the dorsolateral striatum.

The role of the critic is to calculate the values of states, $V(s)$, which it then uses to calculate the prediction error (Equation 4). To find the neural correlates of the critic, one therefore needs to look for areas that represent value. Such areas should show neuronal activity during the expectation of reward: Value corresponds to the expected sum of future reinforcements, so value is positive when a reward is expected. Prediction errors, in contrast, correspond to transient activations that signal a change in value, so they occur when an unexpected reward or a reward-predicting stimulus is presented, but not in the period between the presentation of a reward-predicting stimulus and the reward. For an area to represent the critic, it should also project to, and receive projections from, the dopaminergic system, because values are used to calculate prediction errors, which are used to update values.

The ventral (limbic) striatum fulfills all of these requirements. It shows activity during the expectation of reward (Schultz, Apicella, Scarnati, & Ljungberg, 1992; Schultz, Tremblay, & Hollerman, 2000; Setlow, Schoenbaum, & Gallagher, 2003; Wan & Peoples, 2006) and it projects to, and receives projections from, the dopaminergic system (Joel & Weiner, 2000; Oades & Halliday, 1987). Furthermore, unlike other portions of the striatum, the ventral

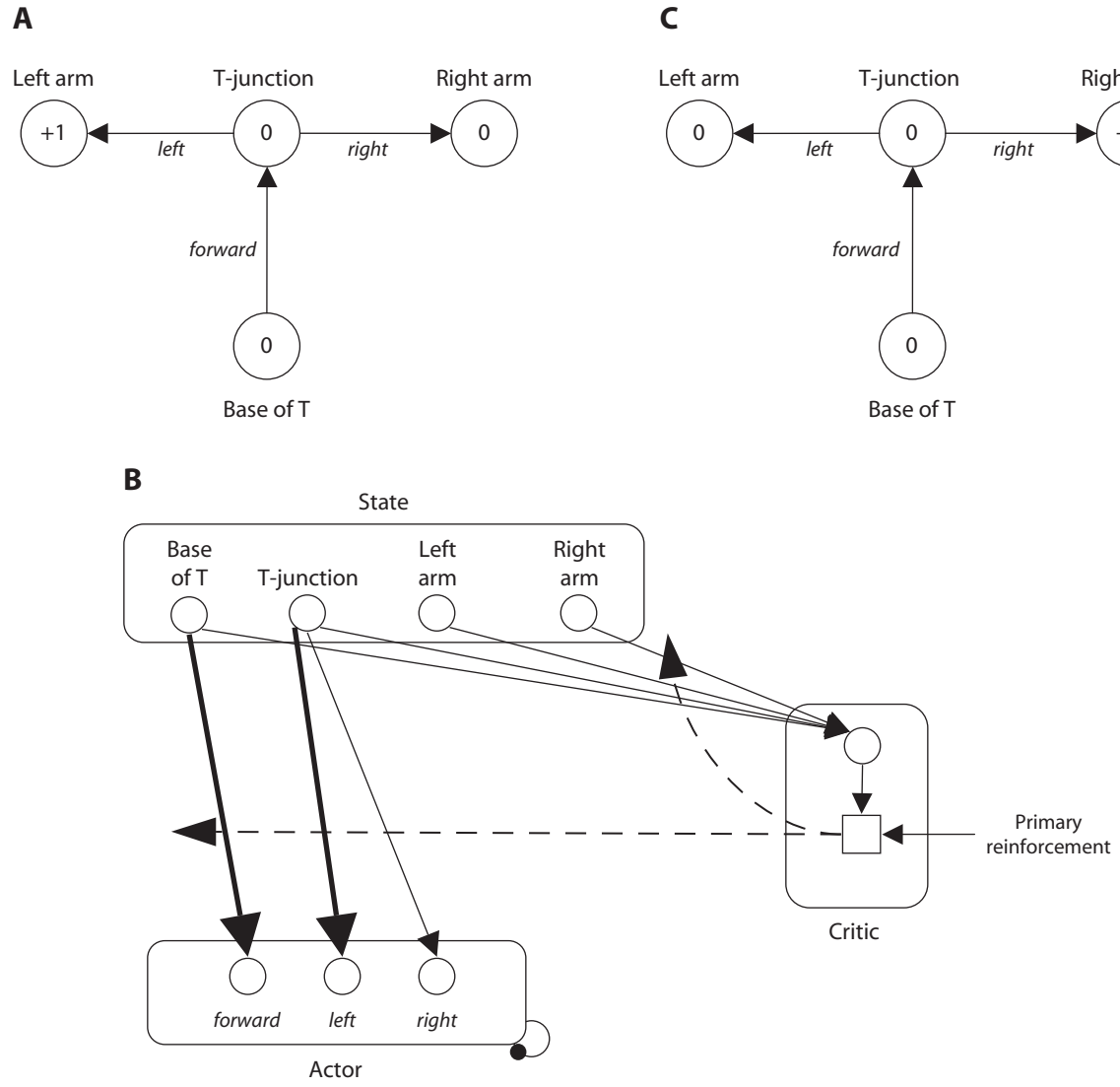

Figure 3. An example illustrating the difference between model-based and model-free approaches. Suppose that an agent is trained in a T-maze in which reinforcement is always available at the end of the left arm, as represented by the Markov decision process (MDP) in panel A. In a model-based approach, the agent learns this MDP and makes a left turn at the T-junction because the model indicates that turning left leads to a state in which there is reinforcement (+1), whereas turning right leads to a state in which there is no reinforcement (0). In contrast, in the model-free approach, the agent turns left because it has learned to associate the T-junction with the action of turning left. This is shown in the actor–critic in panel B, in which the thicker arrow from *T-junction* to *left* indicates a stronger association. (The arrow from *Base of T* to *forward* is also thick because going forward at the base of the T is also reinforced during learning through the propagation of value backward.) Now, suppose that we remove the reinforcement from the end of the left arm and put it instead at the end of the right arm, as shown in the MDP in panel C. Further suppose that we give the agent an opportunity to learn about the new values of the states corresponding to the end of the left arm and the end of the right arm. For example, we might place the agent at the end of each arm but not allow it to consume the reinforcer. If we now place the agent back at the start of the maze, in the model-based approach, the agent will turn right, because it computes the values of states and actions online, according to the updated world model. In contrast, in the model-free approach, the agent will continue to turn left, because that is the action that is associated with the T-junction. The agent would also eventually learn to turn right in the model-free approach, since its cached estimates would eventually catch up with the new contingencies. However, such learning would occur gradually, over multiple trials, since repeated experience would be necessary to update those estimates. In the model-based approach, in contrast, the agent would start turning right immediately.

striatum projects to dopaminergic neurons that innervate all regions of the striatum (Joel & Weiner, 2000). This is consistent with what is required for the critic, which should project not only to dopamine neurons that convey the prediction errors back to itself, but also to dopamine neurons that convey the prediction errors to the actor.

The orbitofrontal cortex and the amygdala also fulfill the three requirements set out above for candidate critic areas. Both the orbitofrontal cortex (Hikosaka & Watanabe, 2000; Schoenbaum, Chiba, & Gallagher, 1998; Schultz, 2000; Schultz et al., 2000; Simmons, Ravel, Shidara, & Richmond, 2007; Tremblay & Schultz,

1999, 2000) and the amygdala (Belova, Paton, & Salzman, 2008; Paton, Belova, Morrison, & Salzman, 2006) show activity during the expectation of reward. Both also project to, and receive projections from, the dopaminergic system (Geisler, Derst, Veh, & Zahm, 2007; Haber & Fudge, 1997; Morecraft, Geula, & Mesulam, 1992; Oades & Halliday, 1987; Ongur, An, & Price, 1998). The orbitofrontal cortex, amygdala, and ventral striatum are closely interconnected anatomically and interrelated functionally (Alexander et al., 1986; Cardinal, Parkinson, Hall, & Everitt, 2002; Cavada, Company, Tejedor, Cruz-Rizzolo, & Reinoso-Suarez, 2000; Gray, 1999; Middleton & Strick, 2001; Rempel-Clower, 2007; Schoenbaum & Roesch, 2005; Zald & Kim, 2001), so these three areas may work together to implement the critic.

An fMRI study in humans directly addressed the distinction between actor and critic by using an instrumental conditioning task and a yoked Pavlovian conditioning task that involved the same value predictions as the instrumental conditioning task but no action selection (O'Doherty et al., 2004). BOLD activation in the ventral striatum correlated with prediction errors in both tasks, but activation in the dorsal striatum correlated only with prediction errors in the instrumental conditioning task. This is consistent with the idea that the critic and actor are related to the ventral and dorsal striatum, respectively: The critic would be expected to be involved in both tasks, whereas the actor would be expected to be involved only in the instrumental task (which required action learning and selection). Studies using fMRI without an explicit reinforcement-learning model have also found that activity in the dorsal striatum depends on (Tricomi, Delgado, & Fiez, 2004), or is at least enhanced by (Elliott, Newman, Longe, & Deakin, 2004), the existence of an instrumental contingency. An electrophysiological study in rats provided further evidence for the idea that the critic is related to the ventral striatum and the actor is related to the dorsal striatum (Daw, 2003): Neurons in the ventral striatum represented predicted rewards rather than actions, whereas neurons in the dorsal striatum represented actions rather than predicted rewards.

The idea that the striatum is involved in implementing both the actor and the critic is consistent with extensive evidence that dopamine plays a role in corticostriatal plasticity (for reviews, see Calabresi, Pisani, Centonze, & Bernardi, 1997; Jay, 2003; Reynolds & Wickens, 2002). If, as is commonly assumed, states are represented in the cortex, the role of dopamine in modulating corticostriatal plasticity is precisely what is required for prediction errors to modulate plasticity in the projections from the state to the actor and to the critic, as in the actor–critic.

The striatum can be divided into two compartments: discrete patches, called *striosomes*, and an extrastriosomal matrix, which surrounds them (Graybiel, 1990; Graybiel & Ragsdale, 1978). An influential idea, distinct from the hypothesized dorsal–ventral organization of the actor–critic, is that the actor and critic might be implemented by the matrix and striosomes, respectively (Houk, Adams, & Barto, 1995). Consistent with this idea, self-stimulation of the striosomes, but not the matrix, leads to rapid response acquisition (White & Hiroi, 1998), suggesting that the striosomes may represent reward or value.[8] This assignment of actor and critic to matrix and striosomes remains controversial, though (Joel, Niv, & Ruppin, 2002). Its main impetus was evidence from studies in rats that neurons in the striosomes, but not the matrix, project to dopaminergic neurons in the substantia nigra pars compacta (SNc; Gerfen, 1984, 1985). Such neuroanatomical organization, however, has never been demonstrated in primates, and even the evidence in rats has been limited to connections between striosomes and a small group of dopamine neurons (Joel et al., 2002). In fact, a more recent study using single-axon tracing in monkeys reported that neurons in the striosomes project to the substantia nigra pars reticulata (SNr) rather than to the SNc (Lévesque & Parent, 2005). This study traced only six striosomal neurons, however, and they were all from the same striosome, so additional research is needed to confirm the generalizability of these findings. No firm conclusions are therefore currently possible regarding the putative involvement of striosomes and matrix in implementing the critic and actor, respectively.

It is also worth noting that the connectivity of striosomes is not uniform throughout the striatum. In particular, the orbitofrontal cortex projects strongly to striosomes in the anterior and ventromedial striatum, but not to those in the dorsolateral striatum (Eblen & Graybiel, 1995). Given the evidence implicating the orbitofrontal cortex in value representation, this pattern of connectivity suggests that, even if striosomes are involved in implementing the critic, such a role could be limited to striosomes in the anterior and ventromedial striatum.[9] This would be generally consistent with the idea that the critic is implemented in the ventral striatum, or, more generally, in the limbic striatum, which includes both ventromedial and anterior portions (Haber, 2003).

Negative Prediction Errors

Recent findings have demonstrated that the firing of neurons in the lateral habenula may represent something akin to a negative prediction error (Matsumoto & Hikosaka, 2007, 2009a). Neurons in the lateral habenula are excited by punishments, by CSs that predict punishments, by the omission of rewards, and by CSs that predict the omission of rewards; they are inhibited by rewards and by CSs that predict rewards (Gao, Hoffman, & Benabid, 1996; Matsumoto & Hikosaka, 2007, 2009a). Furthermore, such responses are modulated by probabilistic expectancy (Matsumoto & Hikosaka, 2009a). For example, neurons in the lateral habenula are more excited by a CS that is followed by punishment 100% of the time than by a CS that is followed by punishment 50% of the time; conversely, they are more inhibited by a CS that predicts reward 100% of the time than by a CS that predicts reward 50% of the time. The responses to unconditioned stimuli also seem to reflect a (probabilistically modulated) prediction error. For example, neurons do not respond to a fully predicted reward, they are inhibited by a reward that was predicted with 50% probability, and they are inhibited even more by an unpredicted reward. Similarly, they are more excited by an unpredicted punishment than by a punishment that

was predicted with 50% probability. All of these findings are consistent with the idea that the lateral habenula reports a negative prediction error. However, some patterns of firing depart slightly from a negative prediction error. For example, neurons in the lateral habenula fire when a punishment is fully expected (Matsumoto & Hikosaka, 2009a).

The pattern of firing of neurons in the lateral habenula is almost exactly the mirror image of the pattern of firing of dopaminergic neurons in the VTA and SNc. This is reminiscent of the type of mirror opponency between appetitive and aversive systems that had previously been suggested on computational grounds (Daw, Kakade, & Dayan, 2002). Furthermore, excitation in the lateral habenula precedes inhibition in dopaminergic neurons (Matsumoto & Hikosaka, 2007), stimulation of the lateral habenula inhibits dopaminergic neurons (Christoph, Leonzio, & Wilcox, 1986; Ji & Shepard, 2007; Matsumoto & Hikosaka, 2007), and simultaneous recordings from the lateral habenula and SNc reveal negative cross-correlations between the activities in the two nuclei (Gao et al., 1996). The lateral habenula, with its direct projections to the VTA and SNc (Geisler & Trimble, 2008), may therefore play an important role in determining the response of dopaminergic neurons. Consistent with the symmetry in the patterns of firing of the lateral habenula and dopaminergic system and with the idea of opponency between these systems, there is also evidence that the dopaminergic system inhibits the lateral habenula: Activity in the lateral habenula is decreased following administration of dopamine agonists (McCulloch, Savaki, & Sokoloff, 1980), and it is increased following administration of dopamine antagonists (McCulloch et al., 1980; Ramm, Beninger, & Frost, 1984) and following dopamine denervation (Kozlowski & Marshall, 1980; Wooten & Collins, 1981). Whether the inhibitory effects of dopamine on the lateral habenula act via the direct VTA projection to the lateral habenula (Geisler & Trimble, 2008) or, for example, via the basal ganglia (Brown & Wolfson, 1983; Wooten & Collins, 1981) is not known. Regardless, these findings suggest that the dopaminergic system and the lateral habenula may interact via mutual inhibition to compute both positive and negative prediction errors.

An important open question is how the negative prediction errors reported by neurons in the lateral habenula are conveyed to target structures involved in valuation and action selection. One possibility is that negative prediction errors are conveyed by the inhibition of dopamine neurons. Due to the low baseline firing rate of dopamine neurons, their inhibition does not seem able to represent negative prediction errors quantitatively when one looks at a short, fixed postevent interval (Bayer & Glimcher, 2005). However, the duration of the pause in firing of dopamine neurons does seem to code negative prediction errors quantitatively (Bayer, Lau, & Glimcher, 2007).

The findings concerning the response of dopamine neurons to punishments or to CSs that predict punishment have been mixed, with evidence for inhibition, excitation, and no response (Guarraci & Kapp, 1999; Horvitz, 2000; Mantz, Thierry, & Glowinski, 1989; Mirenowicz & Schultz, 1996; Schultz & Romo, 1987). One study in anesthetized rats suggested that dopamine neurons might be uniformly inhibited by aversive stimuli and that previous findings to the contrary could have been recording from nondopaminergic neurons (Ungless, Magill, & Bolam, 2004). More recent findings, however, suggest that two distinct populations of dopamine neurons may exist: one that is inhibited by, and one that is excited by, aversive events and stimuli that predict aversive events (Brischoux, Chakraborty, Brierley, & Ungless, 2009; Matsumoto & Hikosaka, 2009b). Both populations are excited by rewards and stimuli that predict rewards (Matsumoto & Hikosaka, 2009b), so the excitation of dopamine neurons in the population that is excited by aversive events is not sufficient to signal a negative prediction error: Such excitation could reflect either a positive or a negative prediction error. Negative prediction errors could, however, be efficiently decoded from the activity of both populations, by subtracting the activity of the population that is inhibited by aversive events from the activity of the population that is excited by aversive events. This would cancel out positive prediction errors and simultaneously enhance the signal for negative prediction errors. Whether such a decoding mechanism might exist in the brain is currently unknown.

The lateral habenula has prominent projections to the dorsal and median raphe (Geisler & Trimble, 2008; Herkenham & Nauta, 1979), so another possibility is that negative prediction errors are conveyed to target structures by serotonin, as has been hypothesized on the basis of computational considerations (Daw et al., 2002). No direct evidence for serotonergic signaling of negative prediction errors exists yet, though. One study showed an improvement in punishment prediction, but not in reward prediction, in healthy volunteers under acute tryptophan depletion (Cools, Robinson, & Sahakian, 2008). This was interpreted as being consistent with a role for serotonin in signaling negative prediction errors, under the assumption that tryptophan depletion would lower tonic levels of serotonin, thereby increasing the signal-to-noise ratio of phasic serotonin responses. Although a similar interaction between tonic and phasic levels has been proposed for dopamine (Grace, 1991, 2000), such an effect for acute tryptophan depletion remains hypothetical, so the evidence provided by this study for the hypothesis that serotonin conveys negative prediction errors is indirect. Additional evidence consistent with that hypothesis comes from the finding that neurotoxin lesions of serotonergic pathways preclude punishment from reducing the probability of the actions that result in such punishment (Thiébot, Hamon, & Soubrié, 1983; Tye, Everitt, & Iversen, 1977). In temporal-difference-based reinforcement learning, the probability of an action is reduced when the action is followed by a negative prediction error. If lesions of serotonergic pathways eliminate such errors, punishments would not result in the suppression of actions, as has been found (Thiébot et al., 1983; Tye et al., 1977). One study reported that neurons in the dorsal raphe did not respond significantly to stressors that elicited strong sympathetic and defensive reactions (Wilkinson & Jacobs, 1988) and that would be expected to produce a negative prediction error. However,

that study reported firing rates that were averaged across minutes of exposure to the stressor and could therefore have missed the effects of brief, phasic responses elicited by the onset of the stressor. Indeed, a recent abstract reported phasic activation of serotonin neurons in the dorsal raphe in response to noxious foot shocks (Schweimer, Brierley, & Ungless, 2008).

Despite this circumstantial evidence for a role of serotonin in reporting negative prediction errors (see also Daw et al., 2002), the extant evidence is mostly consistent with an inhibitory, rather than excitatory, effect of the lateral habenula on the dorsal and median raphe. Electrical stimulation of the lateral habenula has generally been reported to strongly inhibit serotonergic neurons in the raphe (Park, 1987; Stern, Johnson, Bronzino, & Morgane, 1979; Wang & Aghajanian, 1977), resulting in decreased serotonin release in target structures (Reisine, Soubrié, Artaud, & Glowinski, 1982). Furthermore, lesions of the lateral habenula increased serotonin levels in the dorsal raphe in rat models of depression in which such levels were decreased (Yang, Hu, Xia, Zhang, & Zhao, 2008). In contrast to these findings, though, one microdialysis study found an increase in the release of serotonin in the striatum with stimulation of the lateral habenula (Kalen, Strecker, Rosengren, & Bjorklund, 1989), and another study reported that stimulation of the lateral habenula may result in either inhibition or excitation of serotonergic neurons in the raphe, depending on the frequency of the stimulation (Ferraro, Montalbano, Sardo, & La Grutta, 1996). Importantly, the computational theory that proposed that phasic serotonin responses may signal negative prediction errors suggested a very different role for tonic serotonin (Daw et al., 2002). Since the effects of stimulation of the lateral habenula on tonic versus phasic responses of serotonergic neurons are unknown, and since lesions of the lateral habenula likely affect both types of responses, electrophysiological recordings that distinguish between phasic and tonic responses are necessary to determine whether serotonin plays a role in conveying negative prediction errors.

One study recorded from neurons in the dorsal raphe (with unconfirmed serotonergic status) and from dopamine neurons in the SNc while monkeys performed a saccade task with biased rewards (Nakamura, Matsumoto, & Hikosaka, 2008). Monkeys had to fixate a central point, and after a delay, a target would come on. The location of the target (left or right) indicated whether the reward available would be small or large. Monkeys then had to saccade to the target (either immediately or after a delay, depending on the version of the task), at which point they received the reward. The study found that dorsal raphe neurons responded tonically both after the receipt of reward and during the delay between the target onset and the receipt of reward, with some neurons being activated by small rewards and others by large rewards. The study was not geared toward testing the hypothesis that phasic activation of dorsal raphe neurons conveys negative prediction errors, but one finding from the study seems inconsistent with that hypothesis. After a variable number of trials, the location–reward contingencies were switched. This should elicit a positive prediction error for the large reward that was unexpectedly delivered at the location that was previously associated with a small reward, and it should elicit a negative prediction error for the small reward that was unexpectedly delivered at the location that was previously associated with a large reward. Consistent with their role in reporting prediction errors, dopamine neurons exhibited an increase in activity in the former case and a decrease in activity in the latter case. Importantly, these changes in activity lasted for only one trial; in subsequent trials, the amount of reward was already predicted correctly, so dopamine neurons stopped responding. Neurons in the dorsal raphe, in contrast, did not show such transient changes in activity. Instead, neurons that previously responded to small rewards continued to respond to small rewards, and neurons that previously responded to large rewards continued to respond to large rewards. In other words, dorsal raphe neurons seemed to be coding the size of the received reward rather than any prediction error. More definitive conclusions regarding a possible role of serotonin in conveying negative prediction errors will, however, have to wait for an electrophysiological study aimed directly at testing that hypothesis. Such a study would ideally use conditioning with aversive events and confirm the serotonergic status of recorded neurons.

States, Actions, and Rewards in the Brain

States, actions, and rewards are key theoretical constructs in reinforcement learning. This section elaborates these concepts from a neurobiological perspective.

States. In the computational reinforcement-learning community, states are understood to be potentially complex representations that may include a wealth of information beyond the current sensory stimuli (see, e.g., Sutton & Barto, 1998). Most experimental work in neuroscience, in contrast, has equated states with sensory stimuli (e.g., CSs), typically extending the notion of state only to deal with temporal representations. One study, however, highlights the importance of a more comprehensive view of states to account for the firing of dopamine neurons in more complex experimental paradigms (Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004). In that study, the probability of a trial being a rewarded trial increased with the number of trials since the last rewarded trial. The firing of dopamine neurons was consistent with this conditional probability: The fewer the trials since the last rewarded trial (i.e., the more an unrewarded trial was expected), the more dopamine neurons fired to a rewarded trial and the less they were suppressed by an unrewarded trial. A temporal-difference model whose states included information about the number of trials since the last rewarded trial successfully modeled the behavior of dopamine neurons; a temporal-difference model that did not include that information failed to capture the behavior of dopamine neurons (Nakahara et al., 2004). If the representation of states can include something as complex as the number of trials since the last rewarded trial, it may conceivably include a variety of other types of information when relevant, possibly including motivations, emotions, memories, inferences, and so on. Future research should address this possibility.

States and their representation are understudied neuroscientifically. Key topics that have been explored at some length in machine learning have not yet been addressed experimentally in neuroscience. One important question is whether states (and actions) in the brain are discrete, as is the case in standard reinforcement learning, or continuous, as is the case in continuous reinforcement learning (Doya, 1996; Santamaria, Sutton, & Ram, 1998; Smart & Kaelbling, 2000). Another important question is how the brain deals with situations in which complete information about the state is not available. Computationally, this is addressed by partially observable Markov decision processes, in which the agent does not know exactly which state it is in, but instead keeps a probability distribution over states (Cassandra, Kaelbling, & Littman, 1994; Kaelbling, Littman, & Cassandra, 1998; Monahan, 1982). Future research should seek to address these issues experimentally.

Actions. In standard reinforcement learning, actions are elemental, indivisible constructs. In real life, however, actions are structured hierarchically (Botvinick, Niv, & Barto, in press; Botvinick & Plaut, 2004; Miller, Galanter, & Pribram, 1960). For example, making a cup of tea can be divided into several subactions, one of which is pouring water into the cup, which in turn can be subdivided into several lower-level subactions, one of which is lifting the teapot, which in turn can be subdivided further, and so on. Hierarchical reinforcement learning (Barto & Mahadevan, 2003; Parr, 1998; Sutton, Precup, & Singh, 1999) addresses this hierarchical structure.

Neurons in the prefrontal cortex (Fujii & Graybiel, 2005) and striatum (Barnes, Kubota, Hu, Jin, & Graybiel, 2005; Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999) code the beginnings and ends of action sequences, suggesting that frontostriatal circuits may participate in the "chunking" of action sequences (Graybiel, 1998) that characterizes hierarchical reinforcement learning. Indeed, the interaction between dorsolateral prefrontal cortex and striatum to create action chunks has been modeled computationally using hierarchical reinforcement learning (De Pisapia & Goddard, 2003). It has also been suggested that, to support hierarchical reinforcement learning, the actor may include the dorsolateral prefrontal cortex and the critic may include the orbitofrontal cortex (Botvinick et al., in press).

One consequence of the hierarchical structure of actions is that actions can be described at several levels of abstraction (e.g., *making tea* vs. a specific set of motor commands). Some evidence suggests that the dorsolateral prefrontal cortex and associative areas of the striatum may be involved in more abstract action representations, whereas sensorimotor cortices and dorsolateral striatum may be involved in more motor representations that are effector specific (e.g., specific to the hand that was trained on a given task; Yin & Knowlton, 2006). Furthermore, with habitization, actions shift from associative to sensorimotor cortico-basal ganglia networks and become increasingly specific to the effector with which they were trained (Yin & Knowlton, 2006). Future research should seek to determine whether this means that true habits consist of sequences of relatively elemental actions or whether instead habits can also include actions at higher levels in the hierarchy.

Rewards. Most experimental work on reinforcement learning in animals uses primary rewards, such as food or juice. Experiments with humans sometimes use primary rewards, but more often use money, which is typically considered a secondary reward. Secondary rewards are stimuli that acquire rewarding properties by virtue of being paired with primary rewards, a process that is well captured by value learning in the critic.

Neuroimaging work in humans suggests that cognitive feedback also engages the dopaminergic system and striatum (Aron et al., 2004; Rodriguez, Aron, & Poldrack, 2006). These studies used a probabilistic classification learning task, in which participants had to learn to associate stimulus features with responses (or categories). The feedback, however, consisted of presentation of the correct response, rather than primary or secondary rewards or punishments. Nevertheless, BOLD activity in the midbrain and striatum appeared to reflect prediction errors in the task. Consistent with these findings, patients with Parkinson's disease are impaired in probabilistic classification learning tasks (Knowlton, Mangels, & Squire, 1996; Shohamy et al., 2004), which further suggests a role for dopamine in this type of learning. The notion of reward should therefore probably also include, at a minimum, internal signals comparing one's response with the correct response. In fact, it would not be surprising if, at least in humans, the notion of reward were rather broad. In part, this seems obvious. Humans constantly deal with rewards (e.g., doing well on an exam, getting a promotion at work) that seem fairly removed from primary rewards. The idea that such abstract rewards may tap into the same reinforcement-learning machinery as primary rewards do, and that they may therefore have similar effects on behavior, is a powerful one.

Violations of the Normative Assumptions of Reinforcement Learning

The subjective value of a predicted reward is a function of its magnitude, delay, and probability (see, e.g., Ho, Mobini, Chiang, Bradshaw, & Szabadi, 1999), and dopamine neurons have been shown to be sensitive to these three variables (Fiorillo, Tobler, & Schultz, 2003; Kobayashi & Schultz, 2008; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004; Tobler, Fiorillo, & Schultz, 2005). Reinforcement-learning models capture the influence of all three variables on subjective value. Furthermore, they do so according to normative assumptions consistent with classical economic theories. However, much research in behavioral economics has documented ubiquitous departures from such normative assumptions in human behavior (e.g., Camerer & Loewenstein, 2004; Hastie & Dawes, 2001; Prelec & Loewenstein, 1991), and similar findings have been obtained with a variety of other species. This section addresses behavioral and neural findings concerning the effects of magnitude, delay, and probability on subjective value, and it discusses their implications for reinforcement-learning models of behavior and the brain.

Magnitude and subjective value. In the same way that, in psychophysics, the perceived intensity of a stimulus is not a linear function of objective stimulus intensity (e.g., Stevens, 1957), the subjective value (or utility) of a reinforcement also is not a linear function of its objective value (e.g., Bernoulli, 1738/1954; Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Although this is typically ignored in reinforcement-learning models, incorporating this nonlinearity into the models would be straightforward: It would simply require passing the primary reinforcements through an appropriate nonlinear function, such as the S-shaped value function from prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992).

Delay and subjective value. Time in temporal-difference learning is often represented by a tapped delay line (Figure 4). Use of the tapped delay line, together with the standard equation for the prediction error, implies that the value of future reinforcements is discounted exponentially (see section "The Mathematics of Temporal Discounting" in the supplemental materials). Exponential discounting is also used in the discounted utility model of classical economics (Samuelson, 1937). Exponential discounting is often considered normative in classical economics because it applies the same discount at each time (see Frederick, Loewenstein, & O'Donoghue, 2002, for a critical review).

Despite the normative appeal of exponential discounting, humans, monkeys, rats, and pigeons all discount future reinforcements hyperbolically, not exponentially (Ainslie, 1975; Frederick et al., 2002; Green & Myerson, 2004; Kable & Glimcher, 2007; Kim, Hwang, & Lee, 2008; Kobayashi & Schultz, 2008; Mazur, 1987, 2001, 2007; Myerson & Green, 1995; Richards, Mitchell, de Wit, & Seiden, 1997). Furthermore, activity related to delayed rewards in the ventral striatum, medial prefrontal cortex, and posterior cingulate cortex of humans seems to represent hyperbolically discounted values (Kable & Glimcher, 2007). Consistent with these findings, recent electrophysiological results suggest that dopamine neurons also discount future reinforcements hyperbolically (Kobayashi & Schultz, 2008). When monkeys were presented with CSs that predicted reward at different delays, dopamine neurons responded most for CSs associated with rewards at short delays; furthermore, although both exponential and hyperbolic discounting models provided good fits to the neuronal responses, the hyperbolic model fit better overall (Kobayashi & Schultz, 2008).

In temporal-difference models with tapped delay lines, exponential discounting results from the repeated application of the discount factor γ (see section "The Mathematics of Temporal Discounting" in the supplemental materials). Other forms of discounting would also be consistent with such models, as long as one could apply discounting iteratively, when the prediction error is calculated (see Equation 4). In temporal-difference models, the overall shape of the discounting function results from repeated application of that equation. In hyperbolic discounting, however, the discount factor is not constant; instead, it is a function of the time until the reward (see section "The Mathematics of Temporal Discounting" in the supplemental materials). Given that the time of the reward is known only when the reward is eventually received, the information needed for the calculation of the prediction error would not be available when the system transitions between states prior to eventually receiving the reward. Thus, prediction errors could not be calculated at the right time. Achieving hyperbolic discounting in temporal-difference systems is therefore far from straightforward.

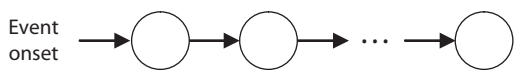

Figure 4. The tapped delay line. Several units are connected in series; activation is transmitted from one unit to the next, but with a delay. On event onset, the first unit is activated; at each time step, that activation moves forward one unit. By seeing which unit is active at a given time, one can determine the time since event onset.

One possible solution comes from the idea that discounting results from the interaction of two value systems, one rational and one emotional (Loewenstein, 1996; McClure, Laibson, Loewenstein, & Cohen, 2004; Metcalfe & Mischel, 1999). These ideas find formal expression in models that behave quasi-hyperbolically, even though both systems discount exponentially (Kable & Glimcher, 2007), or, alternatively, one system discounts exponentially and the other evaluates only rewards that are available immediately (Elster, 1979; Laibson, 1997; McClure et al., 2004). Such models are consistent with a wide variety of behavioral findings (Frederick et al., 2002; Kable & Glimcher, 2007; O'Donoghue & Rabin, 1999), and there is even some fMRI evidence to support them (McClure et al., 2004; but see Kable & Glimcher, 2007). Given that these models are based on exponential discount functions, they do not pose the same difficulties for temporal-difference learning as true hyperbolic discounting does. Attempting to integrate these models with temporal-difference learning is therefore likely to be a fruitful avenue for future research.

Adapting temporal-difference models to be consistent with hyperbolic-like discounting might be seen by some as moving these models into the realm of descriptive models, effectively abandoning their normative motivations. However, hyperbolic discounting can also be justified normatively, because it can be seen as maximizing the rate of gain in repetitive choices (Kacelnik, 1997) or as resulting from Bayesian updating when the hazard rate is uncertain (Sozou, 1998).

Probability and subjective value. Expected utility theory, first proposed by Bernoulli in the 18th century (Bernoulli, 1738/1954) and later reinterpreted and axiomatized by von Neumann and Morgenstern (1944), proposes that the utility of probabilistic outcomes is combined according to expected value.[10] Despite the normative appeal of this idea, humans and other animals weight the probabilities of reinforcement nonlinearly (Green & Myerson, 2004; Ho et al., 1999; Kahneman & Tversky, 1979; Rachlin, Raineri, & Cross, 1991; Trepel, Fox, & Poldrack, 2005; Tversky & Kahneman, 1992). Furthermore,

substantial evidence now exists for nonlinear coding of reinforcement probability in several brain regions (Berns, Capra, Chappelow, Moore, & Noussair, 2008; Hsu, Krajbich, Zhao, & Camerer, 2009; Paulus & Frank, 2006; Tobler, Christopoulos, O'Doherty, Dolan, & Schultz, 2008; but see Abler, Walter, Erk, Kammerer, & Spitzer, 2006, and Preuschoff, Bossaerts, & Quartz, 2006).

The firing of dopamine neurons is modulated by outcome probability (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005). If different CSs predict reward with different probabilities, phasic responses to the CS increase with increasing probability of reward; conversely, phasic responses to the reward itself decrease with increasing probability of reward (Fiorillo et al., 2003; Morris et al., 2004). Consistent with these findings, BOLD activity in the human VTA is also modulated by outcome probability (D'Ardenne et al., 2008). Furthermore, dopamine neurons can integrate information about the probability and magnitude of reward (Tobler et al., 2005). However, it is not known whether the firing of dopamine neurons reflects linear or nonlinear weighting of probabilities. Existing plots of neuronal firing in relation to reward probability or expected value (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005) are not sufficient to reach a definitive conclusion because it is not always clear whether they are truly linear or have a nonlinear component.

Reinforcement-learning models are generally concerned with maximizing expected value, ignoring other characteristics of the reinforcement distribution. Even in model-based reinforcement learning, information about reinforcements typically is limited to their expected value: As noted in the "Markov Decision Processes" section above, $R(s, a, s')$ represents the expected value of reinforcements, not their full distribution. Model-based reinforcement-learning methods therefore appear likely to weight probabilities linearly.

Temporal-difference models also do not represent information about the probabilities of reinforcement. For example, the only quantities stored by the actor–critic are the values of states, $V(s)$, and the preferences for actions, $p(s, a)$. These quantities approximate running averages due to the gradual nature of learning. At first sight, then, it would seem that temporal-difference models might also behave according to expected value, weighting probabilities linearly. However, the averages obtained in temporal-difference models weight older reinforcements exponentially less. This leads to interesting effects of rates of reinforcement. Since rates of reinforcement reflect experientially learned probabilities of reinforcement, this may introduce interesting nonlinearities into the weighting of probabilities. For example, temporal-difference models may tend to underweight small probabilities: Low rates of reinforcement result in old, highly discounted reinforcements, which influence current estimates only weakly. Similarly, temporal-difference models may tend to overweight large probabilities: High rates of reinforcement result in recent reinforcements, which disproportionately boost current estimates.[11] Such patterns of probability discounting are consistent with human behavior: When outcomes are learned experientially, as is the case in temporal-difference models, humans also underweight small probabilities and overweight large probabilities (see Tobler et al., 2008, for discussion and relevant citations).[12] Temporal-difference models may therefore offer an explanation for the underweighting of small probabilities and the overweighting of large probabilities with experientially learned outcomes, although this needs to be tested in future research. More generally, future research should seek to clarify the relation between reinforcement-learning models and probability discounting in humans and in other animals.

Other Challenges for Temporal-Difference Models

Several other findings are inconsistent with standard temporal-difference learning models. Those findings, however, are consistent with more advanced reinforcement-learning techniques, such as semi-Markov dynamics or Bayesian reinforcement learning. This section reviews these findings and their interpretation in reinforcement-learning terms.

Variable reward timing. When a reward is delivered earlier than expected, dopamine neurons fire to the reward but do not then show a depression at the time at which the reward had originally been expected (Hollerman & Schultz, 1998). Although the standard temporal-difference model with a tapped delay line correctly predicts the firing to the early reward, it predicts a depression (negative prediction error) at the time at which the reward had originally been expected (see, e.g., Daw, Courville, & Touretzky, 2006; Suri & Schultz, 1999). One possible solution to this problem is to assume that the early reward somehow resets the tapped delay line, possibly via an attentional mechanism (Suri & Schultz, 1999). However, this solution seems somewhat ad hoc (Daw, Courville, & Touretzky, 2006; Suri, 2002). A more principled account of these findings can be obtained using semi-Markov dynamics and partial observability (Daw, Courville, & Touretzky, 2006), reinforcement-learning techniques that are beyond the scope of the present article (see Bradtke & Duff, 1995; Kaelbling et al., 1998; Puterman, 2005). Other representations of time in temporal-difference models can also account for these findings (Ludvig, Sutton, & Kehoe, 2008).

Semi-Markov dynamics provide a rich representation of time as a continuous variable (Bradtke & Duff, 1995; Puterman, 2005), which may also prove useful in future research to account for other findings in which the timing of events varies, including the effects of variable reward delays on temporal discounting by dopamine neurons (Kobayashi & Schultz, 2008). In fact, by replacing the tapped delay line with a continuous representation of time, semi-Markov systems may be able to more seamlessly include hyperbolic temporal discounting.

Adaptive coding by dopamine neurons. Dopamine neurons do not seem to code the value of prediction errors in absolute terms. Instead, they seem to change their sensitivity (i.e., gain) depending on the anticipated range or standard deviation of reward magnitudes (Tobler et al., 2005). When three different CSs predict 0.05, 0.15, and 0.50 ml of juice, each with probability .5, the firing of do-

pamine neurons to the juice delivery is similar in the three cases, even though the values of the three prediction errors are quite different (Tobler et al., 2005). Dopamine neurons seem to adapt to the range or standard deviation of possible rewards, which may help them discriminate among the likely values of the anticipated rewards (Tobler et al., 2005). Temporal-difference models, in contrast, do not show such adaptive coding. The prediction error in those models reflects the magnitude of the difference between actual and expected rewards, unscaled by the anticipated range or standard deviation of rewards (but see Preuschoff & Bossaerts, 2007).

The effect that this adaptive coding by dopamine neurons may have on target structures is not known. Do target structures receive information about the anticipated range or standard deviation of rewards, and are they therefore capable of recovering the raw, unscaled prediction errors? Some evidence suggests that dopamine neurons themselves may convey information about the anticipated range or standard deviation of rewards to target structures. When one CS predicts a small or medium reward, another predicts a medium or large reward, and a third predicts a small or large reward, each with probability .5, sustained firing of dopamine neurons during the expectation of reward is larger for the third CS than for the other two (Fiorillo et al., 2003). Since the third CS is associated with a larger range and standard deviation of predicted rewards, such sustained firing may convey information about that range or standard deviation to target structures.

It is not known whether target structures use the information about the range or standard deviation of rewards possibly conveyed by dopamine neurons to recover the unscaled prediction error. A formal analysis of the classical Rescorla–Wagner model (Rescorla & Wagner, 1972) in terms of least-squares learning theory shows that value predictions can be updated directly using scaled prediction errors (Preuschoff & Bossaerts, 2007). It seems likely that scaled prediction errors could similarly be used to update the preferences for actions, which would make the recovery of the unscaled prediction errors unnecessary. Future computational research should seek to formalize a complete reinforcement-learning system that uses such scaled prediction errors to learn the values of both states and actions. Such a model could then be used to make predictions that could be tested neuroscientifically.

Risk in the brain. The aforementioned sustained response of dopamine neurons during the anticipation of rewards has been interpreted as a risk (i.e., variance) signal (Fiorillo et al., 2003; Schultz et al., 2008).[13] When different CSs predict a reward with different probabilities, the greater the risk (i.e., the closer the probability is to .5), the stronger that sustained response (Fiorillo et al., 2003; but see Morris et al., 2004). Consistent with these findings, BOLD activation in the midbrain in humans has also been found to reflect risk (Preuschoff et al., 2006). In addition, a circuit-based model has suggested a possible biological mechanism for the calculation of this risk signal, via the corelease of gamma-aminobutyric acid (GABA) and substance P from axons of medium spiny neurons that synapse on GABAergic neurons in the SNr and on dopaminergic neurons in the SNc (Tan & Bullock, 2008).

An alternative account of the sustained, ramping response of dopamine neurons during the expectation of reward has been proposed (Niv, Duff, & Dayan, 2005; but see Fiorillo, Tobler, & Schultz, 2005; Preuschoff & Bossaerts, 2007). That account is based on the observation that phasic dopaminergic responses code positive and negative prediction errors asymmetrically, because of the low baseline firing rate of dopamine neurons. Such asymmetry implies that averaging prediction errors across trials in experiments with probabilistic rewards would produce the type of ramping activity that was observed (Niv et al., 2005). The ramping activity was therefore suggested to be an artifact of averaging across trials, rather than to reflect a within-trial risk signal (Niv et al., 2005). However, the ramping activity has also been observed in single trials (Fiorillo et al., 2005). Furthermore, the averaging of prediction errors across trials results in ramping activity only under certain representations of time, such as a tapped delay line (Fiorillo et al., 2005). More sophisticated representations of time that seem to provide better accounts for other findings concerning dopamine neurons, such as models that include semi-Markov dynamics (Daw, Courville, & Touretzky, 2006), would not show such artifactual ramping.

Studies using fMRI have shown that activity in the human orbitofrontal cortex and striatum correlates with outcome risk (Preuschoff et al., 2006; Tobler, O'Doherty, Dolon, & Schultz, 2007). Furthermore, risk-related activity in the orbitofrontal cortex correlates with participants' risk attitudes (Tobler et al., 2007). The orbitofrontal cortex and striatum have strong bidirectional connections with the dopaminergic system, so the risk-related BOLD activity in these areas could potentially reflect an incoming dopaminergic risk signal (Preuschoff et al., 2006) or, alternatively, contribute to determine the dopaminergic risk signal. Risk-related activity has also been found in the posterior cingulate cortex in monkeys (McCoy & Platt, 2005) and humans (Huettel, Stowe, Gordon, Warner, & Platt, 2006). It has been hypothesized that the putative dopaminergic risk signal could influence such activity via the anterior cingulate cortex (McCoy & Platt, 2005).

To summarize, although some controversy remains around the idea that dopamine neurons convey a risk signal, much evidence suggests that the brain represents risk. Standard reinforcement-learning algorithms, in contrast, do not calculate or use risk (variance). However, the representation of both the expected value and the variance of predicted reinforcements in the brain provides a richer characterization of the reinforcement distribution than does the use of expected value alone. Such richer characterization is consistent with Bayesian approaches to reinforcement learning, which work with the full probability distributions over values (Daw et al., 2005; Dearden, Friedman, & Russell, 1998; Engel, Mannor, & Meir, 2003). Bayesian reinforcement-learning methods, although computationally challenging, have the advantage of providing a principled, normative solution to the important exploration-versus-exploitation dilemma (Dearden

et al., 1998; Poupart, Vlassis, Hoey, & Regan, 2006). The explicit representation of variance in the brain may therefore offer important computational advantages.

CONCLUSIONS

In the last decade, ideas and techniques from reinforcement learning have played a central role in advancing our understanding of conditioning and choice. Reinforcement learning provides an integrated understanding of how agents can learn to behave so as to maximize rewards and minimize punishments—an ability that is the very pillar of survival and functioning in the animal world. Reinforcement-learning agents and animals face formally similar problems, so it is not surprising that reinforcement learning provides conceptual insights into animal learning. More remarkably, though, actual mechanisms, architectures, and algorithms developed in the reinforcement-learning community seem to map directly onto neural processes in animals.[14] The use of reinforcement learning to advance our understanding of the neural bases of conditioning can fairly be considered to be one of the prime examples in cognitive and behavioral neuroscience of the power of an integrated theoretical and empirical approach.

Despite these remarkable successes, some challenges remain. Reinforcement learning is in many respects similar to neoclassical economic theories. Both provide mathematically elegant, coherent, and normatively motivated accounts of behavior.[15] However, much research has documented systematic deviations in human and animal behavior and neurophysiology from the normative assumptions shared by both theories. In economics, this led to a schism between neoclassical and behavioral economics. Among researchers interested in reinforcement-learning models of the brain and behavior, it seems likely that theorists of different persuasions will also address these challenges differently. Some may abandon the normative assumptions of reinforcement learning and strive to make the models more accurate descriptively. Others may seek to maintain a normative or seminormative perspective, striving to explain how the empirical findings are consistent with a rational or semirational (Daw, Courville, & Dayan, 2008) system. Both approaches are likely to prove useful, and indeed they may cross-fertilize each other. For example, the hyperbolic model of temporal discounting was developed on descriptive grounds (Ainslie, 1975; Mazur, 1987), but subsequent work suggested how it could be justified normatively (Kacelnik, 1997; Sozou, 1998).

One of the key advantages of reinforcement-learning approaches is that they address all aspects of reinforcement-related learning and behavior in an integrated, coherent system. Reinforcement learning was developed to build integrated agents that maximize rewards and minimize punishments. Reinforcement-learning systems therefore integrate all pieces of the puzzle that are necessary to achieve that goal (from value learning, to exploration, to action selection, etc.). The main challenge in seeking to bring reinforcement-learning models into alignment with the empirical findings will likely not be in finding ways of explaining individual findings; in fact, ways of reconciling reinforcement-learning models with many of the discordant findings were suggested above. The key will be to ensure that the explanations of the individual findings maintain the coherence of the entire system. An integrative theory that explains all of the challenging findings in an integrated, coherent system is currently lacking and should be a priority for future research.

AUTHOR NOTE

I am very grateful to the anonymous reviewers for their many valuable suggestions and to Nathaniel Daw for very useful discussions concerning the relation of habits and goal-directed actions to model-free and model-based reinforcement learning. Correspondence concerning this article should be addressed to T. V. Maia, Department of Psychiatry, Columbia University, 1051 Riverside Drive, Unit 74, New York, NY 10032 (e-mail: tmaia@columbia.edu).

REFERENCES

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, **31**, 790-795.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, **34B**, 77-98.

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, **82**, 463-496.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, **9**, 357-381.

Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., & Poldrack, R. A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology*, **92**, 1144-1152.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, **437**, 1158-1161.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215-232). Cambridge, MA: MIT Press.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory & Applications*, **13**, 343-379.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, & Cybernetics*, **13**, 834-846.

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

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology*, **98**, 1428-1439.

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

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *Journal of Neuroscience*, **28**, 10023-10030.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, **22**, 23-36. (Original work published 1738)

Berns, G. S., Capra, C. M., Chappelow, J., Moore, S., & Noussair, C. (2008). Nonlinear neurobiological probability weighting functions for aversive outcomes. *NeuroImage*, **39**, 2047-2057.

Botvinick, M. M., Niv, Y., & Barto, A. G. (in press). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*. doi:10.1016/j.cognition.2008.08.011

Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, **111**, 395-429.

Bradtke, S. J., & Duff, M. O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In G. Tesauro, D. S.

Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 393-400). Cambridge, MA: MIT Press.

Bray, S., & O'Doherty, J. (2007). Neural coding of reward-prediction error signals during classical conditioning with attractive faces. *Journal of Neurophysiology*, **97**, 3036-3045.

Brischoux, F., Chakraborty, S., Brierley, D. I., & Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proceedings of the National Academy of Sciences*, **106**, 4894-4899.

Brown, L. L., & Wolfson, L. I. (1983). A dopamine-sensitive striatal efferent system mapped with [^{14}C]deoxyglucose in the rat. *Brain Research*, **261**, 213-229.

Calabresi, P., Pisani, A., Centonze, D., & Bernardi, G. (1997). Synaptic plasticity and physiological interactions between dopamine and glutamate in the striatum. *Neuroscience & Biobehavioral Reviews*, **21**, 519-523.

Camerer, C. F., & Loewenstein, G. (2004). Behavioral economics: Past, present, future. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 3-51). Princeton, NJ: Princeton University Press.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, **26**, 321-352.

Cassandra, A. R., Kaelbling, L. P., & Littman, M. L. (1994). Acting optimally in partially observable stochastic domains. In *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 1023-1028). Menlo Park, CA: AAAI Press.

Cavada, C., Company, T., Tejedor, J., Cruz-Rizzolo, R. J., & Reinoso-Suarez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex: A review. *Cerebral Cortex*, **10**, 220-242.

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

Cools, R., Robinson, O. J., & Sahakian, B. (2008). Acute tryptophan depletion in healthy volunteers enhances punishment prediction but does not affect reward prediction. *Neuropsychopharmacology*, **33**, 2291-2299.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, **319**, 1264-1267.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 431-452). Oxford: Oxford University Press.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, **18**, 1637-1677.

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

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, **8**, 1704-1711.

Daw, N. D., Niv, Y., & Dayan, P. (2006). Actions, policies, values, and the basal ganglia. In E. Bezard (Ed.), *Recent breakthroughs in basal ganglia research* (pp. 111-130). New York: Nova Science.

Day, J. J., Roitman, M. F., Wightman, R. M., & Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, **10**, 1020-1028.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 761-768). Menlo Park, CA: AAAI Press.

De Pisapia, N., & Goddard, N. H. (2003). A neural model of frontostriatal interactions for behavioural planning and action chunking. *Neurocomputing*, **52-54**, 489-495. doi:10.1016/S0925-2312(02)00753-1

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B*, **308**, 67-78.

Dickinson, A. (1994). Instrumental conditioning. In N. J. Mackintosh (Ed.), *Animal learning and cognition* (pp. 45-79). San Diego: Academic Press.

Domjan, M. (2003). *The principles of learning and behavior* (5th ed.). Belmont, CA: Thomson/Wadsworth.

Doya, K. (1996). Temporal difference learning in continuous time and space. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 1073-1079). Cambridge, MA: MIT Press.

Eblen, F., & Graybiel, A. M. (1995). Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *Journal of Neuroscience*, **15**, 5999-6013.

Elliott, R., Newman, J. L., Longe, O. A., & Deakin, J. F. W. (2004). Instrumental responding for rewards is associated with enhanced neuronal response in subcortical reward systems. *NeuroImage*, **21**, 984-990.

Elster, J. (1979). *Ulysses and the sirens: Studies in rationality and irrationality*. Cambridge: Cambridge University Press.

Engel, Y., Mannor, S., & Meir, R. (2003). Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In *Proceedings of the 20th International Conference on Machine Learning* (pp. 154-161). Menlo Park, CA: AAAI Press.

Ferraro, G., Montalbano, M. E., Sardo, P., & La Grutta, V. (1996). Lateral habenular influence on dorsal raphe neurons. *Brain Research Bulletin*, **41**, 47-52. doi:10.1016/0361-9230(96)00170-0

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

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioral & Brain Functions*, **1**, 7.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, **40**, 351-401.

Fujii, N., & Graybiel, A. M. (2005). Time-varying covariance of neural activities recorded in striatum and frontal cortex as monkeys perform sequential-saccade tasks. *Proceedings of the National Academy of Sciences*, **102**, 9032-9037.

Gao, D. M., Hoffman, D., & Benabid, A. L. (1996). Simultaneous recording of spontaneous activities and nociceptive responses from neurons in the pars compacta of substantia nigra and in the lateral habenula. *European Journal of Neuroscience*, **8**, 1474-1478.

Geisler, S., Derst, C., Veh, R. W., & Zahm, D. S. (2007). Glutamatergic afferents of the ventral tegmental area in the rat. *Journal of Neuroscience*, **27**, 5730-5743.

Geisler, S., & Trimble, M. (2008). The lateral habenula: No longer neglected. *CNS Spectrums*, **13**, 484-489.

Gerfen, C. R. (1984). The neostriatal mosaic: Compartmentalization of corticostriatal input and striatonigral output systems. *Nature*, **311**, 461-464. doi:10.1038/311461a0

Gerfen, C. R. (1985). The neostriatal mosaic. I. Compartmental organization of projections from the striatum to the substantia nigra in the rat. *Journal of Comparative Neurology*, **236**, 454-476.

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

Grace, A. A. (2000). The tonic/phasic model of dopamine system regulation and its implications for understanding alcohol and psychostimulant craving. *Addiction*, **95**(Suppl. 2), S119-S128.

Gray, T. S. (1999). Functional and anatomical relationships among the amygdala, basal forebrain, ventral striatum, and cortex: An integrative discussion. In J. F. McGinty (Ed.), *Advancing from the ventral striatum to the amygdala: Implications for neuropsychiatry and drug abuse* (Annals of the New York Academy of Sciences, Vol. 877, pp. 439-444). New York: New York Academy of Sciences.

Graybiel, A. M. (1990). Neurotransmitters and neuromodulators in the basal ganglia. *Trends in Neurosciences*, **13**, 244-254.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning & Memory*, **70**, 119-136.

Graybiel, A. M., & Ragsdale, C. W., Jr. (1978). Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining. *Proceedings of the National Academy of Sciences*, **75**, 5723-5726.

GREEN, L., & MYERSON, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, **130**, 769-792.

GUARRACI, F. A., & KAPP, B. S. (1999). An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential Pavlovian fear conditioning in the awake rabbit. *Behavioural Brain Research*, **99**, 169-179.

HABER, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, **26**, 317-330.

HABER, S. N., & FUDGE, J. L. (1997). The interface between dopamine neurons and the amygdala: Implications for schizophrenia. *Schizophrenia Bulletin*, **23**, 471-482. doi:10.1093/schbul/23.3.471

HASTIE, R., & DAWES, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. New York: Sage.

HERKENHAM, M., & NAUTA, W. J. (1979). Efferent connections of the habenular nuclei in the rat. *Journal of Comparative Neurology*, **187**, 19-47.

HERTWIG, R., BARRON, G., WEBER, E. U., & EREV, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, **15**, 534-539. doi:10.1111/j.0956-7976.2004.00715.x

HIKOSAKA, K., & WATANABE, M. (2000). Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral Cortex*, **10**, 263-271.

HINTON, G. E., MCCLELLAND, J. L., & RUMELHART, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 77-109). Cambridge, MA: MIT Press.

HO, M.-Y., MOBINI, S., CHIANG, T.-J., BRADSHAW, C. M., & SZABADI, E. (1999). Theory and method in the quantitative analysis of "impulsive choice" behaviour: Implications for psychopharmacology. *Psychopharmacology*, **146**, 362-372.

HOLLERMAN, J. R., & SCHULTZ, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, **1**, 304-309.

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

HOUK, J. C., ADAMS, J. L., & BARTO, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249-270). Cambridge, MA: MIT Press.

HSU, M., KRAJBICH, I., ZHAO, C., & CAMERER, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience*, **29**, 2231-2237. doi:10.1523/jneurosci.5296-08.2009

HUETTEL, S. A., STOWE, C. J., GORDON, E. M., WARNER, B. T., & PLATT, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron*, **49**, 765-775.

JAY, T. M. (2003). Dopamine: A potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology*, **69**, 375-390. doi:10.1016/S0301-0082(03)00085-6

JI, H., & SHEPARD, P. D. (2007). Lateral habenula stimulation inhibits rat midbrain dopamine neurons through a $GABA_A$ receptor-mediated mechanism. *Journal of Neuroscience*, **27**, 6923-6930. doi:10.1523/jneurosci.0958-07.2007

JOEL, D., NIV, Y., & RUPPIN, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, **15**, 535-547.

JOEL, D., & WEINER, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, **96**, 451-474.

JOG, M. S., KUBOTA, Y., CONNOLLY, C. I., HILLEGAART, V., & GRAYBIEL, A. M. (1999). Building neural representations of habits. *Science*, **286**, 1745-1749.

JOHNSON, A., VAN DER MEER, M. A. A., & REDISH, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, **17**, 692-697.

KABLE, J. W., & GLIMCHER, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, **10**, 1625-1633.

KACELNIK, A. (1997). Normative and descriptive models of decision making: Time discounting and risk sensitivity. In G. R. Bock & G. Cardew (Eds.), *Characterizing human psychological adaptations* (Ciba Foundation Symposium, No. 208, pp. 51-70). New York: Wiley.

KAELBLING, L. P., LITTMAN, M. L., & CASSANDRA, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, **101**, 99-134.

KAELBLING, L. P., LITTMAN, M. L., & MOORE, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, **4**, 237-285.

KAHNEMAN, D., & TVERSKY, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263-291.

KALEN, P., STRECKER, R. E., ROSENGREN, E., & BJORKLUND, A. (1989). Regulation of striatal serotonin release by the lateral habenula-dorsal raphe pathway in the rat as demonstrated by in vivo microdialysis: Role of excitatory amino acids and GABA. *Brain Research*, **492**, 187-202.

KILLCROSS, S., & COUTUREAU, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, **13**, 400-408.

KIM, S., HWANG, J., & LEE, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, **59**, 161-172.

KIRKLAND, K. L. (2002). High-tech brains: A history of technology-based analogies and models of nerve and brain function. *Perspectives in Biology & Medicine*, **45**, 212-223. doi:10.1353/pbm.2002.0033

KNIGHT, F. H. (1921). *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

KNOWLTON, B. J., MANGELS, J. A., & SQUIRE, L. R. (1996). A neostriatal habit learning system in humans. *Science*, **273**, 1399-1402.

KNUTSON, B., & GIBBS, S. E. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, **191**, 813-822.

KOBAYASHI, S., & SCHULTZ, W. (2008). Influence of reward delays on responses of dopamine neurons. *Journal of Neuroscience*, **28**, 7837-7846. doi:10.1523/jneurosci.1600-08.2008

KOZLOWSKI, M. R., & MARSHALL, J. F. (1980). Plasticity of [^{14}C]2-deoxy-D-glucose incorporation into neostriatum and related structures in response to dopamine neuron damage and apomorphine replacement. *Brain Research*, **197**, 167-183.

LAIBSON, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, **112**, 443-477.

LÉVESQUE, M., & PARENT, A. (2005). The striatofugal fiber system in primates: A reevaluation of its organization based on single-axon tracing studies. *Proceedings of the National Academy of Sciences*, **102**, 11888-11893. doi:10.1073/pnas.0502710102

LOEWENSTEIN, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior & Human Decision Processes*, **65**, 272-292.

LOGOTHETIS, N. K., PAULS, J., AUGATH, M., TRINATH, T., & OELTERMANN, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, **412**, 150-157. doi:10.1038/35084005

LUDVIG, E. A., SUTTON, R. S., & KEHOE, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, **20**, 3034-3054.

MANTZ, J., THIERRY, A. M., & GLOWINSKI, J. (1989). Effect of noxious tail pinch on the discharge rate of mesocortical and mesolimbic dopamine neurons: Selective activation of the mesocortical system. *Brain Research*, **476**, 377-381.

MATSUMOTO, M., & HIKOSAKA, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, **447**, 1111-1115.

MATSUMOTO, M., & HIKOSAKA, O. (2009a). Representation of negative motivational value in the primate lateral habenula. *Nature Neuroscience*, **12**, 77-84.

MATSUMOTO, M., & HIKOSAKA, O. (2009b). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, **459**, 837-841.

MAZUR, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior: Vol. 5. The effect of delay and of intervening events on reinforcement value* (pp. 55-73). Hillsdale, NJ: Erlbaum.

MAZUR, J. E. (2001). Hyperbolic value addition and general models of animal choice. *Psychological Review*, **108**, 96-112.

MAZUR, J. E. (2007). Choice in a successive-encounters procedure and hyperbolic decay of reinforcement. *Journal of the Experimental Analysis of Behavior*, **88**, 73-85.

MCCLURE, S. M., BERNS, G. S., & MONTAGUE, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, **38**, 339-346.

MCCLURE, S. M., LAIBSON, D. I., LOEWENSTEIN, G., & COHEN, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, **306**, 503-507.

MCCOY, A. N., & PLATT, M. L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience*, **8**, 1220-1227.

MCCULLOCH, J., SAVAKI, H. E., & SOKOLOFF, L. (1980). Influence of dopaminergic systems on the lateral habenular nucleus of the rat. *Brain Research*, **194**, 117-124.

METCALFE, J., & MISCHEL, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, **106**, 3-19.

MICHIE, D. (1961). Trial and error. In S. A. Barnett & A. McLaren (Eds.), *Science survey* (Part 2, pp. 129-145). Harmondsworth, U.K.: Penguin.

MIDDLETON, F. A., & STRICK, P. L. (2001). A revised neuroanatomy of frontal–subcortical circuits. In D. G. Lichter & J. L. Cummings (Eds.), *Frontal–subcortical circuits in psychiatric and neurological disorders*. New York: Guilford.

MILLER, G. A., GALANTER, E., & PRIBRAM, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

MINSKY, M. (1963). Steps toward artificial intelligence. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 406-450). New York: McGraw-Hill.

MIRENOWICZ, J., & SCHULTZ, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, **72**, 1024-1027.

MIRENOWICZ, J., & SCHULTZ, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, **379**, 449-451.

MONAHAN, G. E. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, **28**, 1-16.

MONTAGUE, P. R., DAYAN, P., & SEJNOWSKI, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, **16**, 1936-1947.

MORECRAFT, R. J., GEULA, C., & MESULAM, M. M. (1992). Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. *Journal of Comparative Neurology*, **323**, 341-358.

MORRIS, G., ARKADIR, D., NEVET, A., VAADIA, E., & BERGMAN, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, **43**, 133-143.

MYERSON, J., & GREEN, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, **64**, 263-276.

NAKAHARA, H., ITOH, H., KAWAGOE, R., TAKIKAWA, Y., & HIKOSAKA, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron*, **41**, 269-280.

NAKAMURA, K., MATSUMOTO, M., & HIKOSAKA, O. (2008). Reward-dependent modulation of neuronal activity in the primate dorsal raphe nucleus. *Journal of Neuroscience*, **28**, 5331-5343.

NIV, Y., DUFF, M. O., & DAYAN, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral & Brain Functions*, **1**, 6. doi:10.1186/1744-9081-1-6

NIV, Y., & SCHOENBAUM, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, **12**, 265-272.

OADES, R. D., & HALLIDAY, G. M. (1987). Ventral tegmental (A10) system: Neurobiology. 1. Anatomy and connectivity. *Brain Research*, **434**, 117-165.

O'DOHERTY, J., DAYAN, P., FRISTON, K., CRITCHLEY, H., & DOLAN, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, **38**, 329-337.

O'DOHERTY, J., DAYAN, P., SCHULTZ, J., DEICHMANN, R., FRISTON, K., & DOLAN, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452-454.

O'DONOGHUE, T., & RABIN, M. (1999). Doing it now or later. *American Economic Review*, **89**, 103-124.

ONGUR, D., AN, X., & PRICE, J. L. (1998). Prefrontal cortical projections to the hypothalamus in macaque monkeys. *Journal of Comparative Neurology*, **401**, 480-505.

PACKARD, M. G., & KNOWLTON, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, **25**, 563-593.

PAGNONI, G., ZINK, C. F., MONTAGUE, P. R., & BERNS, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, **5**, 97-98.

PARK, M. R. (1987). Monosynaptic inhibitory postsynaptic potentials from lateral habenula recorded in dorsal raphe neurons. *Brain Research Bulletin*, **19**, 581-586.

PARR, R. (1998). *Hierarchical control and learning for Markov decision processes.* Unpublished doctoral dissertation, University of California, Berkeley.

PATON, J. J., BELOVA, M. A., MORRISON, S. E., & SALZMAN, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, **439**, 865-870.

PAULUS, M. P., & FRANK, L. R. (2006). Anterior cingulate activity modulates nonlinear decision weight function of uncertain prospects. *NeuroImage*, **30**, 668-677.

PESSIGLIONE, M., SEYMOUR, B., FLANDIN, G., DOLAN, R. J., & FRITH, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, **442**, 1042-1045.

POUPART, P., VLASSIS, N., HOEY, J., & REGAN, K. (2006). An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 697-704). New York: ACM.

PRELEC, D., & LOEWENSTEIN, G. (1991). Decision making over time and under uncertainty: A common approach. *Management Science*, **37**, 770-786.

PREUSCHOFF, K., & BOSSAERTS, P. (2007). Adding prediction risk to the theory of reward learning. In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), *Reward and decision making in corticobasal ganglia networks* (Annals of the New York Academy of Sciences, Vol. 1104, pp. 135-146). New York: New York Academy of Sciences.

PREUSCHOFF, K., BOSSAERTS, P., & QUARTZ, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, **51**, 381-390.

PUTERMAN, M. L. (2001). Dynamic programming. In R. A. Meyers (Ed.), *Encyclopedia of physical science and technology* (3rd ed., Vol. 4, pp. 673-696). San Diego: Academic Press.

PUTERMAN, M. L. (2005). *Markov decision processes: Discrete stochastic dynamic programming*. Hoboken, NJ: Wiley-Interscience.

RACHLIN, H., RAINERI, A., & CROSS, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior*, **55**, 233-244.

RAMM, P., BENINGER, R. J., & FROST, B. J. (1984). Functional activity in the lateral habenular and dorsal raphe nuclei following administration of several dopamine receptor antagonists. *Canadian Journal of Physiology & Pharmacology*, **62**, 1530-1533.

REDISH, A. D., & JOHNSON, A. (2007). A computational model of craving and obsession. In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), *Reward and decision making in corticobasal ganglia networks* (Annals of the New York Academy of Sciences, Vol. 1104, pp. 324-339). New York: New York Academy of Sciences.

REISINE, T. D., SOUBRIÉ, P., ARTAUD, F., & GLOWINSKI, J. (1982). Involvement of lateral habenula-dorsal raphe neurons in the differential regulation of striatal and nigral serotonergic transmission in cats. *Journal of Neuroscience*, **2**, 1062-1071.

REMPEL-CLOWER, N. L. (2007). Role of orbitofrontal cortex connections in emotion. In G. Schoenbaum, J. A. Gottfried, E. A. Murray, & S. J. Ramus (Eds.), *Linking affect to action: Critical contributions of the orbitofrontal cortex* (Annals of the New York Academy of Sciences, Vol. 1121, pp. 72-86). New York: New York Academy of Sciences.

RESCORLA, R. A., & WAGNER, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

REYNOLDS, J. N., & WICKENS, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, **15**, 507-521. doi:10.1016/S0893-6080(02)00045-X

RICHARDS, J. B., MITCHELL, S. H., DE WIT, H., & SEIDEN, L. S. (1997). Determination of discount functions in rats with an adjusting-amount procedure. *Journal of the Experimental Analysis of Behavior*, **67**, 353-366.

RODRIGUEZ, P. F., ARON, A. R., & POLDRACK, R. A. (2006). Ventral-striatal/nucleus-accumbens sensitivity to prediction errors during classification learning. *Human Brain Mapping*, **27**, 306-313.

SAMUELSON, P. (1937). A note on measurement of utility. *Review of Economic Studies*, **4**, 155-161.

SANTAMARIA, J. C., SUTTON, R. S., & RAM, A. (1998). Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive Behavior*, **6**, 163-218.

SCHOENBAUM, G., CHIBA, A. A., & GALLAGHER, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, **1**, 155-159.

SCHOENBAUM, G., & ROESCH, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron*, **47**, 633-636.

SCHÖNBERG, T., DAW, N. D., JOEL, D., & O'DOHERTY, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, **27**, 12860-12867.

SCHULTZ, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, **80**, 1-27.

SCHULTZ, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, **1**, 199-207.

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

SCHULTZ, W., APICELLA, P., SCARNATI, E., & LJUNGBERG, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, **12**, 4595-4610.

SCHULTZ, W., DAYAN, P., & MONTAGUE, P. R. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593-1599.

SCHULTZ, W., & DICKINSON, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, **23**, 473-500.

SCHULTZ, W., PREUSCHOFF, K., CAMERER, C., HSU, M., FIORILLO, C. D., TOBLER, P. N., & BOSSAERTS, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B*, **363**, 3801-3811. doi:10.1098/rstb.2008.0152

SCHULTZ, W., & ROMO, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, **57**, 201-217.

SCHULTZ, W., TREMBLAY, L., & HOLLERMAN, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, **10**, 272-283.

SCHWEIMER, J. V., BRIERLEY, D. I., & UNGLESS, M. A. (2008). Phasic nociceptive responses in dorsal raphe serotonin neurons. *Fundamental & Clinical Pharmacology*, **22**, 119.

SETLOW, B., SCHOENBAUM, G., & GALLAGHER, M. (2003). Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron*, **38**, 625-636.

SHOHAMY, D., MYERS, C. E., GROSSMAN, S., SAGE, J., GLUCK, M. A., & POLDRACK, R. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain*, **127**, 851-859.

SIMMONS, J. M., RAVEL, S., SHIDARA, M., & RICHMOND, B. J. (2007). A comparison of reward-contingent neuronal activity in monkey orbitofrontal cortex and ventral striatum: Guiding actions toward rewards. In G. Schoenbaum, J. A. Gottfried, E. A. Murray, & S. J. Ramus (Eds.), *Linking affect to action: Critical contributions of the orbitofrontal cortex* (Annals of the New York Academy of Sciences, Vol. 1121, pp. 376-394). New York: New York Academy of Sciences.

SMART, W. D., & KAELBLING, L. P. (2000). Practical reinforcement learning in continuous spaces. In *Proceedings of the 17th International Conference on Machine Learning* (pp. 903-910). San Francisco: Morgan Kaufmann.

SOZOU, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society B*, **265**, 2015-2020.

STERN, W. C., JOHNSON, A., BRONZINO, J. D., & MORGANE, P. J. (1979). Effects of electrical stimulation of the lateral habenula on single-unit activity of raphe neurons. *Experimental Neurology*, **65**, 326-342.

STEVENS, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153-181.

SURI, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks*, **15**, 523-533.

SURI, R. E., & SCHULTZ, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, **91**, 871-890.

SUTTON, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9-44.

SUTTON, R. S., & BARTO, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. R. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497-537). Cambridge, MA: MIT Press.

SUTTON, R. S., & BARTO, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

SUTTON, R. S., PRECUP, D., & SINGH, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, **112**, 181-211.

TAN, C. O., & BULLOCK, D. (2008). A local circuit model of learned striatal and dopamine cell responses under probabilistic schedules of reward. *Journal of Neuroscience*, **28**, 10062-10074.

THIÉBOT, M. H., HAMON, M., & SOUBRIÉ, P. (1983). The involvement of nigral serotonin innervation in the control of punishment-induced behavioral inhibition in rats. *Pharmacology, Biochemistry & Behavior*, **19**, 225-229.

THORNDIKE, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplements*, **2**(4, Whole No. 8).

TOBLER, P. N., CHRISTOPOULOS, G. I., O'DOHERTY, J. P., DOLAN, R. J., & SCHULTZ, W. (2008). Neuronal distortions of reward probability without choice. *Journal of Neuroscience*, **28**, 11703-11711.

TOBLER, P. N., FIORILLO, C. D., & SCHULTZ, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, **307**, 1642-1645.

TOBLER, P. N., O'DOHERTY, J. P., DOLAN, R. J., & SCHULTZ, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology*, **97**, 1621-1632.

TOLMAN, E. C. (1932). *Purposive behavior in animals and men*. New York: Appleton Century.

TREMBLAY, L., & SCHULTZ, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, **398**, 704-708.

TREMBLAY, L., & SCHULTZ, W. (2000). Reward-related neuronal activity during go–nogo task performance in primate orbitofrontal cortex. *Journal of Neurophysiology*, **83**, 1864-1876.

TREPEL, C., FOX, C. R., & POLDRACK, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research*, **23**, 34-50.

TRICOMI, E. M., DELGADO, M. R., & FIEZ, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron*, **41**, 281-292.

TVERSKY, A., & KAHNEMAN, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk & Uncertainty*, **5**, 297-323.

TYE, N. C., EVERITT, B. J., & IVERSEN, S. D. (1977). 5-Hydroxytryptamine and punishment. *Nature*, **268**, 741-743.

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., & MORGENSTERN, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

WAN, X., & PEOPLES, L. L. (2006). Firing patterns of accumbal neurons during a Pavlovian-conditioned approach task. *Journal of Neurophysiology*, **96**, 652-660.

WANG, R. Y., & AGHAJANIAN, G. K. (1977). Physiological evidence for habenula as major link between forebrain and midbrain raphe. *Science*, **197**, 89-91.

WHITE, N. M., & HIROI, N. (1998). Preferential localization of self-stimulation sites in striosomes/patches in the rat striatum. *Proceedings of the National Academy of Sciences*, **95**, 6486-6491.

WICKENS, J. R., BUDD, C. S., HYLAND, B. I., & ARBUTHNOTT, G. W. (2007). Striatal contributions to reward and decision making: Making sense of regional variations in a reiterated processing matrix. In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), *Reward and decision making in corticobasal ganglia networks* (Annals of the New York Academy of Sciences, Vol. 1104, pp. 192-212). New York: New York Academy of Sciences.

WILKINSON, L. O., & JACOBS, B. L. (1988). Lack of response of serotonergic neurons in the dorsal raphe nucleus of freely moving cats to stressful stimuli. *Experimental Neurology*, **101**, 445-457.

WITTEN, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. *Information & Control*, **34**, 286-295.

WOOTEN, G. F., & COLLINS, R. C. (1981). Metabolic effects of unilateral lesion of the substantia nigra. *Journal of Neuroscience*, **1**, 285-291.

YANG, L.-M., HU, B., XIA, Y.-H., ZHANG, B.-L., & ZHAO, H. (2008). Lateral habenula lesions improve the behavioral response in depressed rats via increasing the serotonin level in dorsal raphe nucleus. *Behavioural Brain Research*, **188**, 84-90.

YIN, H. H., & KNOWLTON, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, **7**, 464-476.

ZALD, D. H., & KIM, S. W. (2001). The orbitofrontal cortex. In S. P. Salloway, P. F. Malloy, & J. D. Duffy (Eds.), *The frontal lobes and neuropsychiatric illness* (pp. 33-69). Washington, DC: American Psychiatric Publishing.

NOTES

1. Some form of future discounting is necessary mathematically. Without such discounting, the sum would rarely converge, because we are summing an infinite number of terms. In so-called "episodic tasks," however, which terminate after a finite number of steps, the discount factor is not mathematically necessary, because the sum consists of a finite number of terms.

2. Here, I am assuming that the reinforcement for an action taken at time t is given at time t; r_t is therefore the reinforcement for action a_t taken in state s_t. Sometimes it is assumed instead that the reinforcement for the action taken at time t is given at time $t + 1$. This is simply a matter of notation.

3. This is more general than it might seem because, when necessary, relevant information about the system's history can be encapsulated in the states.

4. The initial value tends to matter only early in learning. As one gets more samples, the influence of the initial value on the current estimate diminishes rapidly because of the exponential discounting of early influences.

5. This is true only with localist state representations. More generally, the weight-update rule for the weight w_s from state-unit s to the value-prediction unit is $w_s \leftarrow w_s + \alpha\delta x_s$. With localist state representations, $x_s = 1$ if s was the state at the previous time step and $x_s = 0$ otherwise, so this equation reduces to Equation 5.

6. This is why lateral inhibition is indicated in the actor layer in Figure 2.

7. As in the case of the critic, this is true only with localist state representations. More generally, the weight-update rule for the weight w_{sa} from state-unit s to the action-unit a corresponding to the selected action is $w_{sa} \leftarrow w_{sa} + \beta\delta x_s$. With localist state representations, $x_s = 1$ if s was the state at the previous time step and $x_s = 0$ otherwise, so this equation reduces to Equation 11.

8. It seems unlikely that self-stimulation could have produced this effect by electrically stimulating the actor. For that, the stimulation would have to be delivered precisely in the part of the striatum responsible for the response being learned, which seems unlikely to occur by chance.

9. Interestingly, the electrodes in the self-stimulation experiment mentioned above tended to be concentrated in anterior, ventral, and medial portions of the striatum (White & Hiroi, 1998). If the electrodes had instead been concentrated in the dorsolateral striatum, self-stimulation of dorsolateral striosomes conceivably could have failed to reinforce the response.

10. In expected utility theory, the utility of an outcome is a nonlinear function of its magnitude; expected utility is the expected value of the utilities of probabilistic outcomes. As noted above, in reinforcement learning primary reinforcements typically are not passed through a nonlinear function to obtain their utility, but doing so would be straightforward.

11. The idea that recency effects may produce underweighting of small probabilities and overweighting of large probabilities has also been proposed in behavioral economics (Hertwig, Barron, Weber, & Erev, 2004).

12. Strangely, humans exhibit the reverse pattern when outcomes are described verbally, as is the case in most behavioral economics experiments (Kahneman & Tversky, 1979; Trepel et al., 2005; Tversky & Kahneman, 1992).

13. In economics, *risk* refers to the variance of outcomes when the probabilities are known; *uncertainty* and *ambiguity*, in contrast, are often used when the probabilities are unknown or are not well defined (Knight, 1921). In the electrophysiological and neuroimaging literature, *uncertainty* is sometimes used to refer to variance due to probabilities that the subject knows because of extensive prior training (Fiorillo et al., 2003; Tobler, O'Doherty, Dolan, & Schultz, 2007). I refer to such variance as *risk*, for consistency with the economics literature.

14. Of course, we should not forget the old adage "When all you have is a hammer, everything looks like a nail." History is rich in what, from our present-day perspective, seem amusingly naive attempts to explain the functioning of the brain in terms of the technological advances of the time (Kirkland, 2002).

15. Reinforcement learning is not always fully normative because it also takes into consideration computational tractability and efficiency. However, it is based on normative assumptions (e.g., maximization of expected value, exponential discounting).

SUPPLEMENTAL MATERIALS

Additional information about (1) learning Markov decision processes, (2) determining their value functions, and (3) the mathematics of temporal discounting may be downloaded from http://cabn.psychonomic-journals.org/content/supplemental.

(Manuscript received December 6, 2008;
revision accepted for publication July 23, 2009.)