48 Dopamine Reward Prediction Errors: The Interplay between Experiments and Theory

CLARA K. STARKWEATHER AND NAOSHIGE UCHIDA

ABSTRACT Reinforcement-learning theories provide a normative perspective on learning and decision-making. In the 1990s, neurophysiology experiments revealed an exceptional correspondence between the activity of midbrain dopamine neurons and the reward prediction error (RPE) signal used to train computers in a reinforcement-learning algorithm called *temporal difference* (TD) learning. Studies of midbrain dopamine neurons play a pivotal role at the interface of empirical and theoretical studies. A theoretical framework for reinforcement learning has facilitated the interpretation of neurophysiology data and has guided the design of future studies. Here we discuss recent developments in the interplay between experimental findings and theories of dopamine signaling. In particular, recent studies emphasize the importance of state uncertainty in the neurobiological implementation of reinforcement learning.

An inexperienced shopper brings home a variegated bag of plums. After finding that the green and red plums are sour and the darker plums are sweet, she develops an eye for deep purple plums. She selects only ripe plums during future grocery trips. Animals learn to predict the value of outcomes in order to adapt successfully to their environments. The process of trial and error—and, crucially, learning from errors—reinforces the pairing of value with otherwise neutral sensory stimuli.

This chapter covers models and neural mechanisms of reinforcement learning. We will focus on the midbrain dopaminergic system, thought to be at the center of reinforcement learning in the mammalian brain. Research on the dopamine system occupies a unique space at the interface of empirical and theoretical studies. A theoretical framework for reinforcement learning provides twofold utility: it has aided in the interpretation of neurophysiology data and guides the design of future studies. In this spirit, we will frame this chapter using these normative models.

Pioneering Models and Neural Correlates of Reinforcement Learning

Certain outcomes (*unconditioned stimuli*, or US) trigger a biological response in the absence of learning, such as salivation accompanying a bite of tasty food. During Pavlovian conditioning, animals learn to predict a US, following a neutral sensory stimulus that does not evoke an automatic biological response (*conditioned stimulus*, or CS). Evidence of this prediction is seen in the animal's level of conditioned responding. Classical-conditioning tasks elicit conditioned responding that matches the automatic reaction to the anticipated US. The level of conditioned responding is measured during the time between the CS and US (the *interstimulus interval*, or ISI) and gradually increases across trials of learning.

An early theory describing the formation of a CS-US association (Rescorla & Wagner, 1972) set the change in associative strength between the CS and the US ($\Delta\hat{V}$) directly proportional to the discrepancy between the actual US level (V) and the predicted US level ($\hat{V}$):

$$\Delta\hat{V} = \alpha(V - \hat{V}) \tag{48.1}$$

In equation 48.1, α is a constant with a value between 0 and 1 that reflects the animal's learning rate. Prior to learning, $\hat{V}$ is 0. Thus, $(V - \hat{V})$ is large once an unexpected reward is received, driving large changes in $\hat{V}$ during initial CS-US pairings. As $\hat{V}$ increases over trials, smaller incremental changes in $\hat{V}$ occur on each subsequent trial. No additional associative strength is gained unless an existing prediction is violated. The Rescorla-Wagner model formalized the idea that an error between the actual and predicted outcome $(V - \hat{V})$ is needed in order to drive learning.

Subsequently, a learning theory was born in computer science (Sutton, 1988; Sutton & Barto, 1990). This theory, called temporal difference (TD) learning, also used the discrepancy between actual and expected outcome to drive learning, similar to the Rescorla-Wagner

model. The model learns values, formally defined as the discounted sum of future reward (Sutton & Barto, 1990) associated with each "state" (figure 48.1*A*):

$$V(s_t) = \sum_{\tau=t}^{\infty} \gamma^{\tau-t}\, r(\tau) \tag{48.2}$$

$$= r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \cdots$$

where s_t is the state at time t, t is the current time, $r(\tau)$ is the reward at time τ, and γ is a discount factor ($0 \leq \gamma \leq 1$) that down weights future rewards. States represent the environment, or *task space*, in a way that is useful to performing the current task and here represent the world at different moments in time following some observable stimuli (although these may be used in other ways, as seen later in this chapter). States s, and transitions between these states, form a Markov process, allowing equation 48.2 to be written recursively using the Bellman equation:

$$V(s_t) = r(t) + \gamma V(s_{t+1}). \tag{48.3}$$

After transitioning from the state t to the state $t+1$, the agent obtains new feedback, $r(t)$. Both sides of equation 48.3 are the predicted value for s_t—that is, $\hat{V}(s_t)$. Due to the new information ($r(t)$) obtained by taking the state transition, the right-hand side can be seen as a more accurate prediction. The discrepancy between the left-hand side and the right-hand side is called the TD error because it corresponds to the discrepancy between the predictions at consecutive time points:

$$\delta(t) = r(t) + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t), \tag{48.4}$$

where $\delta(t)$ is the TD error at time t. The model then updates the predicted value for state s_t according to the following update rule:

$$\hat{V}(s_t) \leftarrow \hat{V}(s_t) + \alpha\delta(t), \tag{48.5}$$

where α is the learning rate. We will illustrate how TD learning would operate for a CS-US pairing, with a constant time delay. Before learning, the TD model has not learned any values. When the model experiences US (reward) delivery at time t following CS onset, this produces a positive prediction error $\delta(t)$. This positive prediction error increases the value associated with the corresponding state s_t, resulting in a positive value estimate at time t. During the subsequent CS-US pairing, there is a positive TD error at time $t-1$ because $\gamma \hat{V}(s_{t+1})$ is now positive, producing a positive prediction error $\delta(t-1)$ and increasing the value associated with the previous state s_{t-1}. In this way, the positive prediction errors propagate back in time over trials, increasing the values of states that fully tile the ISI, until the only positive prediction error occurs at the time of the CS (if CS onset is unpredictable). The TD algorithm eventually learns a sustained value estimate that precisely spans the time between the CS and US (figure 48.1*B*).

Although TD learning was originally developed for computer science applications, neurophysiology experiments in the 1990s by Schultz and colleagues revealed an exceptional correspondence between the error signals predicted by TD learning and those signaled by midbrain dopamine neurons (Bayer & Glimcher, 2005; Hollerman & Schultz, 1998; Mirenowicz & Schultz, 1994; Schultz, Dayan, & Montague, 1997; figure 48.1*B, C*). The authors trained monkeys on a classical-conditioning task in which a CS was followed by a US, roughly 1 s later. Before learning, dopamine neurons showed phasic activation only at the time of the US. After learning, this phasic dopamine response instead occurred at the time of the CS, and the signal at the time of the US was much smaller than before learning. Strikingly, if the US was unexpectedly omitted, dopamine neurons briefly paused their tonic firing exactly at the time of the usual US delivery. These signals, collectively referred to as reward prediction error (RPE), match the error signals proposed by the TD-learning algorithm (figure 48.1*B, C*):

(1) Before learning, the only positive TD error $\delta(t)$ is at the time of the US due to positive excitation from reward $r(t)$.
(2) After learning, the sustained value signal commences following the CS, resulting in a positive TD error and therefore a positive TD error at CS onset.
(3) After learning, because the sustained value prediction drops to zero at the time of the predicted US, the TD error is negative at the time of the US and cancels out positive excitation from reward $r(t)$, resulting in a smaller response at the time of a predicted reward.
(4) After learning, if the US is not delivered at the predicted time, the negative TD error is not canceled out by the arrival of a reward, resulting in a negative signal.

In these ways, dopamine signals recapitulate TD error signals. While the Rescorla-Wagner model predicts a smaller error signal at the time of a predicted US, it lacks a timing mechanism within a trial and therefore would not necessarily capture the CS response or the time-locked negative signal observed on reward omission with a nonzero CS-US delay.

The idea that dopamine neurons convey TD error signals has been substantiated by further experiments in a variety of species, including rats (Flagel et al., 2011; Hart, Rutledge, Glimcher, & Phillips, 2014; Pan, Schmidt, Wickens, & Hyland, 2005; Roesch, Calu, & Schoenbaum, 2007; Stuber et al., 2008), mice (Cohen, Haesler, Vong, Lowell, & Uchida, 2012; Eshel et al., 2015; Eshel, Tian,

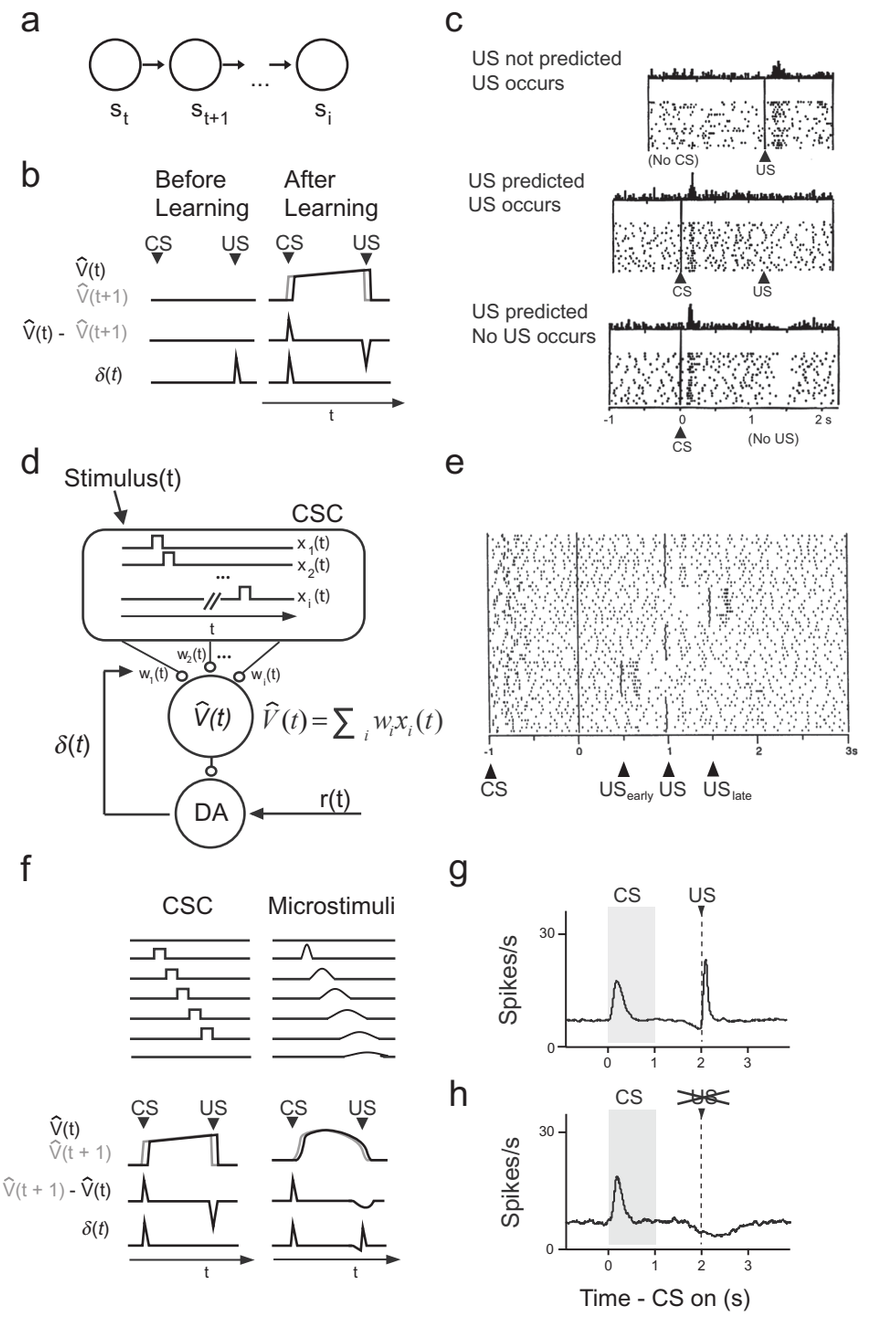

FIGURE 48.1 Temporal difference (TD) learning and dopamine signaling. *A*, States can be modeled as each corresponding to an arbitrary unit of time. *B*, Value signal, TD of value signal, and error signal produced by a simple TD model. *C*, Firing pattern of a putative dopaminergic neuron during a classical-conditioning task. From Schultz, Dayan, and Montague (1997). *D*, Neural circuit hypothesis for implementing TD learning in the brain, using the complete serial compound (CSC) feature representation. Value is computed as the linear sum of features (each "CSC" component, indexed from 1–i) multiplied by their weights. *E*, The timing of US delivery following CS onset was jittered (either earlier or later) by 500 ms in probe trials. Note there was no reward omission response if US was given early. From Hollerman and Schultz (1998). *F*, CSC versus microstimulus features. *G*, Averaged data from dopamine neurons showed a small dip in the tonic rate of dopamine firing prior to reward delivery. From Tian and Uchida (2015). *H*, Averaged data from dopamine neurons showed a temporally spread dip upon reward omission that lasted roughly 1 s. From Tian and Uchida (2015).

—-1
—0
—+1

Bukwich, & Uchida, 2016; Menegas, Babayan, Uchida, & Watabe-Uchida, 2017; Parker et al., 2016), and humans (D'Ardenne, McClure, Nystrom, & Cohen, 2008). Importantly, recent studies applied optogenetic methods to unambiguously identify dopamine neurons while recording both in mice (Cohen et al., 2012; Eshel et al., 2015, 2016) and in monkeys (Stauffer et al., 2016). Eshel et al. (2015, 2016) demonstrated that reward expectation reduces dopamine reward responses in a purely subtractive fashion, and dopamine neurons in the lateral ventral tegmental area (VTA) exhibit very similar response properties, indicating that these neurons signal TD errors homogeneously. The correspondence between TD errors and dopamine signals supports the hypothesis that TD learning approximates how reinforcement learning is actually implemented in the brain. This hypothesis is further supported by optogenetic stimulation experiments, which show that the temporally precise manipulation of dopamine signaling at the time of expected outcomes leads to increased behavioral responding to preceding CSs, just as TD learning would predict (Steinberg et al., 2013).

A classic adaptation of TD learning to a biological circuit model utilizes a *complete serial compound* (CSC) feature representation that tracks elapsed time relative to observable stimuli (Schultz, Dayan, & Montague, 1997). Each CSC feature $x_i(t)$ is a vector of zeros, except for a value of 1 at timepoint i relative to cue onset (figure 48.1*B*). Each CSC feature can be conceived as representing the occupancy of a single state s described in the example above. $x_1(t)$ would have a value of 1 at time point 1, corresponding to occupancy of state s_1. A neuron corresponding to $x_1(t)$ would only fire at time point 1. A population of analogous neurons could show sequential activations and represent the entire interval between the CS and US. The value function estimate is modeled as a linear combination of these features (figure 48.1*D*):

$$\hat{V}(t) = \sum_i w_i x_i(t), \tag{48.6}$$

where $x(t)$ represents CSC features, and w_i is a predictive weight associated with feature i. The weights are updated according to the following learning rule:

$$\Delta w_i = \alpha x_i(t)\, \delta(t), \tag{48.7}$$

where α is a learning rate ($0 \le \alpha \le 1$), and $\delta(t)$ is the prediction error in the value signal. The discrepancy between actual and predicted value is computed similar to that above (48.4), according to

$$\delta(t) = r(t) + \gamma \hat{V}(t+1) - \hat{V}(t), \tag{48.8}$$

where $r(t)$ represents reward at time t. The TD error increases the weights of sequential neural activations (modeled as CSC features) such that a downstream area can compute an accurate value prediction. In line with this algorithmic hypothesis, other theoretical works speculated that the cortex conveys the CSC temporal features into the striatum, where value is computed (Houk, Adams, & Barto, 1995). By modulating the weights of corticostriatal synapses, dopamine signals could plausibly shape the striatal value representation. The TD-learning model provided a cornerstone for understanding how reinforcement learning could be biologically implemented.

Timing in Temporal Difference Models

The CSC TD model captures the exquisite temporal specificity of dopamine RPEs. First, dopamine neurons show a negative prediction error exactly at the time of an expected reward. Second, dopamine responses to reward are suppressed only at the time of the expected reward. In a 1998 study, Hollerman and Schultz (1998) occasionally shifted the timing of reward by just 500 ms earlier or later than the time of the usual reward delivery (figure 48.1*E*). This small temporal jitter evoked a larger dopamine response than if the reward were delivered at its usual time. This finding is consistent with the errors produced by the TD model because the temporal difference of the value signal is most negative at the cue-reward delay time that the animal was trained on (thereby canceling excitation from reward only at that time; see figure 48.1*B*). Another timing-related characteristic of dopamine RPEs, captured by the TD model, is delay discounting. Because value corresponds to the discounted sum of future rewards (equation 48.2), the magnitude of the value signal at cue onset is inversely related to the delay between cue and reward. Consistent with this, several studies in both rodents and primates have observed delay discounting in the cue response of dopamine neurons (Fiorillo, Newsome, & Schultz, 2008; Kobayashi & Schultz, 2008; Roesch, Calu, & Schoenbaum, 2007; Starkweather, Babayan, Uchida, & Gershman, 2017). Cues followed by late rewards result in smaller dopamine responses than cues followed by early rewards. Thus, the TD framework matches several findings relating to timing and the dopamine system.

However, additional experimental observations suggest that the TD model—particularly, the CSC implementation—does not provide a complete account of timing in the dopamine system. The first observation is that dopamine responses are not thoroughly suppressed at the time of expected rewards, particularly for CS-US pairings involving long ISIs (Fiorillo et al., 2008; Kobayashi & Schultz, 2008). A CSC TD model would suppress the US responses equally irrespective of

-1—
0—
+1—

the delay time. The second observation is that the "dip" observed upon reward omission is temporally extended for up to 1 s (Cohen et al., 2012; Matsumoto & Hikosaka, 2007; Schultz, Dayan, & Montague, 1997; Tian & Uchida, 2015). In contrast, a CSC TD model produces a sharp dip only at the time of the expected reward (figure 48.1*B*).

Based on these experimental observations, Ludvig, Sutton, and Kehoe (2008) proposed a TD model that swaps the CSC representation for Gaussian distributions, whose widths increase over elapsed time, called *microstimuli* (figure 48.1*F*). The increasing widths of the microstimuli respect Weber's law, which stipulates that the variance of timing estimation increases linearly with elapsed time (Balsam & Gallistel, 2009). If a reward occurs at a long delay from the CS onset, the reward increases the weights of more than one microstimulus feature (contrasting with the CSC TD model) because the Gaussian distribution of multiple features will have nonzero values at the moment a reward is received. Accordingly, the TD model increases the weights of multiple microstimulus features that are active at that time. Over many trials this produces a value signal that is less sharply resolved in time than the CSC TD model (figure 48.1*F*). Specifically, the value function will be positive at more time points than just the exact moment that reward is received, and it will have a smaller amplitude (later microstimuli are modeled as having smaller heights). Therefore, the TD error, $\gamma\widehat{V}(t+1)-\widehat{V}(t)$, will be shallower and more spread out over time. As a consequence, when a reward is given, the response cannot be completely canceled out. The reward response is greater for longer delays because the model's ability to cancel out excitation from the reward becomes less precise for longer intervals. Furthermore, if a predicted reward is omitted, the "dip" is more spread out across time than predicted by the TD model with the CSC (figure 48.1*H*). Finally, a common observation from our lab is that the baseline tonic firing rate of dopamine neurons decreases slightly before a reward is received (figure 48.1*G*; Starkweather et al., 2017; Tian et al., 2016; Tian & Uchida, 2015). This prereward decrease in firing is reproduced by the microstimulus model because the temporal imprecision results in a negative prediction error commencing prior to the exact time of the reward. Therefore, the microstimulus TD model explains common observations of dopamine recordings, which were inconsistent with the CSC.

Other updates to the TD model have been proposed, which we will discuss in the next section. However, it is important to keep in mind that all these models possess a timekeeping mechanism. Dopamine RPEs are likely computed on an imperfect and noisy timing mechanism. Therefore, any modeling refinements proposed in future work should also contain a timing representation constrained by scalar timing uncertainty.

State Inference in the Temporal Difference Model

In 1998 Hollerman and Schultz made another important experimental observation in monkeys trained on a CS-US pairing separated by a constant delay time. In a small proportion of probe trials, rewards were given earlier than predicted. As predicted by the TD model, this early reward produced a large dopamine response because the timing of the reward was unexpected. However, there was no omission response (i.e., negative RPE) at the time of the usual reward (figure 48.1*E*). In contrast, the CSC TD model produced an omission response at the time of the usual reward (figure 48.2*A*). Based on this experimental observation, two main categories of modifications were proposed to the TD model.

The first proposal was that the TD model simply "resets" after a reward is received. Future errors could be fixed at zero after a reward is received (Suri & Schultz, 1998, 1999). Alternatively, a reward receipt could terminate the set of CSC features activated during the ISI and trigger a new set of CSC features activated during the intertrial interval, or ITI (Brown, Bullock, & Grossberg, 1999). Implicit in this reset model is that the animal knows (upon reward) that a reward is no longer expected. This means that the animal infers the "state" of the task based on observable stimuli (the reward). If one imagines the task is divided into two states—the ISI state during which a reward is expected and the ITI state during which a reward is not expected—the receipt of a reward initiates CSC features that correspond to the ITI state and terminates those that correspond to the ISI state. However, this is an ad hoc modification to the TD model with the CSC. It adds a simple form of state inference to the TD model to match the data but does not acknowledge other instances in which state inference comes into play. For example, what if reward were omitted? Would the ISI features simply continue on indefinitely because no reward is received? Or would the animal infer over time that a reward is not coming, thereby terminating the ISI features? These issues remain unresolved with the reset model.

The second proposed modification is that the TD model's features themselves represent the inferred state of the environment (Daw, Courville, & Touretzky, 2006; Rao, 2010). We will refer to this as a *belief-state* TD model. The temporal CSC or microstimulus representation is swapped with a belief state, which is an inferred probability distribution over possible states. Value is no longer proportional to the weight assigned to a particular active state. Rather, value is equal to the weight assigned

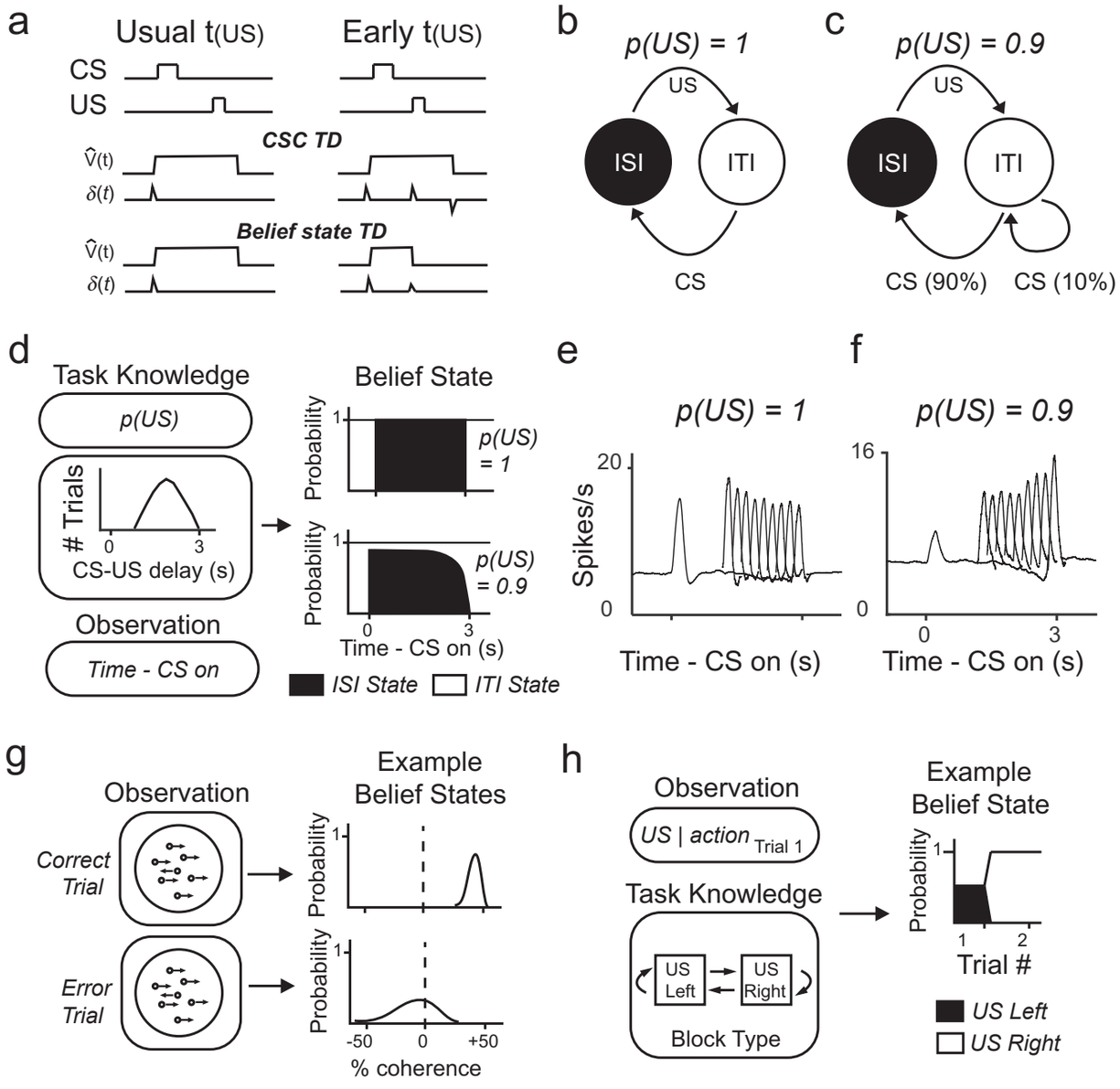

FIGURE 48.2 Belief-state features for reinforcement learning. *A*, Schematic for values and error signals produced by early reward delivery, in the belief-state TD model versus CSC TD model. CSC TD model produces a spurious reward omission "dip" after an early reward delivery, which is not observed in the data shown in figure 48.1*E*. *B*, Semi-Markov schematic for computing the belief state under fully observable task conditions. Adapted from Daw et al. (2006). *C*, Semi-Markov schematic for computing the belief state under partially observable task conditions. Adapted from Daw et al. (2006). *D*, In a task invoking variable delay times and particular reward contingencies (illustrated for 100%—and 90%–rewarded contingencies), the agent uses its observation of time passed since cue onset in addition to its knowledge of reward timing/contingency to compute a probability distribution over possible states from figure 48.2*B, C*. This probability distribution is called a *belief state*. *E*, Dopamine neurons produced smaller error signals as a function of time when reward is delivered if the task illustrated in figure 48.2*D* is 100% rewarded. *F*, Dopamine neurons produced larger error signals as a function of time when reward is delivered if the task illustrated in figure 48.2*D* is 90% rewarded. This can be explained by taking the belief state into account—as belief favors the possibility of a reward omission trial at later time points, reward evokes a larger error signal if it is actually delivered. *G*, Based on ambiguous sensory stimuli (random dot motion), the agent may compute a belief state over possible coherences and direction of movement. Adapted from Lak et al. (2017). *H* Based on observation of the first trial type within a block, an agent may use its knowledge of the task structure to compute a belief state. Schematized based on Bromberg-Martin et al. (2010).

to a state, scaled by the probability (from the belief state) allotted to that state, summed over all states:

$$\hat{V}(t) = \sum_i w_i b_i(t), \tag{48.9}$$

where $b(t)$ represents the belief state (i.e., the probability of being in each state i) at time t, with i indexing individual states, and w_i is a predictive weight associated with state i. Furthermore, weights are updated proportionally to the probability with which the agent "believes" it occupies a particular state:

$$\Delta w_i = \alpha b_i(t)\, \delta(t) \tag{48.10}$$

The belief-state TD model was originally conceived as a semi-Markov process involving just two states: the ISI and the ITI (figure 48.2*B*). Semi-Markov dynamics imply that the time spent in a state is probabilistic and is defined by a probability distribution called a *dwell time distribution* (contrasting with a Markov process, where each "state" represents one arbitrary unit of time). For that reason, rewards are discounted by however much time elapses within a state once received because this time interval varies from task to task: prediction errors should be larger, therefore driving up the value of a particular state more if the reward arrives early, whereas the opposite should be true if the reward arrives late.

The belief-state TD model accounts for the Hollerman and Schultz result and captures other experimental findings (Daw, Courville, & Touretzky, 2006). The Hollerman and Schultz experiment was rewarded in 100% of trials, meaning that the task is fully observable because the state of the task (ISI or ITI) can be deciphered based on sensory cues alone. Upon observing the CS, the belief state allots 100% probability into the ISI state and 0% probability into the ITI state. Upon observing the US, the belief state allots 100% probability into the ITI state and 0% probability into the ISI state (figure 48.2*D*). If the ISI accrues a larger weight (as it should, because rewards are only received when the belief state favors the ISI), a transition into the ITI upon receiving the US would eliminate future reward expectation. This would abolish the spurious reward omission response reported by the CSC TD model upon receiving an early reward (figure 48.2*A*). A second set of experimental results, compatible with the belief-state TD model, consists of dopamine responses recorded in 100%-rewarded task contingencies with variable delay times. In these experiments, on any given trial, the delay between the CS and US (or a reward-predicting CS) was drawn from a uniform distribution (Fiorillo et al., 2008; Nomoto, Schultz, Watanabe, & Sakagami, 2010; Pasquereau & Turner, 2015). Rewards delivered earliest in the variable delay interval produced the largest dopamine responses, and rewards delivered later in the interval produced smaller dopamine responses. While the authors argued that these results are explained by expectancy over time resembling a hazard function (the momentary likelihood that an event will occur, given that it hasn't occurred yet), this result is also compatible with the belief-state TD model (Starkweather et al., 2017). Because late rewards *within a particular state* are discounted more heavily in a semi-Markov framework, the belief-state TD model also captures this pattern of prediction errors across time. In contrast, a CSC TD model, which lacks an explicit representation of state space (ISI vs. ITI), would produce the same prediction error at each possible time of reward delivery because the uniform distribution of timings teaches the model to apply the same value prediction at each time point.

The belief-state TD model captures experimental findings from tasks that jitter the timing of reward relative to cue. However, these tasks do not implicate state uncertainty, which is a core tenant of the belief-state TD model. A simple modification that renders the task states partially observable is reducing the probability of reward from 100% to 90% (Starkweather et al., 2017; Starkweather, Gershman, & Uchida, 2018). Once the cue comes on, the animal no longer knows for certain whether it is in the ISI or the ITI. This is because in 10% of trials (the unrewarded trials), a cue onset leads to a hidden state transition from the ITI to the ITI (figure 48.2*C*). Whereas the belief state was uniform, following cue onset, under deterministic task conditions, the belief state evolves over time as the animal gathers evidence in favor of one hidden state over the other. After cue onset, the belief state is 90%–10%, with 90% allotted to the ISI. As time elapses and the animal does not receive a reward, the belief state shifts more probability into the ITI state, yielding to the possibility of a reward omission trial (figure 48.2*D*). Because the belief state shifts toward the ITI state, the value prediction is low, and thus, the prediction error is high if a reward is actually received at a later time point. This has been demonstrated experimentally and is a key prediction of the belief-state TD model: state uncertainty should dramatically affect how reward expectation evolves over time. In a 100%-rewarded deterministic scenario (figure 48.2*C*), later rewards evoke smaller prediction errors due to discounting over a lengthy dwell time, whereas in a 90%-rewarded nondeterministic scenario, later rewards evoke larger prediction errors due to the belief favoring the unrewarded state (figure 48.2*F*).

Experimental results in other tasks implicating state uncertainty are also explained by the belief-state TD model. One study used a belief-state TD model to

capture dopamine signals observed in primates during a perceptual decision-making task (Lak, Nomoto, Keramati, Sakagami, & Kepecs, 2017; figure 48.2*G*). In this task a random-dot motion stimulus with variable coherence was presented following a fixation cue. Based on the perceived direction of the random dot motion, the animal made an action that if correct resulted in reward. The belief state was modeled as a probability distribution over a range of motion directions and coherences. On average the belief state should be centered at the true motion direction and coherence. However, the belief state would be different from trial to trial, as it was assumed there was sampling noise in the perception of the stimulus. Action value was computed based on the belief state, with the model choosing the action on the higher-valued side of the left-versus-right decision boundary. The value prediction was proportional to the probability of receiving a reward on a particular trial, given the belief state and corresponding choice—equivalent to a confidence signal—and would produce a difference in dopamine signals between correct (higher confidence) and error (lower confidence) trials. In contrast, the CSC TD model does not have access to a distribution representing the uncertainty in perceptual stimuli and would therefore predict dopamine signals of similar magnitude for "correct" and "error" trials. The authors found that dopamine signals were well matched by a belief-state TD model.

While both of the studies discussed above invoke a belief state that is computed on every trial, a belief state could pertain to more global aspects of task structure involving a block of trials. In one study (figure 48.2*I*), the cued side (left vs. right) on which a reward was presented switched between blocks (Bromberg-Martin, Matsumoto, Hong, & Hikosaka, 2010). So, if a formerly rewarded side went unrewarded, this signaled a block change. On a subsequent trial, animals should infer that the formerly unrewarded side was now rewarded. This inferred switch was reflected in dopamine signals. Dopamine responses to the inferred rewarded cue were larger than in the previous block, indicating a higher value prediction even if the animal had not yet experienced a reward on the newly rewarded side. This finding cannot be explained by a simple cached-value system (*cached*, meaning that the value must be assigned to a state based on direct experience) because the dopamine response reflected the new inferred value without the animal directly experiencing the new cue-outcome pairing. However, it may be explained by a belief state that shifts to reflect the rewarded side of the new block (Costa, Tran, Turchi, & Averbeck, 2015; Fuhs & Touretzky, 2007; Hampton, Bossaerts, & O'Doherty, 2006). Finally, a recent study similarly showed that dopamine responses reflect the inferred value of cues (Babayan, Uchida, & Gershman, 2018). In this study, blocks consisted of identical trials of a cue followed either by "big" or "small" rewards. On rare blocks, rewards were of intermediate size. After experiencing one of these intermediate rewards, the dopamine signal at the time of the reward on the subsequent trial appeared to subtract either the value of the "big" or "small" reward from the value of the received intermediate reward, as if the value prediction conveyed to dopamine neurons reflected whether the current block was inferred to be either "big" or "small," the two states in this task.

The belief-state TD model is needed to account for a growing number of empirical observations. In addition, belief-state TD models solve other existing controversies. For instance, in the original CSC model the value signal and the TD error signal propagate backward from the time of reward to the time of cue only gradually during learning. Neurophysiology data have not supported this gradual shift (Menegas et al., 2017; Pan et al., 2005). These results have been taken as evidence against TD models. However, belief-state models do not require this gradual shift because they must learn whether a cue is associated with a reward to begin with (in theoretical terms, learning about the state space), separately from when exactly the reward will occur.

Model-Based Feature Representation for Reinforcement Learning

Model-free systems predict discounted future reward without an explicit model of the environment. For instance, the CSC TD model lacks knowledge that a reward signifies a state transition into the ITI, hence the spurious "reward omission" dip after an early reward was delivered. Knowledge of the transition structure between states, and of the corresponding observations and probabilities of triggering these state occupancies, constitute *model-based* information. Therefore, the feature representation for the belief-state TD model is model-based. However, the TD-learning rules, including the computation of the TD error and the update of "states" weights, remain model-free. This model-free component means that the belief-state TD model stores, or "caches," weights for particular states, to be deployed every time a particular state is occupied (meaning that the belief state allots a nonzero probability to that state). These weights must be learned through the direct experience of occupying certain states.

The belief-state TD model differs from fully model-based reinforcement learning, which uses a forward model of the environment to compute values for various states. Consider a maze in which an animal receives

different rewards at each possible exit (Niv, Joel, & Dayan, 2006). The animal may be hungry and therefore place greater value on food versus liquid rewards at the end of the maze. The animal simulates mental paths through the maze and computes values for states (each corresponding to various locations in the maze). States forming trajectories leading to food rewards would have higher values than other states perhaps leading to liquid rewards. If the animal was thirsty, a different set of state values would be simulated based on the higher utility of paths leading to liquid rewards. Importantly, the animal would not have to be trained while it was thirsty in order to use its model of the environment to reassign values for various states. This contrasts with the belief-state TD model, or any system that uses a model-free-learning rule, in which a state must be directly experienced in order to update its value (*cached value*). Work utilizing a two-step decision-making task has shown that humans, to varying degrees, combine model-based and model-free value computation (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Blood oxygen level-dependent signals in the ventral striatum, thought to correlate with dopamine activity, signal prediction errors in value predictions that reflect both model-based and model-free computation. Furthermore, while the computations that feed into the value prediction may involve a complex cognitive model of the environment, fully model-based accounts still endorse the idea that dopamine signals errors in value prediction.

Some recent studies argue that dopamine may not exclusively signal discrepancies in value prediction. In one experiment rats were trained to associate a cue with a reward. Characteristics of the reward (e.g., flavor of milk) were varied across blocks (Takahashi et al., 2017). Changes in milk flavor—even though rats preferred chocolate and vanilla equally—evoked a positive dopamine signal at the beginning of a block switch. In another experiment (Sharpe et al., 2017) an otherwise blocked sensory association between two cues could be "unblocked" by optogenetically activating dopamine neurons. These studies lead to the proposal that dopamine signals carry prediction errors along axes other than value (Langdon, Sharpe, Schoenbaum, & Niv, 2018). This differs from the belief-state TD model, which maintains the model-free computation of the TD error and update rules and places model-based, belief-state computations upstream of the value function. We chose to focus on the belief-state TD model, which contains model-free prediction errors, because it is widely accepted that dopamine signals are sensitive to errors in value prediction and drive learning about value. Furthermore, dopaminergic axons in the ventral striatum convey canonical model-free TD errors (Menegas et al., 2017; Parker et al., 2016). The unpredicted presentation of neutral stimuli caused no response from these axons (Menegas et al., 2017), incongruent with a pure "sensory" surprise signal. Furthermore, the activation or inactivation of dopamine signals exerts effects consistent with positive or negative RPEs (Chang et al., 2016; Steinberg et al., 2013). In these ways the model-free "value" axis of dopamine's role in learning is well established. By contrast, while it is possible that errors in sensory prediction (such as chocolate vs. vanilla milk) drive learning about state space, it is unclear how these errors could be separated along multiple axes by downstream circuitry. While recent work has shown that dopamine signals broadcast different types of information to subregions of the striatum (Howe & Dombeck, 2016; Menegas et al., 2017; Parker et al., 2016), it remains unknown whether dopamine signals containing state prediction errors can be separated downstream to drive learning about a world model. This is an important experimental hurdle that must be addressed before theories further integrate model-based prediction errors.

Implementing a Belief-State Temporal Difference Model

Neural implementations of a belief state may be modeled as partially observable Markov decision processes (POMDP; Daw et al., 2006; Lak et al., 2017; Rao, 2010; Starkweather et al., 2017). While the resulting belief state derived from a POMDP is equivalent to a belief state in the semi-Markov model, a benefit of a POMDP is that the belief state will propagate between nodes of the model as time elapses, potentially providing a closer analogy to neural activity. Rao (2010) harnessed this property to propose a simple implementation of a belief state using a recurrent neural network. The belief state is computed as follows:

$$b_i(t) \propto p(o(t)|i) \sum_j p(i|j)\, b_j(t-1), \qquad (48.11)$$

where $b_i(t)$ is the posterior probability that the animal is in substate i at time t, $p(o(t)|i)$ is the likelihood of the observation $o(t)$ under hypothetical substate i, and $p(i|j)$ is the probability of transitioning from substate j to substate i. It is possible to compute the belief state recursively from the belief state computed at $(t-1)$ because of the Markov property. In this way, a simple recurrent circuit that maintains feedback from the previous time point (in addition to incorporating feedforward information for new observations) could implement a belief state. We illustrated how the Starkweather et al. (2017) task could map onto this proposed implementation (figure 48.3*A*). Individual units of the feedforward layer would convey the likelihood of each observation,

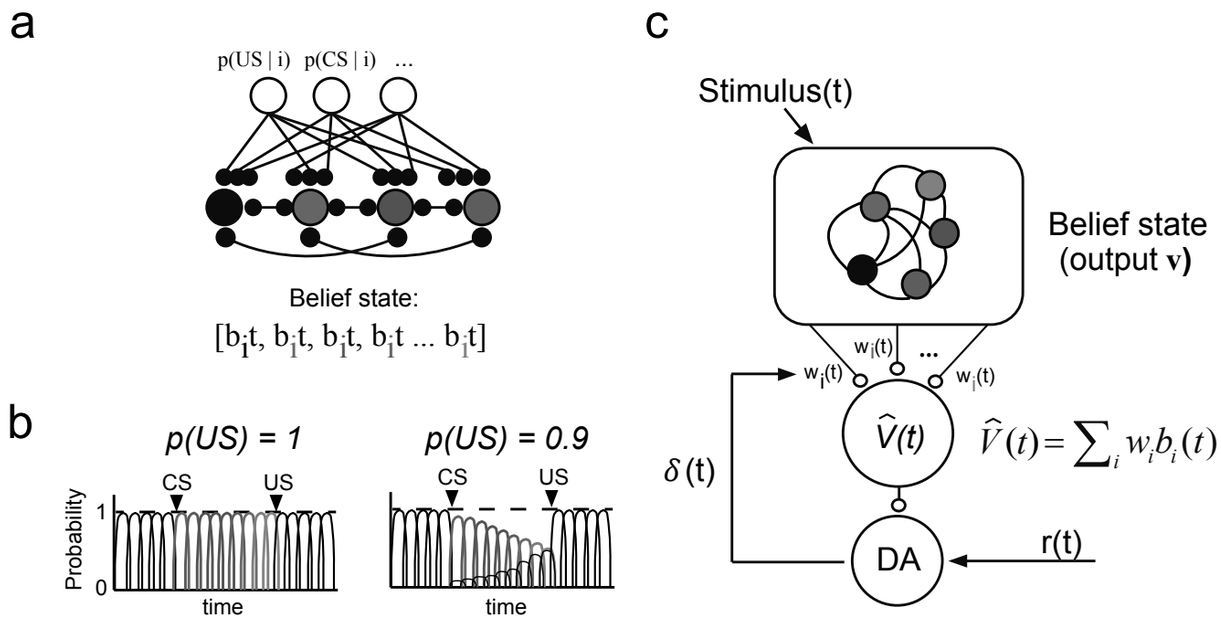

FIGURE 48.3 Recurrent network implementation of a belief state. *A*, A simple network with feedforward connections and a recurrently connected layer, adapted to computing a belief state for the task described in figure 48.2*D*. *B*, Predicted firing rates for output layer neurons in the task described in figure 48.2*A* if these neurons each signal one value within the belief-state vector. *C*, Revised neural circuit hypothesis of TD-learning implementation. Belief-state features could be represented by a recurrent network. The outputs from this recurrent network, multiplied by their weights, would be linearly summed to produce the value estimate. (See color plate 63.)

given a particular state occupancy; individuals units of the output recurrent layer would each correspond to a particular state and would each fire in proportion to the probability allotted to that state, collectively reading out the belief state (figure 48.3*B*). These output units would provide inputs to the striatum, where they could shape value predictions.

One important question in reinforcement learning is how an agent knows in the first place which features, or "states," to learn. Even in a simple TD model such as the CSC TD model, time could be tracked from the onset of any observable stimuli, meaning that the number of temporal "states" the TD model could erroneously assign weight is enormous. The same sort of problem exists when considering the belief-state TD model: there are an unconstrained number of possible states to compute inferences over in any given environment. How does the brain select the right belief state appropriate to maximizing reward in the current task? One possibility is that units projecting from cortex to striatum are maximally active when they represent higher belief in a particular state—and, critically, only in states that are relevant to the task. One way the cortex may be able to accomplish this is by using spike-timing-dependent synaptic mechanisms to hone a state representation based on temporal coincidences during a task. This was postulated as a mechanism for learning about state space during vocal learning in the songbird (Mackevicius & Fee, 2018). If the brain actively anticipates state transitions (e.g., tries to predict reward timing), temporal coincidences that commonly occur during the task may automatically hone the belief state representation into only that relevant for the current task. A second possibility is that the cortex uses the dopamine signal to compute a lower-dimensional belief state that helps the animal maximize discounted future reward. Rao proposed feeding the belief state outputs into a hidden layer containing fewer units and computing value using the outputs of this smaller hidden layer (Rao, 2010). The weights for inputs into this "hidden" layer would be tuned to the relevant belief-state representation by being trained on dopamine-like TD error signals. A third possibility is that the cortex simultaneously computes many belief states (some of these irrelevant for the current task) and feeds all of these into the striatum. Then, only those cortical inputs carrying belief states relevant to value prediction on the current task achieve synaptic potentiation with their striatal targets. One mechanism by which this could occur is dopamine-dependent modulation of the spike-timing-dependent plasticity (STDP) rule (Brzosko, Zannone, Schultz, Clopath, & Paulsen, 2017). The timing of dopamine release itself also

-1—
0—
+1—

modulates the strength of STDP-evoked dendritic spine enlargement in the striatum (Yagishita et al., 2014). In these ways, dopamine itself may play a role in strengthening only the belief-state inputs into the striatum that are temporally coincident with rewards, leading to greater weights for only the subset of belief-state inputs that allow the animal to maximize future rewards.

The belief-state TD model extends the theoretical framework for reinforcement learning in the brain. It connects the cortex's ability to represent probabilistic models of the environment with the goal of computing accurate future values. Further experiments should probe the neural implementation of a belief state and identify ways in which the brain efficiently knows which belief state to use in order to maximize expected future reward.

Acknowledgments

The authors thank Dr. Samuel Gershman for the development of the authors' studies. The authors thank all the members of the Uchida lab for discussions and for their contributions to various aspects of the studies cited in this article.

REFERENCES

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature Communications, 9*(1), 1891. https://doi.org/10.1038/s41467-018-04397-0.

Balsam, P. D., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neurosciences, 32*(2), 73–78. https://doi.org/10.1016/j.tins.2008.10.004.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47*(1), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology, 104*(2), 1068–1076. https://doi.org/10.1152/jn.00158.2010.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience, 19*(23), 10502–10511.

Brzosko, Z., Zannone, S., Schultz, W., Clopath, C., & Paulsen, O. (2017). Sequential neuromodulation of Hebbian plasticity offers mechanism for effective reward-based navigation. *eLife, 6*. https://doi.org/10.7554/eLife.27756.

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., & Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nature Neuroscience, 19*(1), 111–116. https://doi.org/10.1038/nn.4191.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature, 482*(7383), 85–88. https://doi.org/10.1038/nature10754.

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2015). Reversal learning and dopamine: A Bayesian perspective. *Journal of Neuroscience, 35*(6), 2407–2416. https://doi.org/10.1523/JNEUROSCI.1989-14.2015.

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*(5867), 1264–1267. https://doi.org/10.1126/science.1150605.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation, 18*(7), 1637–1677. https://doi.org/10.1162/neco.2006.18.7.1637.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., & Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature, 525*(7568), 243–246. https://doi.org/10.1038/nature14855.

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience, 19*(3), 479–486. https://doi.org/10.1038/nn.4239.

Fiorillo, C. D., Newsome, W. T., & Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience, 11*(8), 966–973. https://doi.org/10.1038/nn.2159.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., … Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. *Nature, 469*(7328), 53–57. https://doi.org/10.1038/nature09588.

Fuhs, M. C., & Touretzky, D. S. (2007). Context learning in the rodent hippocampus. *Neural Computation, 19*(12), 3173–3215. https://doi.org/10.1162/neco.2007.19.12.3173.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience, 26*(32), 8360–8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006.

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience, 34*(3), 698–704. https://doi.org/10.1523/JNEUROSCI.2489-13.2014.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4), 304–309. https://doi.org/10.1038/1124.

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.), *Computational neuroscience: Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Howe, M. W., & Dombeck, D. A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature, 535*(7613), 505–510. https://doi.org/10.1038/nature18942.

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *Journal of Neuroscience,*

28(31), 7837–7846. https://doi.org/10.1523/JNEUROSCI.1600-08.2008.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology, 27*(6), 821–832. https://doi.org/10.1016/j.cub.2017.02.026.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology, 49*, 1–7. https://doi.org/10.1016/j.conb.2017.10.006.

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*(12), 3034–3054. https://doi.org/10.1162/neco.2008.11-07-654.

Mackevicius, E. L., & Fee, M. S. (2018). Building a state space for song learning. *Current Opinion in Neurobiology, 49*, 59–68. https://doi.org/10.1016/j.conb.2017.12.001.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature, 447*(7148), 1111–1115. https://doi.org/10.1038/nature05860.

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife, 6*. https://doi.org/10.7554/eLife.21886.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology, 72*(2), 1024–1027.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences, 10*(8), 375–381. https://doi.org/10.1016/j.tics.2006.06.010.

Nomoto, K., Schultz, W., Watanabe, T., & Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *Journal of Neuroscience, 30*(32), 10692–10702. https://doi.org/10.1523/JNEUROSCI.4828-09.2010.

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience, 25*(26), 6235–6242. https://doi.org/10.1523/JNEUROSCI.1478-05.2005.

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., … Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience, 19*(6), 845–854. https://doi.org/10.1038/nn.4287.

Pasquereau, B., & Turner, R. S. (2015). Dopamine neurons encode errors in predicting movement trigger occurrence. *Journal of Neurophysiology, 113*(4), 1110–1123. https://doi.org/10.1152/jn.00401.2014.

Rao, R. P. N. (2010). Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience, 4*, 146. https://doi.org/10.3389/fncom.2010.00146.

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

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10*(12), 1615–1624. https://doi.org/10.1038/nn2013.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., … Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience, 20*(5), 735–742. https://doi.org/10.1038/nn.4538.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience, 20*(4), 581–589. https://doi.org/10.1038/nn.4520.

Starkweather, C. K., Gershman, S. J., & Uchida, N. (2018). The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron, 98*(3), 616–629.e6. https://doi.org/10.1016/j.neuron.2018.03.036.

Stauffer, W. R., Lak, A., Yang, A., Borel, M., Paulsen, O., Boyden, E. S., & Schultz, W. (2016). Dopamine neuron-specific optogenetic stimulation in rhesus macaques. *Cell, 166*(6), 1564–1571.e6. https://doi.org/10.1016/j.cell.2016.08.024.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience, 16*(7), 966–973. https://doi.org/10.1038/nn.3413.

Stuber, G. D., Klanker, M., de Ridder, B., Bowers, M. S., Joosten, R. N., Feenstra, M. G., & Bonci, A. (2008). Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. *Science, 321*(5896), 1690–1692. https://doi.org/10.1126/science.1160873.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research, 121*(3), 350–354.

Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience, 91*(3), 871–890.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*(1), 9–44. https://doi.org/10.1007/BF00115009.

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

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron, 95*(6), 1395–1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025.

Tian, J., Huang, R., Cohen, J. Y., Osakada, F., Kobak, D., Machens, C. K., … Watabe-Uchida, M. (2016). Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron, 91*(6), 1374–1389. https://doi.org/10.1016/j.neuron.2016.08.018.

Tian, J., & Uchida, N. (2015). Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors. *Neuron, 87*(6), 1304–1316. https://doi.org/10.1016/j.neuron.2015.08.028.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science, 345*(6204), 1616–1620. https://doi.org/10.1126/science.1255514.