Chapter 10: Learning to predict

Learning to predict significant future events is a fundamental task facing all animals. In this chapter, we study reward and punishment prediction as a paradigmatic predictive learning problem. Using classical (Pavlovian) conditioning as a case study, we show how simple reinforcement learning algorithms can explain many aspects of how animals learn to predict reward/punishment. These algorithms rely on the gradient-based optimization principles introduced in the last chapter. They can also be generalized to compute full posterior distributions over parameters, rather than just point estimates, thereby explaining the sensitivity of animals to uncertainty. Finally, we show how reinforcement learning algorithms can be implemented in the basal ganglia, where dopamine provides the critical error signal for learning.

The last chapter introduced a general framework for learning algorithms in the brain. We now leverage this framework to understand how the brain learns to predict future reward. The technical study of this problem in engineering is known as *reinforcement learning* (RL), which forms the basis of many impressive AI achievements, such as training autonomous robots (Tang et al., 2025) and systems that play human-level Atari games (Mnih et al., 2015). We will show how similar algorithms appear to be used by the brain. We will also see in Chapter 12 how variations of these algorithms can be applied to learn more general predictions beyond reward.

See Sutton and Barto (2018) for a general introduction to RL.

To contain the scope of our treatment, we will focus on "pure" reward/punishment prediction in this chapter, deferring consideration of control (how to intervene on the environment to maximize reward and minimize punishment) until the next chapter. Fortunately, pure prediction has been extensively studied in animals, using experimental protocols where animals are exposed to reward-predictive stimuli without being able to control the stimuli or rewards. The most famous of these protocols is *classical conditioning*, which we will describe in the next section. This will provide a rich empirical testbed for thinking about what kinds of algorithms the brain might use to predict reward. We will then connect these algorithmic ideas with neurobiology, exploring how RL algorithms could be implemented in basal ganglia circuits under the supervision of dopamine signaling.

In the interest of brevity, we will henceforth refer primarily to reward prediction, even though some phenomena concern punishment. Although an oversimplification, we can for now think of punishment as negative reward.

Also known as *Pavlovian conditioning* due to the pioneering contributions of (Pavlov, 1927).

1 *Classical conditioning*

In a standard classical conditioning protocol (see Table 1 for examples), an animal is exposed to a neutral stimulus (the conditioned

stimulus, or CS) followed by an appetitive (good) or aversive (bad) stimulus (the unconditioned stimulus, or US). Most protocols employ *delay conditioning*, where the onset of the US coincides with the offset of the CS. The key variable is the *conditioned response* (CR) to the CS onset. The CR typically increases over the course conditioning. Importantly, this increase is not due merely to repeated stimulus exposure, because it does not occur if the relative timing of the CS and US is randomized (i.e., there is no stable temporal relationship between the two stimuli); this implies that the temporal relationship between the stimuli is fundamental to the emergence of the CR. The CR is thought to reflect the *reinforcement* of the CS by its temporal relationship with the US.

Another common protocol is *trace conditioning*, where the CS offset and US onset are separated by a "trace" interval. Some protocols also present the US at some point during (or even before) the CS.

CS	US	CR
Tone	Food	Salivation
Tone	Shock	Freezing
Light	Air puff	Eyeblink
Taste	Nausea	Taste aversion
Lever	Food	Approach
Light	Food	Approach
Tone	Shock	Suppression

Table 1: Examples of classical conditioning protocols.

A concrete example of classical conditioning (pigeon autoshaping) is schematized in Figure 1. The CS is a keylight which predicts the delivery of food (the US) into the hopper. With repeated pairings, the pigeon begins to peck at the keylight (the CR). Note that food delivery is independent of pecking.

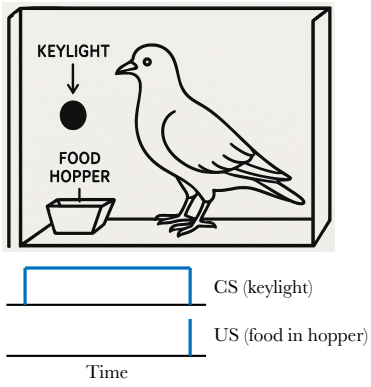

Figure 1: **Pigeon autoshaping**.

What are animals learning during classical conditioning? A natural hypothesis is that the CR reflects a prediction about upcoming reward. This hypothesis can explain several aspects of conditioned responding. All other things being equal, the CR rate is greater when the CS-US delay is shorter (except for very short delays, at which point anticipation may not be useful) and when the reinforcement rate (the CS-conditional US rate) is greater (Harris and Carpenter, 2011). This suggests that the CR is closely tied to the expected rate of reinforcement in the near future following the appearance of the CS. Another aspect compatible with a predictive view is the fact that prediction errors drive learning: an unexpected US following the CS tends to increase the CR, whereas the omission of an expected US following the CS tends to decrease the CR. The role of prediction errors in learning is particularly striking when one examines protocols with multiple simultaneously presented CSs (known as *compound conditioning*). For example, the CR to a CS is weaker if that CS is paired with a previously reinforced CS, compared to reinforcing the CS alone. In both cases, the CS was reinforced the same number of times, but in the former case the previously reinforced CS "blocks" the new CS ostensibly because it already adequately predicts the US—there is no prediction error to drive learning.

This *blocking effect* is called "Kamin" (or forward) blocking, after Kamin (1968).

It will be useful to establish some standard notation for describing these kinds of experiments. We'll use uppercase letters (A, B, C, etc.) to denote CSs; compound CSs will be denoted by concatenations (e.g., AB denotes the compound presentation of A and B). A reinforced CS will be denoted by A+, and an unreinforced CS will be denoted A-. A test stimulus (typically presented without reinforce-

ment) is denoted by B? and the resulting test CR is the variable of interest. Thus, the compound condition of the Kamin blocking experiment can be described as A+/AB+/B?, which can be compared with a control condition A+/B+/B? (i.e., only single-CS training).

Another striking example of how prediction error drives learning: the *overexepectation effect*. Two CSs are reinforced separately (A+/B+), then reinforced in compound (AB+), and finally the response to each one is tested individually (A? B?). This protocol produces a *reduction* in the CR compared to a protocol in which the compound reinforcement phase is omitted (Rescorla, 1970). In other words, conditioned responding is lower despite the animal receiving more reinforcements! Why? Intuitively, animals come to predict a regular amount of reinforcement to each individual CS, and then predict twice as much when the two CSs are presented together (under the assumption that reward predictions summate across the elements of a compound). The receipt of the same amount of reinforcement produces a negative prediction error, driving a reduction in the predictions for each CS. In the next section, we will formalize these intuitions.

Imagine you're a barista who is regularly tipped $1 by customer A and $1 by customer B. When you see both customers arrive, it's natural to assume that you'll receive $2 total. If you only receive $1, this suggests that your predictions were too high (a negative prediction error) and therefore you should reduce them. The next time one of the customers arrives, you'll predict a reduced tip.

2 *The Rescorla-Wagner model*

The empirical observations described above can be captured by a simple model due to Rescorla and Wagner (1972), which we will refer to as the Rescorla-Wagner model. We will describe it in a slightly simplified form here. A CS configuration is represented by a vector x, where $x_d = 1$ denotes the presence of CS d and $x_d = 0$ denotes its absence. In neural terms, we can think of x as the activity of neurons tuned to different CSs. A US is denoted by r (typically binary). The US prediction $\hat{r}$ is a linear function of the CS vector:

$$\hat{r} = \sum_d w_d x_d, \tag{1}$$

where w_d is an associative strength (or synaptic weight, in neural terms) between CS d and the US, typically initialized to 0. The weight is updated based on the prediction error $r - \hat{r}$:

$$\Delta w_d = \eta x_d (r - \hat{r}), \tag{2}$$

where $\eta \in [0, 1]$ is a learning rate. Intuitively, the weight is increased when the prediction error is positive (more reward was received than predicted) and decreased when the prediction error is negative (less reward was received than predicted). It is also essential for credit (or blame) assignment that the CS be present in order for its weight to change.

Despite its simplicity, the Rescorla-Wagner model can explain a remarkably wide range of classical conditioning phenomena. Consider

the Kamin blocking effect: if CS A has already been paired with the US, then $w_A \approx 1$. This means that during the compound conditioning phase $\hat{r} = w_A + w_B \approx 1 + 0$, and as a consequence the prediction error is approximately 0, preventing learning of a non-zero weight for CS B.

The same principles can be applied to understanding the overexpectation effect. After separate reinforcement of A and B, each weight is approximately 1. When presented in compound, the US prediction is then $\hat{r} = w_A + w_B \approx 1 + 1 = 2$. When $r = 1$ is received during the compound conditioning phase, the prediction error is $r - \hat{r} \approx 1 - 2 = -1$. This leads to a decrement of both w_A and w_B.

Reflecting on the these and other successes of the Rescorla-Wagner model, the most important principles it embodies are:

- Learning driven by prediction errors.
- Additive combination of weights.
- Credit assignment based on CS presence.

As we will see in this chapter and in the next few chapters, these are fairly robust principles of learning in animals—BUT, they do not exhaustively describe the principles of animal learning. We will explore two (not mutually exclusive) ways in which this gap can be addressed. One is to search for more general principles that encompass both the original principles and their violations. Another is to posit the coexistence of several learning systems in the brain.

Before we come to the shortcomings of the Rescorla-Wagner model, let's try to gain a deeper appreciation of its successes, by deriving it from first principles as an optimization algorithm. This will allow us to see how it is connected to the optimization picture developed in the last chapter. A simple way to formalize the problem facing the animal is to define a loss function based on predictive accuracy, such as $L(\hat{r}, r) = (r - \hat{r})^2$. Taking the gradient of the loss with respect to the weights yields $\nabla_w L \propto x(r - \hat{r})$. Thus, the Rescorla-Wagner update (Eq. 2) can be derived as gradient descent on the squared error loss.

3 *Learning in the absence of stimuli*

As we saw above, animals don't learn in the absence of prediction errors. Or do they? Consider this small variation on the standard conditioning protocol: prior to CS-US pairings, the CS is presented repeatedly by itself. These CS-alone trials retard subsequent acquisition of the CR, a phenomenon known as *latent inhibition* (Lubow,

Latent inhibition is also sometimes known as the *CS pre-exposure effect*.

1973). The challenge posed by this effect is that if the weight is initialized to 0, so the US prediction is initially 0 when the CS-alone trials occur, then the prediction error will also be 0. The Rescorla-Wagner model predicts no learning in this case, and yet the animal is clearly learning something.

Another challenge for the Rescorla-Wagner model stems from the requirement that learning only occurs for present stimuli. But consider a small variation on the Kamin blocking protocol: switch the order of A+ and AB+ training, so that AB+ comes first. After AB+ training, the animal produces a CR in response to B, but this response is reduced after A+ training—a phenomenon known as *backward blocking* (Miller and Matute, 1996). This can't happen in the Rescorla-Wagner model, because $x_B = 0$ during A+ training, and therefore Δw_B must equal 0.

There are many other examples of "retrospective revaluation" effects where learning about one CS affects later responses to an absent CS. In the Kamin blocking protocol, presenting A by itself after compound training (A+/AB+/A-) has the effect of "unblocking" B (Blaisdell et al., 1999). Similarly, presenting A by itself after the compound training phase of the overexpectation protocol (A+/B+/AB+/A-) has the effect of rescuing the CR to B (Blaisdell et al., 2001).

Presenting a CS by itself after conditioning is known as *extinction*.

To address these and related problems (see Gershman, 2015), we turn to a probabilistic view of the learning problem facing animals.

4 A probabilistic view

Gradient descent on the squared error loss is a point estimation procedure; it ignores uncertainty about the weights. We can derive a different normative analysis by computing a full posterior over the weights given the CS/US history. To do this, we need to specify a set of assumptions about the animal's internal model of the world. Suppose the animal's internal model assumes that the US is a noisy linear combination of CS features:

$$r = \sum_d w_d x_d + \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_r^2)$. Suppose further that the animal's internal model also assumes that the weights are Gaussian-distributed: $w_d \sim \mathcal{N}(0, \sigma_w^2)$. Under this linear-Gaussian internal model, the posterior is also Gaussian, with mean $\hat{w}$ and covariance matrix Σ, updated recursively according to:

It is possible to generalize this model to a time-varying weight vector (Dayan and Kakade, 2001; Gershman, 2015), but we omit this for simplicity.

$$\Delta \hat{w} = \eta(r - \hat{r}) \tag{4}$$

$$\Delta \Sigma = -\eta x^\top \Sigma, \tag{5}$$

These updates are a special case of the *Kalman filtering* equations, named after Kalman (1960). The learning rate vector η corresponds to the *Kalman gain*.

where η is now vector-valued with CS-specific learning rates (and x now appears projected onto the covariance matrix):

$$\eta = \frac{\Sigma x}{x^\top \Sigma x + \sigma_r^2}. \tag{6}$$

Notice the similarity of the posterior mean update (Eq. 4) to the Rescorla-Wagner update. In fact, if the covariance is the identity matrix, $\Sigma = I$, then the two are equivalent apart from a time-varying learning rate. In general, the covariance will not be diagonal, and this has important implications for classical conditioning.

Let's first consider the case where there is a single CS, so Σ is a scalar. Each time the CS is presented, Σ decreases. This has the effect of also decreasing the learning rate, because

$$\eta = \frac{1}{1 + \frac{\sigma_r^2}{x^2 \Sigma}}. \tag{7}$$

This expression can also explain why partial reinforcement (making the CS a less reliable predictor of the US) typically slows learning (Jenkins and Stanley, 1950; Gottlieb, 2004), by increasing σ_r^2 and thus reducing η.

Intuitively, the animal becomes more confident (the posterior variance shrinks) as it collects more data, which makes it more resistant to learning from future observations. Latent inhibition is a natural consequence of these dynamics.

In the case of compound conditioning, Eq. 5 implies that the off-diagonals of the covariance matrix will become negative. Intuitively, this reflects the zero-sum nature of the linear model: the associative strengths must add to 1, so increasing the weight for one CS requires that the weight for the other be decreased. This naturally produces many retrospective revaluation effects, because the learning rates for absent stimuli will be negative. In the backward blocking protocol (AB+/A+), for example, the strengthening of the weight for A during the second phase will lead to a weakening of the weight for B due to $\eta_B < 0$ during A+ training.

5 *Long-range prediction*

So far, we have been assuming that the computational problem facing the animal is predicting the next US, but this neglects the fact that animals are not completely myopic—they care about events farther in the future. Consider the case of *second-order conditioning*, where one CS (A) is first trained and then another CS (B) is paired with A such that the onset of A coincides with the offset of B. Animals acquire a CR to B even though it is never paired with the US. Apparently the animals treat A as a kind of proxy for future reinforcement.

The models we've introduced so far have no mechanism for explaining second-order conditioning. What's needed is a representation of long-range predictions. This is where we appeal to the RL

concepts that we alluded to at the beginning of the chapter. We start by redefining the computational problem as the goal of predicting *expected discounted return*, or *value*:

$$V_t = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots], \tag{8}$$

where we have introduced the time index t. The parameter $\gamma \in [0, 1)$ is the *temporal discount factor*, which reflects the animal's preference for obtaining rewards sooner rather than later (in keeping with standard RL terminology, we will refer to rewards here rather than to USs). The expectation averages over stochasticity in the reward sequence.

This form of exponential discounting can alternatively be understood in the following way: if at each time step you might die with probability $1 - \gamma$, then the probability of surviving H time steps is γ^H. Eq. 8 is thus equivalent to the expected *undiscounted* return over a stochastic horizon H drawn from a geometric distribution.

At first glance, this seems like a totally impossible problem. How can an animal estimate an expectation of an infinite series? Indeed, the problem is impossible unless some additional structure is assumed. The key move in RL theory is to assume that the rewards are conditionally independent given an underlying state s: the expected reward in state s is given by $R(s)$. The state evolves according to a Markov process with transition probability $T(s'|s)$, where we've adopted the notational convention of writing transitions as $s \to s'$. With these additional assumptions, we can express the value in a recursive form known as the *Bellman equation*:

A process is Markovian if it is memoryless: the probability of the next state depends only on the current state, not the preceding state or observation history.

$$V(s) = R(s) + \gamma \sum_{s'} T(s'|s) V(s'). \tag{9}$$

where we have redefined the value as a function of the state because it is now also Markovian. The principal benefit of the Bellman equation is that it allows us to derive a learning procedure for estimating the value function. Let $\hat{V}(s)$ denote the value function approximator. The Bellman equation stipulates that the *temporal difference* (TD) prediction error

Here we assume that s' is sampled from the transition distribution $T(\cdot|s)$.

$$\delta = r + \gamma \hat{V}(s') - \hat{V}(s) \tag{10}$$

is on average 0 when the approximation is exact, $\hat{V} = V$. When $\delta > 0$, the value function has been underestimated and $\hat{V}(s)$ should be increased; when $\delta < 0$, the value function has been overestimated and $\hat{V}(s)$ should be decreased. More systematically, the TD learning algorithm updates the value function approximator according to:

$$\Delta \hat{V}(s) = \eta \delta, \tag{11}$$

where η is a scalar learning rate. We can understand this update as gradient descent on a TD loss function:

$$L(\hat{V}(s), \hat{V}^*(s)) = [\hat{V}^*(s) - \hat{V}(s)]^2 = \delta^2, \tag{12}$$

where the target $\hat{V}^*(s) = R(s) + \gamma \hat{V}(s')$ is "bootstrapped"—it relies on the estimator for the next state. Nonetheless, the algorithm will still converge to the correct values when provided with sufficient experience (Sutton and Barto, 2018).

Eq. 11 represents the simplest TD learning algorithm, with value estimates stored in a look-up table. For environments with many states, this is not an efficient strategy, since the animal would need to experience each state multiple times to obtain a good estimate (i.e., there's no generalization across states). To address this shortcoming, we can replace the look-up table with a function approximator parametrized by w, and then optimize these parameters using gradient descent on the TD loss (Eq. 12).

To build a bridge to the discussion of classical conditioning, we'll take each state to be represented by a feature vector $x = f(s)$ for some encoding function f (e.g., where each feature represents a single CS presence/absence), and the function approximator to be linear, $\hat{V}(s) = \sum_d w_d x_d$. Gradient descent on the TD loss leads to the following update:

$$\Delta w_d = \eta x_d \delta. \tag{13}$$

We have arrived at an equation that is remarkably similar to the Rescorla-Wagner learning rule. In fact, the two learning rules are identical in the "myopic limit" ($\gamma = 0$), where the animal only cares about predicting immediate reward. TD learning with linear function approximation thus shares all of the key properties of the Rescorla-Wagner model, but goes beyond it by estimating long-range predictions. This allows TD learning to capture phenomena like second-order conditioning, because the future reward term $\gamma \hat{V}(s')$ is greater than 0 during B→A training even though the immediate reward (r) is 0.

You might think that something is amiss here—shouldn't the correct long-range prediction be that B *inhibits* the arrival of reward following A? And you would be right! The elevated value estimate for B is transient, and will eventually become 0 or even negative (conditioned inhibition), depending on our assumptions about how the stimuli are represented. This is consistent with the finding that increasing the number of B→A shifts the pattern of responses from second-order conditioning to conditioned inhibition (Yin et al., 1994).

If A and B share some features, then B's unique features will need to have negative weights to counteract the positive weights learned for the shared features on A-alone trials.

6 *Putting it all together*

We've seen that two different generalizations of the Rescorla-Wagner model can capture a diverse range of phenomena. We can combine the complementary advantages of these generalizations by defining

a probabilistic model for the value function approximator (Gershman, 2015). As above, we assume that the weights are Gaussian-distributed. We also assume that the true values are noisy linear combinations of state features x:

$$V(s) = \sum_d w_d x_d + \epsilon, \tag{14}$$

with $\epsilon \sim \mathcal{N}(0, \sigma_v^2)$. To keep things simple, we'll assume deterministic rewards and transition dynamics. We can then use the Bellman equation to write down the distribution on rewards:

This expression relies on the fact that a linear transformation of a Gaussian random is also Gaussian.

$$\begin{aligned} r &= V(s) - \gamma V(s') + \tilde{\epsilon} \\ &= w^\top x - \gamma w^\top x' + \tilde{\epsilon} \\ &= w^\top h + \tilde{\epsilon}, \end{aligned} \tag{15}$$

where $\tilde{\epsilon} \sim \mathcal{N}(0, (1+\gamma^2)\sigma_v^2)$ and $h = x - \gamma x'$. Because this is another linear-Gaussian system, we can again apply the Kalman filtering equations to obtain the posterior mean and covariance matrix updates:

$$\Delta w = \eta \delta \tag{16}$$

$$\Delta \Sigma = -\eta h^\top \Sigma, \tag{17}$$

where η is once again a vector. These updates are very similar to Eqs. 4 and 5. The main difference is that the prediction error ($r - \hat{r}$) is replaced by the TD error in the mean update, and h replaces x in the covariance update. Both reduce to the original updates when $\gamma = 0$.

This model can explain phenomena that none of the other models described above can explain on their own. For example, consider the following variation on a second-order conditioning protocol. After second-order conditioning, the first-order CS (A, which was previously paired directly with the US) is extinguished by presenting it alone. This causes a reduction not only in responding to A but also in responding to the second-order stimulus B (Rashotte et al., 1977). TD learning cannot explain this, because (like the Rescorla-Wagner model) it doesn't update value estimates for absent stimuli. The probabilistic version of TD learning, in contrast, can explain this effect by virtue of the fact that the onset of one coincides with the offset of the other, and therefore the covariance update implies that they will have positive covariance. Decreasing the weight for A will thereby decrease the weight for B.

Neither the Rescorla-Wagner model nor its probabilistic variant can explain second-order conditioning to begin with, let alone the effects of post-training extinction.

7 *The neural architecture of reinforcement learning*

Representations of value can be found in multiple brain areas (e.g., Hampton and O'Doherty, 2007; Ottenheimer et al., 2023), but the

striatum (a subcortical structure that is part of the basal ganglia) is thought to play a pivotal role in the transformation of stimulus features (represented by cortical inputs) to value estimates. In particular, the ventral subdivision of the striatum, the nucleus accumbens, contains neurons whose activity is sensitive to both US and CS delivery during classical conditioning (Roitman et al., 2005; Day et al., 2006). An example neuron is shown in Figure 2, firing in response to the CS+ and the delivery of a sucrose reward, but not to the CS-.

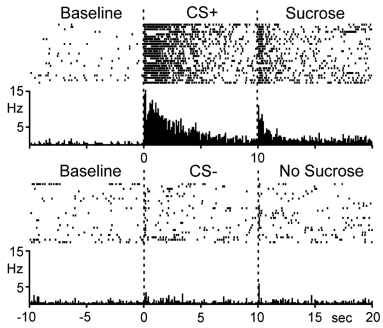

Figure 2: **Responses of a single neuron in the nucleus accumbens following classical conditioning**. Reproduced from Day et al. (2006).

Lesions of the nucleus accumbens (in particular, its "core" region that is functionally and anatomically distinct from its "shell") impair both the acquisition (Parkinson et al., 2000) and expression (Cardinal et al., 2002) of conditioned responding. Local antagonism of NMDA receptors, which are typically necessary for the induction of long-term potentiation, also impairs acquisition (Di Ciano et al., 2001), indicating the necessity of synaptic plasticity in the nucleus accumbens for classical conditioning.

If cortical inputs to the striatum encode the feature vector $x = f(s)$, and striatal neurons encode $\hat{V}(s)$, then (under the linear function approximation assumption) corticostriatal synapses correspond to the weights w. The weight updates should therefore follow the learning rules described above. In particular, we can understand the TD update (Eq. 13) as a form of "predictive Hebbian learning" (Montague et al., 1996), where weights are updated in proportion to the coincidence of presynaptic activity with a prediction error signal (Eq. 10). The question is then where this prediction error comes from. Considerable evidence suggests that phasic (fast timescale) signaling via the neuromodulator dopamine conveys prediction errors (Gershman et al., 2024), as we review next.

7.1 Dopamine signaling of prediction errors

Prediction errors of the form required by TD learning (Eq. 10) have several signatures that are also exhibited by dopamine neurons in the midbrain (Figure 3). First, they increase in response to unexpected reward, such as at the beginning of conditioning or in the absence of a preceding CS. Second, predicted rewards elicit no response at the time of reward, but do elicit a response at the onset of the earliest reward-predicting CS. Third, the omission of an expected reward causes a suppression of dopamine activity.

Beyond these qualitative characterizations, the activity of dopamine neurons appears to quantitatively conform to a prediction error. At the time for reward, the value estimate should approximately equal a weighted average of recent rewards, with the weights decaying exponentially. To see this, note that at the time of reward (which we de-

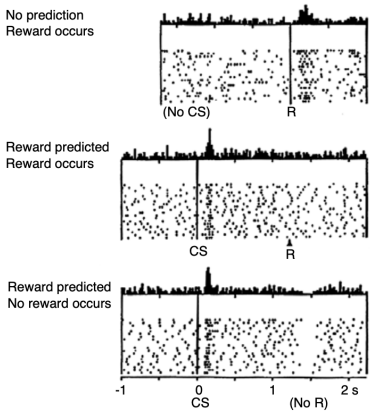

Figure 3: **Activity of dopamine neurons during classical conditioning**. "R" indicates reward delivery. Reproduced from Schultz et al. (1997).

note by s_r), the prediction error is given by $\delta = r - \hat{V}(s_r)$, under the assumption that the second term in the prediction error definition, $\gamma\hat{V}(s')$, is negligible (due to the fact that the next reward will happen far in the future). Letting $\hat{V}_n$ denote $\hat{V}(s_r)$ after n conditioning trials and r_n denote the reward received on trial n, we have:

Notice that the learning rate η determines the slope of the exponential decay.

$$\begin{aligned}\hat{V}_n &= \eta r_n + (1-\eta)\hat{V}_{n-1} \\ &= \eta r_n + \eta(1-\eta)r_{n-1} + (1-\eta)^2\hat{V}_{n-2} \\ &= \eta \sum_{k=0}^{n}(1-\eta)^k r_{n-k}. \end{aligned} \tag{18}$$

where we have assumed that $\hat{V}_0 = 0$. This implies that if dopamine quantitatively reports a prediction error, then we should be able to predict activity at the time of reward on trial n as a linear function of past rewards:

$$\delta_n = r_n - \sum_{k=0}^{n} b_k r_{n-k}, \tag{19}$$

with a log-transformed regression coefficient given by:

$$\log b_k = k\log(1-\eta) + \text{const.} \tag{20}$$

which is a negative linear function of trial lag k. Bayer and Glimcher (2005) fit a lagged regression model to recordings from dopamine neurons, confirming that the coefficients do indeed decay exponentially (Figure 4).

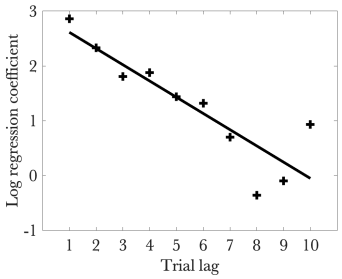

Figure 4: **Log-transformed lag regression coefficients**. Based on regression coefficients taken from Bayer and Glimcher (2005).

Other experiments have verified that dopamine activity matches the functional form of a prediction error, namely the difference between observed and predicted reward. Eshel et al. (2015) showed that the activity of dopamine neurons quantitatively matches the predictions of a subtractive model (Figure 5); this holds true across a range of reward magnitudes in both the presence and absence of a CS. The same data were not well-matched by alternative models assuming division rather than subtraction. Further work by Eshel et al. (2016) showed that a subtractive model also quantitatively accounts for the parametric suppression of dopamine neuron responses by different levels of reward expectation.

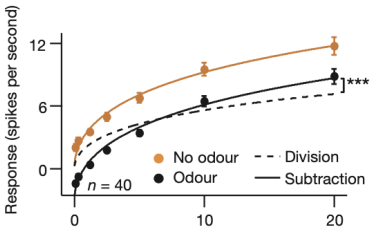

Figure 5: **Dopamine neuron activity reports the difference between observed and predicted reward**. Reproduced from Eshel et al. (2015).

As discussed above, prediction errors play a crucial role in explaining many learning phenomena. We can interrogate these explanations by looking directly at the prediction errors putatively signaled by dopamine. The explanation of Kamin blocking, for example, hinged on the absence of prediction errors during compound training. This suggests that dopamine responses should be similarly suppressed during compound training. To test this, Waelti et al. (2001) first paired one CS with reward (A+) and another without reward (B-). They then trained new stimuli in compound with each

pretrained CS (AX+ and BY+). When subsequently tested on X and Y by themselves, animals produced a stronger CR to Y than to X—the classic Kamin blocking effect. During compound conditioning, dopamine neurons showed substantial responses to reward following the BY compound, but not following the AX compound (Figure 6), consistent with the absence of a positive prediction error when a new CS is reinforced in compound with a pretrained CS.

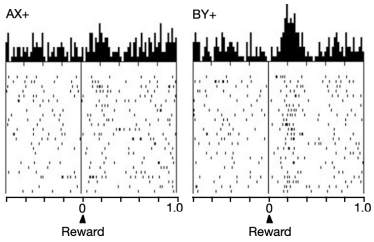

Figure 6: **Dopamine responses during compound training**. Reproduced from Waelti et al. (2001).

The role of dopamine in blocking is causal: artificially stimulating dopamine neurons during compound training "unblocks" learning (Steinberg et al., 2013), consistent with the hypothesis that this produces a positive prediction error capable of driving learning. Stimulating dopamine neurons can also offset the reductions in conditioned responding during extinction (when a previously trained CS is presented in the absence of reward) or when the reward is reduced by shifting from sucrose to water.

7.2 *Stimulus representation*

We have glossed over an important detail about the mapping of classical conditioning experiments onto TD learning. The Rescorla-Wagner model is a trial-based model, but TD learning is typically applied to these experiments at finer-grained timescale; it can model the effects of different temporal arrangements on both conditioned responding and neural activity. This requires making assumptions about how the temporal arrangements are represented by the feature vector (x).

A standard assumption, known as the *complete serial compound* (CSC), breaks each stimulus down into a contiguous set of binary temporal features. Only one feature is active ($x_d = 1$) during each time interval relative to the stimulus onset (Figure 7). This representation allows the function approximator to learn value estimates for each interval (specifically, the weight w_d will converge to the value of the interval during which feature d is active). With the CSC representation, dopamine signals should propagate backwards over time from the US to the CS, as the weight for each feature is progressively updated. Decisive evidence for this was presented by Amo et al. (2022), as shown in Figure 8.

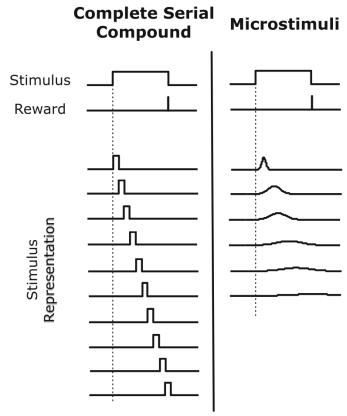

Figure 7: **Stimulus representations**. Reproduced from Ludvig et al. (2012).

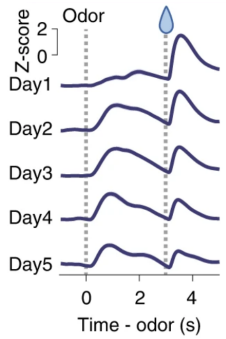

Figure 8: **Gradual backward shift of the dopamine response over the course of classical conditioning**. Reproduced from Amo et al. (2022).

There are, however, practical problems with the CSC: a large number of such features is needed to approximate the value function well, and they don't afford any temporal generalization. For example, if an animal is trained with one CS-US interval and then tested on a slightly longer interval, the weights learned for the CSC representation will abruptly drop after the expected US time. In contrast, animals show a gradual decline in conditioned responding when

tested on longer intervals (Figure 9). This suggests that animals use a representation which allows some degree of temporal generalization.

An alternative *microstimulus* representation was proposed by Ludvig et al. (2008). Each microstimulus corresponds to a radial tuning function with a particular preferred time interval (Figure 7). These correspond closely to *time cells* found in many areas of the brain (Figure 10). Like time cells, the tuning width of microstimuli grows wider with teh preferred interval, reflecting the decline in temporal resolution for longer intervals—a property mirrored in behavior by weaker temporal control for longer intervals (see for example the broader temporal generalization gradient in Figure 9).

Dopamine neurons fire more strongly in response to cued rewards delivered after a longer delay (Figure 11). This finding is incompatible with the CSC, which predicts independence of the delay after sufficient training. In contrast, the microstimulus model predicts this phenomenon (Gershman et al., 2014), due to the fact that their declining temporal precision causes value estimates to get smeared out in time. This suppresses the value at the time of reward, making the prediction error larger.

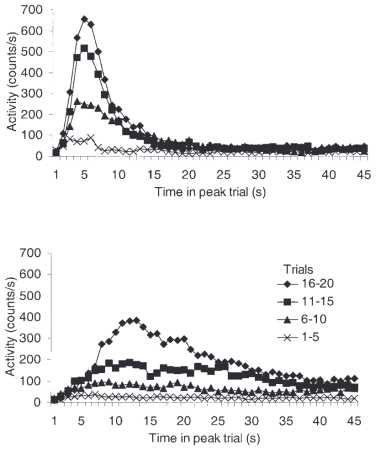

Figure 9: **Conditioned responding on trials where the US is omitted**. Top panel shows data from animals trained with a 5-second CS-US interval; bottom panel shows data from animals trained with a 15-second CS-US interval. Timed responding gradually emerges across trials. Adapted from Drew et al. (2005).

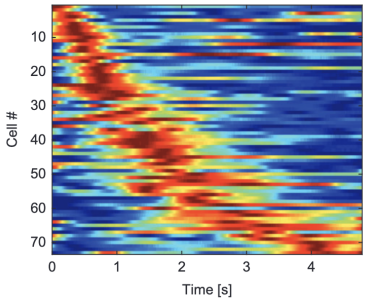

Figure 10: **Time cells in the medial prefrontal cortex**. Each row corresponds to a single neuron, sorted by preferred time interval. The color shows firing rate. Reproduced from Tiganj et al. (2017).

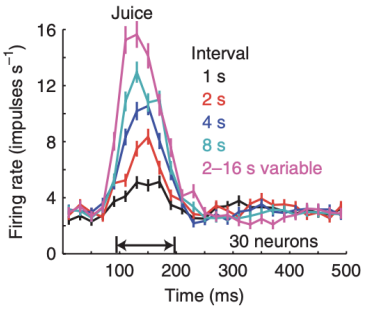

Figure 11: **Dopamine neurons increase their firing rate for more delayed reward delivery**. Reproduced from Fiorillo et al. (2008).

7.3 *Three-factor plasticity rules at corticostriatal synapses*

Let's now return to the plasticity rule implied by dopamine (Eq. 13). It asserts that a sufficient condition for plasticity at corticostriatal synapses is the coincidence of pre-synaptic (cortical) activity with the prediction error (putatively dopamine). It is true that both are necessary, but it turns out that they are not sufficient—post-synaptic (striatal) activity is also necessary (Pennartz et al., 1993). Corticostriatal synapses thus obey a 3-factor Hebbian learning rule: pre-synaptic activity $\times$ post-synaptic activity $\times$ dopamine (Reynolds and Wickens, 2002).

As we'll see in the next chapter, 3-factor Hebbian rules can be derived from policy learning (updating action probabilities to improve reward), a process thought to occur in dorsal striatum. However, this doesn't help explain the discrepancy between the TD learning model and the plasticity rules in nucleus accumbens (ventral striatum). One way to resolve the discrepancy is to posit that the value function is parametrized non-linearly, such as:

$$\hat{V}(s) = \left(\sum_d w_d x_d\right)^{\alpha}, \tag{21}$$

where α is a nonlinear transformation parameter. Under this assump-

tion, the gradient of the TD loss is given by:

$$\nabla_w L \propto x\delta \left(\sum_d w_d x_d \right)^{\alpha-1}. \tag{22}$$

The third term is monotonically increasing in $\hat{V}(s)$ for $\alpha > 1$. This means that the weight update now depends on the post-synaptic activity—a 3-factor learning rule.

There is some evidence for a nonlinear transformation in nucleus accumbens: when synaptic inputs arrive synchronously, medium spiny neurons (the principal cells in the striatum) combine their inputs supralinearly due to activation of NMDA receptors and voltage-dependent calcium channels (Carter et al., 2007). This is precisely the setting in which synaptic plasticity is expected to occur, as discussed in the last chapter.

Neural responses with expansive ($\alpha > 1$) power-law nonlinearities are widely observed in cortex, though less well-studied in subcortical structures like the striatum. These nonlinearities can arise from noise in the subthreshold regime (see Chapter 2), where the average membrane potential is close to but below the firing threshold, so that firing is fluctuation-driven (Miller and Troyer, 2002).

7.4 *A neural implementation of probabilistic TD learning*

Earlier we showed how the TD learning algorithm could be generalized to track the full posterior distribution over weights, allowing it to explain phenomena like latent inhibition and learning about absent stimuli. Here we show how this probabilistic version can be approximated using the combination of two mechanisms: variance normalization of prediction errors, and pre-processing the inputs to the striatum using a recurrent neural network.

It will be useful to reparametrize the probabilistic TD algorithm slightly (Gershman, 2017). Let $\alpha = \Sigma h$ be a vector of "associabilities" (unnormalized learning rates) and let $\lambda = x^\top \Sigma x + \sigma_r^2$ denote the marginal predictive variance (the overall expected error). The learning rate vector η is given by the ratio of these two quantities. With these expressions, we can write Eq. 16 as:

$$\Delta w = \alpha \bar{\delta}, \tag{23}$$

where $\bar{\delta} = \delta/\lambda$ is the TD error normalized by the predictive variance. Several studies have suggested that the activity of dopamine neurons is variance normalized (i.e., that dopamine neurons report $\bar{\delta}$ rather than δ). First, dopamine neuron responses are lower to the same rewards following a CS associated with a wider range of reward magnitudes (Tobler et al., 2005); a wider range effectively increases variance. A more direct experimental test (Figure 12) has shown that high variance decreases the reward sensitivity of dopamine neurons, even when the magnitudes and ranges are held fixed (Rothenhoefer et al., 2021).

We now turn to the computation of the associability vector α. Rather than representing the covariance matrix Σ explicitly, we can

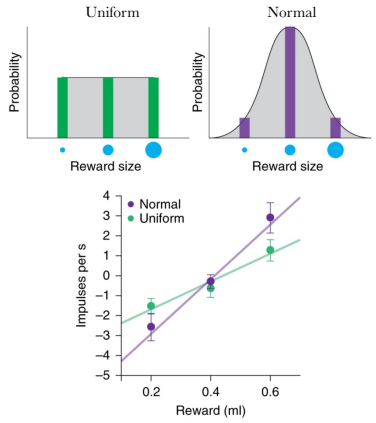

Figure 12: **The reward sensitivity of dopamine neurons for a low-variance (Gaussian) and high-variance (uniform) reward distributions**. Adapted from Rothenhoefer et al. (2021).

represent it implicitly using a recurrent network (Dayan and Kakade, 2001; Gershman, 2017). Let y be the activity of a recurrent network with feedforward inputs h and a recurrent weight matrix B. The linear firing rate dynamics are given by:

$$\tau \dot{y} = -y + h + By, \tag{24}$$

where τ is a time constant. These dynamics converge to:

$$y^{\infty} = (I - B)^{-1} h, \tag{25}$$

where I is the identity matrix. If B is initialized to 0 and updated according to an "anti-Hebbian" rule,

$$\Delta B \propto -h y^{\top} + I - B, \tag{26}$$

then asymptotically $(I - B)^{-1} = \mathbb{E}[\Sigma]$. Thus, $y^{\infty} = \mathbb{E}[\alpha]$. We can then apply the update $\Delta w = y^{\infty} \bar{\delta}$, which is just the standard TD update, but using the transformed inputs (y) and the normalized TD error ($\bar{\delta}$).

One candidate for the implementation of the recurrent transformation is the orbitofrontal cortex, which sends input to the nucleus accumbens (Eblen and Graybiel, 1995). Neurons in this area are sensitive to stimulus covariance. For example, when two stimuli are sequentially paired prior to reinforcement (a procedure known as *sensory preconditioning*), the responses of orbitofrontal neurons to the two stimuli become correlated (Sadacca et al., 2018). Lesioning this region impairs the ability of animals to produce a CR to A after undergoing B+ training (Jones et al., 2012). This pattern of results can be reproduced by the probabilistic TD model described above. When the onset of B coincides with the offset of A, the model learns a *positive* covariance, such that presenting B by itself also activates the representation of A, allowing it to be reinforced even when absent.

8 Conclusion

Several core principles emerge from this chapter. First, learning from prediction errors is a common algorithmic motif across several different models. Second, credit assignment is based on the active representation of stimuli (both those that are present and those that are linked to the present stimuli). Third, learning is sensitive to uncertainty. Fourth, the prediction target for learning is likely long-range (not just immediate upcoming reward). These principles can be realized neurally using fairly simple mechanisms, including variants of Hebbian plasticity rules and recurrent firing rate dynamics.

These principles are only part of the story. In the next chapter, we will invoke some new principles to explain how animals adapt their

actions to maximize reward. And we'll see in Chapter 12 how brains are capable of learning in ways that go beyond simply predicting future reward.

Study questions

1. In what sense can temporal difference learning be viewed as a generalization of the Rescorla-Wagner model? How does the discount factor γ expand the predictive horizon of learning?
2. Why does the probabilistic Kalman filter model naturally account for latent inhibition, while the Rescorla-Wagner model does not?
3. How might nonlinear integration in medium spiny neurons help align neural plasticity with reinforcement learning theory?

References

Amo, R., Matias, S., Yamanaka, A., Tanaka, K. F., Uchida, N., and Watabe-Uchida, M. (2022). A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nature Neuroscience*, 25:1082–1092.

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

Blaisdell, A. P., Denniston, J. C., and Miller, R. R. (2001). Recovery from the overexpectation effect: Contrasting performance-focused and acquisition-focused models of retrospective revaluation. *Animal Learning & Behavior*, 29:367–380.

Blaisdell, A. P., Gunther, L. M., and Miller, R. R. (1999). Recovery from blocking achieved by extinguishing the blocking CS. *Animal Learning & Behavior*, 27:63–76.

Cardinal, R. N., Parkinson, J. A., Lachenal, G., Halkerston, K. M., Rudarakanchana, N., Hall, J., Morrison, C. H., Howes, S. R., Robbins, T. W., and Everitt, B. J. (2002). Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behavioral Neuroscience*, 116:553–567.

Carter, A. G., Soler-Llavina, G. J., and Sabatini, B. L. (2007). Timing and location of synaptic inputs determine modes of subthreshold integration in striatal medium spiny neurons. *Journal of Neuroscience*, 27:8967–8977.

Day, J. J., Wheeler, R. A., Roitman, M. F., and Carelli, R. M. (2006). Nucleus accumbens neurons encode Pavlovian approach behaviors: evidence from an autoshaping paradigm. *European Journal of Neuroscience*, 23:1341–1351.

Dayan, P. and Kakade, S. (2001). Explaining away in weight space. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 451–457. MIT Press.

Di Ciano, P., Cardinal, R. N., Cowell, R. A., Little, S. J., and Everitt, B. J. (2001). Differential involvement of NMDA, AMPA/kainate, and dopamine receptors in the nucleus accumbens core in the acquisition and performance of pavlovian approach behavior. *Journal of Neuroscience*, 21:9471–9477.

Drew, M. R., Zupan, B., Cooke, A., Couvillon, P., and Balsam, P. D. (2005). Temporal control of conditioned responding in goldfish. *Journal of Experimental Psychology. Animal Behavior Processes*, 31:31–39.

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

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525:243–246.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience*, 19:479–486.

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience*, 11:966–973.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, 11:e1004567.

Gershman, S. J. (2017). Dopamine, inference, and uncertainty. *Neural Computation*, 29:3311–3326.

Gershman, S. J., Assad, J. A., Datta, S. R., Linderman, S. W., Sabatini, B. L., Uchida, N., and Wilbrecht, L. (2024). Explaining dopamine through prediction errors and beyond. *Nature Neuroscience*, 27:1645–1655.

Gershman, S. J., Moustafa, A. A., and Ludvig, E. A. (2014). Time representation in reinforcement learning models of the basal ganglia. *Frontiers in Computational Neuroscience*, 7:194.

Gottlieb, D. A. (2004). Acquisition with partial and continuous reinforcement in pigeon autoshaping. *Learning & Behavior*, 32:321–334.

Hampton, A. N. and O'Doherty, J. P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. *Proceedings of the National Academy of Sciences*, 104:1377–1382.

Harris, J. A. and Carpenter, J. S. (2011). Response rate and reinforcement rate in Pavlovian conditioning. *Journal of Experimental Psychology. Animal Behavior Processes*, 37:375–384.

Jenkins, W. O. and Stanley, J. C. (1950). Partial reinforcement: a review and critique. *Psychological Bulletin*, 47:193–234.

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, 338:953–956.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45.

Kamin, L. (1968). Attention-like associative processes in classical conditioning. In *Miami symposium on the prediction of behavior: Aversive stimulation*, pages 9–31. University of Miami Press, Miami, FL.

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin*, 79:398–407.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, 20:3034–3054.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & Behavior*, 40:305–319.

Miller, K. D. and Troyer, T. W. (2002). Neural noise can explain expansive, power-law nonlinearities in neural response functions. *Journal of Neurophysiology*, 87:653–659.

Miller, R. R. and Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, 125:370–386.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski,

G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–533.

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

Ottenheimer, D. J., Hjort, M. M., Bowen, A. J., Steinmetz, N. A., and Stuber, G. D. (2023). A stable, distributed code for cue value in mouse cortex during reward learning. *elife*, 12:RP84604.

Parkinson, J. A., Willoughby, P. J., Robbins, T. W., and Everitt, B. J. (2000). Disconnection of the anterior cingulate cortex and nucleus accumbens core impairs Pavlovian approach behavior: Further evidence for limbic cortical–ventral striatopallidal systems. *Behavioral Neuroscience*, 114:42–63.

Pavlov, I. P. (1927). *Conditioned Reflexes*. Oxford University Press.

Pennartz, C., Ameerun, R., Groenewegen, H., and Lopes da Silva, F. (1993). Synaptic plasticity in an in vitro slice preparation of the rat nucleus accumbens. *European Journal of Neuroscience*, 5:107–117.

Rashotte, M. E., Griffin, R. W., and Sisk, C. L. (1977). Second-order conditioning of the pigeon's keypeck. *Animal Learning & Behavior*, 5:25–38.

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

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

Reynolds, J. N. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507–521.

Roitman, M. F., Wheeler, R. A., and Carelli, R. M. (2005). Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. *Neuron*, 45:587–597.

Rothenhoefer, K. M., Hong, T., Alikaya, A., and Stauffer, W. R. (2021). Rare rewards amplify dopamine responses. *Nature Neuroscience*, 24:465–469.

Sadacca, B. F., Wied, H. M., Lopatina, N., Saini, G. K., Nemirovsky, D., and Schoenbaum, G. (2018). Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *Elife*, 7:e30373.

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

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martin-Martin, R., and Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(Volume 8, 2025):153–188.

Tiganj, Z., Jung, M. W., Kim, J., and Howard, M. W. (2017). Sequential firing codes for time in rodent medial prefrontal cortex. *Cerebral Cortex*, 27:5663–5671.

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

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

Yin, H., Barnet, R., and Miller, R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: operational similarities and differences. *Journal of Experimental Psychology: Animal Behavior Processes*, 20(4):419–428.