Ashby, F. G., Crossley, M. J., & Inglis, J. B. (2023).
Mathematical models of human learning.
In F. G. Ashby, H. Colonius, & E. Dzhafarov
(Eds.), *The new handbook of mathematical
psychology, Volume 3* (pp. 163–217).
Cambridge University Press.

Contents

4
Mathematical Models of Human Learning

F. Gregory Ashby[a], Matthew J. Crossley[b], and Jeffrey B. Inglis[c]

4.1 Early Models of Human Learning

Many early learning theories were seeded by the seminal work of Thorndike. In his famous “puzzle box” experiments, Thorndike placed an animal inside a box with a door that could be opened via a latch accessible to the animal. When the animal learned to operate the latch correctly, the door opened, and it was free to consume a reward placed near the box. Thorndike measured the amount of time it took animals to solve such puzzle boxes and found that the escape time tended to decrease with each trial – that is, the animals learned. From these observations, Thorndike (1927) postulated the *law of effect*, which states that behavior is driven by associations between stimuli and responses, and that these associations are strengthened when a response is followed by a satisfying effect and weakened when followed by a discomforting effect. With this, the field of associative learning was born. Already apparent in this early work is its clear connections to modern-day reinforcement learning (RL) theory.

Russian physiologist Pavlov (1927) pioneered one still modern approach to studying associative learning called *classical conditioning*. His famous experiments studying how the salivation response of dogs could be conditioned to occur to a previously neutral stimulus gave the field a standardized paradigm and a new nomenclature (e.g., unconditioned stimulus [US], unconditioned response [UR], conditioned stimulus [CS], and conditioned response [CR]) that drove research in the field forward. Later, Skinner (1938) pioneered many more of the standard methods in use today for the investigation of associative learning. He created operant conditioning chambers – popularly known as the Skinner box – that were equipped with both a manipulandum

[a] University of California, Santa Barbara, USA
[b] Macquarie University, Australia
[c] University of California, Santa Barbara, USA

(e.g., a lever) and a tool to record lever pulls so that cumulative operant behavior (e.g., pulling the lever) could be measured over an experimental session. This approach came to be known as *operant* or *instrumental conditioning* .

Watson and Guthrie followed many of the basic tenets of associative learning formulated by Thorndike, but each introduced novel refinements (e.g., Guthrie 1935; Watson 1913). Unlike Thorndike, neither thought that reinforcement (i.e., neither a satisfying nor discomforting effect) was necessary for associative learning. Rather, they thought that mere temporal contiguity between stimulus and response was sufficient. Later in this chapter, we will see how this notion is related to a form of two-factor synaptic plasticity proposed by Hebb (1949).

An important theoretical alternative to the dominant theories of instrumental conditioning came from Tolman (1948), who advocated that animals learned "cognitive maps" and used these maps to make flexible and goal-directed actions. This view gained relatively little traction in Tolman's lifetime, but is renewed today by modern model-based RL accounts of learning.

The first attempts to formalize theories of learning focused on building mathematical equations that could fit learning curves from a variety of different conditioning experiments (Gulliksen, 1934; Hull, 1943; Thurstone, 1919). The most systematic attempts were by Hull (1943), who embraced Thorndike's fundamental ideas on associative learning – although he spoke of *habits* instead of stimulus-response associations. More importantly, he expressed his views in the form of explicitly stated assumptions. The resulting equations clearly expressed what Hull believed were driving factors of an animal's behavior (e.g., habit strength, drive reduction, etc.).

In the 1950's, mathematical models of learning began to focus less on curve fitting and more on the psychological processes that mediate the learning. This change in focus began with two *Psychological Review* articles on mathematical learning theory that appeared in quick succession – Estes' (1950) introduction of stimulus-sampling theory and Bush and Mosteller's (1951) description of the linear-operator model. Both of these contributions were hugely influential, partly because they were among the first process models in psychology, and as such, they spurred others to develop their own process models. The excitement created by these efforts played a key role in the birth of modern mathematical psychology. But both articles were also influential in their own right. In particular, the linear-operator model inspired the Rescorla-Wagner model (Rescorla & Wagner, 1972), which is now ubiquitous in the learning literature, and more than a half century later, stimulus-sampling theory continues to motivate new research (e.g., Fanselow, Zelikowsky, Perusini, Barrera, and Hersman 2014; Soto and Wasserman 2010).

Mathematical learning theory played a huge role in the field of mathematical psychology during its first formal decade of existence. Stimulus sampling theory and the linear operator model were both elaborated, and a large number of Markov chain models were proposed that assumed learning was a process of moving between discrete states of knowledge. During the 1960s, interest in cognitive processes saw a shift to models of concept learning, which today would be called rule-based learning, and a new focus on the cognitive components of learning, including attention, storage, and retrieval (e.g., Greeno and Bjork 1973). Much of this work is reviewed in the classic text by Atkinson, Bower, and Crothers (1965).

Today, mathematical models of learning are developed and tested in a wide range of different fields, including for example, machine learning (e.g., Alpaydin 2020; Mohri, Rostamizadeh, and Talwalkar 2018), and learning in simple species such as Drosophila (e.g., Kennedy 2019) and zebrafish (e.g., Ninkovic and Bally-Cuif 2006). A review of all this work is outside the scope of any one chapter. Instead, our focus will be on mathematical models of human learning. In some cases, we will consider developments in machine learning and research with non-human animals, but in all cases the focus will be on how such work has contributed to our understanding of human learning.

4.2 Neuroscience Breakthroughs

Mathematical modeling of human learning began to languish in the late 1960s, partly because of the cognitive revolution that turned interest to other phenomena, and partly because it became apparent that the best existing models were valid for only a narrow and limited set of learning-related phenomena. Furthermore, models that succeeded in different domains often bore little similarity to each other. This landscape remained largely unchanged for the next several decades, until two breakthroughs in neuroscience offered a clear path forward. The first was the discovery of long-term potentiation (LTP) and long-term depression (LTD), which served as promising models of learning at the cellular level. The second breakthrough was the discovery that humans have multiple learning and memory systems that for the most part are functionally and anatomically distinct, and that each control behavior under different experimental conditions. As a result, it is likely that no single mathematical model can describe all human learning. Instead, qualitatively different models are needed for different learning systems.

4.2.1 Discovery of LTP and LTD

In his classic 1949 book, entitled "Organization of Behavior: A Neuropsychological Theory," Donald Hebb proposed a neural mechanism that he thought might mediate learning and memory. Specifically, he postulated that

> Let us assume then that the persistence or repetition of a reverberatory activity (or 'trace') tends to induce lasting cellular changes that add to its stability. The assumption can be precisely stated as follows: When an axon of cell A is near enough to excite a cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased (p. 62, Hebb 1949).

Hebb's hypothesis is now widely known as *Hebbian learning*.

Several decades later, this exact type of neural plasticity was discovered at synapses in the hippocampus (Bliss & Lømo, 1973). Specifically, brief, high-frequency presynaptic activation was found to cause a persistent (at least 1 hour) increase in the post-synaptic response – a phenomenon known as *long-term potentiation* (LTP). Then, 9 years later, the opposite phenomenon of *long-term depression* (LTD) was discovered, in which prolonged, but weak presynaptic activation causes a persistent (at least 1 hour) decrease in the postsynaptic response (Ito, Sakurai, & Tongroach, 1982). LTP and LTD have now been observed and closely studied in many different brain regions and in many different cell types. Furthermore, they are known to occur under a plethora of diverse conditions, and to be driven by numerous intracellular signalling cascade mechanisms. Although a review of the current literature on LTP and LTD is well beyond the scope of this chapter, a noncontroversial conclusion of this literature is that it is now widely accepted that LTP and LTD form the neural basis of learning and memory (e.g., S. Martin, Grimwood, and Morris 2000; Nicoll 2017).

4.2.2 Discovery of multiple learning and memory systems

Early mathematical models of learning assumed that all human learning occurs in the same way, which suggests that all learning should depend on the same neural network and be consolidated into the same memory system. This assumption was inconsistent with the growing body of evidence that began to accumulate in the 1960s showing that the best models seemed valid for only a narrow range of experimental tasks, and this led many mathematical psychologists to turn away from the study of learning. A resurgence in mathematical models of learning was ushered in by the discovery that humans have multiple learning and memory systems that for the most part are functionally and anatomically distinct, that evolved at different times and for different purposes, that are ideally suited to learning different types of information, and that thrive under very different environmental conditions.

The first step in this process was to realize that humans have multiple memory systems (e.g., Eichenbaum and Cohen 2001; Poldrack et al. 2001; Squire 2004; Tulving and Craik 2000). After overwhelming evidence in support of multiple memory systems was documented, it was an easy inference to conclude that humans must therefore also have multiple learning systems. After all, learning is the acquisition of a skill or some form of knowledge, and memory is the storage and/or expression of what was learned. So learning and memory are closely related. Mathematical models of learning focus on how the memory traces are established and consolidated, whereas models of memory focus on the nature of those traces and how they are accessed to produce memory-dependent behaviors (e.g., see the chapter by Howard in this volume). For this reason, an obvious hypothesis is that there are as many learning systems as there are memory systems (e.g., Ashby and Maddox 2005; Ashby and O'Brien 2005).

As soon as the multiple systems hypothesis was formulated, work began to identify the networks that mediate learning in each system and to study the properties of the various systems (for a review, see e.g., Ashby and Valentin 2017). This body of research made it clear that no single model was likely to account for all human learning. For example, basal-ganglia-mediated procedural learning is incremental, whereas prefrontal-cortex mediated rule learning is mostly all-or-none (e.g., J. D. Smith and Ell 2015).

4.3 Modern Approaches to Modeling Human Learning

The birth of mathematical psychology coincided with the first attempts to build process models of learning. The reinterest in learning that occurred with the neuroscience breakthroughs described in the previous section coincided with the development of new types of learning models, and also with the first ever implementational-level models – that is, models that attempt to describe the neural circuits that implement the algorithms described by process models. This section briefly introduces these more modern approaches to building mathematical models of learning, and then the rest of the chapter examines these trends in more detail.

4.3.1 Descriptive- and process-level approaches

Current descriptive and process models of human learning are dominated by two different, but converging approaches – one rooted in the statistics literature and one rooted in the machine-learning and computer-science literatures (as described, e.g., by Alpaydin 2020; Sutton and Barto 1998). Both attempt to build models that optimize some aspect of learning – the former by following principles of Bayesian statistics, and the latter by assuming that human learning depends on some popular machine-learning algorithms.

Normative models have a long history in psychology. For example, ideal observer models have played an important role in psychophysics and signal detection theory since the 1950s (e.g., Green and Swets 1966). Similarly, during the 1980s and 1990s, human classification performance was carefully compared to the performance of optimal classifiers (e.g., Ashby and Alfonso-Reese 1995; Ashby and Maddox 1998). Comparing human performance to the performance of an optimal device is a valuable step in the evolution of model building in any area of psychology. Humans are highly skilled in many behaviors, so an optimal model will often provide a reasonably good fit to human data. Better fits are usually possible by adding certain suboptimal components to the model, such as various types of noise. Carefully documenting which types of added suboptimalities allow the model to provide the best fit provides invaluable information about the underlying processes that mediate the behavior. In the case of human learning, the Bayesian models are objectively optimal, in the sense that they assume the learner chooses the response most likely to be correct, and that these choice probabilities are updated trial-by-trial according to Bayes theorem. In this class of models, learning is typically equivalent to Bayesian updating.

An alternative approach, which is perhaps even more popular and that looks very different on the surface, is to build models that assume human learning follows algorithms that were developed in the computer-science, machine-learning, and artificial-intelligence literatures. In this approach, the models are typically some form of neural network, and learning is a process of adjusting the connection strengths or weights between units. These algorithms fall into one of three general classes: unsupervised, RL, and supervised (e.g., Alpaydin 2020). Unsupervised learning algorithms, which include Hebbian learning as a prominent special case, modify all learning-related weights using the same algorithm and without regard to feedback. RL algorithms also apply the same learning algorithm to every weight, but the algorithm applied depends on the type of feedback that was delivered (e.g., reward versus non-reward). Finally, supervised learning algorithms attempt to compute the unique error of the output unit associated with every modifiable weight in the network, and they then tune that weight according to this unique error. The most prominent examples, such as backpropagation and the delta rule, attempt to implement a gradient descent optimization procedure. Unsupervised learning and RL are *global learning rules* because they apply the same rule to every learning-related weight, whereas supervised learning is a *local learning rule* because it uses a different error to modify every weight.

Early models imported from computer science assumed that learning followed gradient descent trajectories, as implemented for example, by the delta rule and backpropagation. More recently, a large subset of these models apply one of the many RL algorithms that are described in the influential text of Sutton and Barto (1998). Included in this list are temporal-difference learning, actor-critic architectures, Q learning, and SARSA (State-Action-Reward-State-Action).

The models in this class are not objectively optimal, at least not in the sense of the Bayesian models, which try to maximize response accuracy. Even so, the learning algorithms they assume were all developed in attempts to maximize the learning abilities of some artificial system. Therefore, if not objectively optimal, many of them are among the most efficacious learning algorithms ever invented. In this sense, models in this class are similar to the normative models that are constructed using Bayesian statistics.

4.3.2 Implementational-level approaches

Implementational-level models require extensive knowledge about brain function and behavior. Despite this high standard, they date back at least to early work by Marr (e.g., Marr 1969) and Grossberg (e.g., Grossberg 1972). One remarkable aspect of these early models is that they predate the discovery of the forms of synaptic plasticity that they postulated. Despite this early and seminal work, until recently, there were relatively few implementational-level models in psychology.

During the past two decades, the field of neuroscience has exploded, and the number implementational-level models in psychology has grown commensurately. As these models became more popular, new methods were developed to build and test them, and collectively this new field is known as *computational cognitive neuroscience* (Ashby, 2018; O'Reilly, Munakata, Frank, Hazy, et al., 2012). The goal here is to first identify the neural network that mediates the behavior and then build a model that mimics neural activity in this network. In the case of learning, the model should display learning-related synaptic plasticity in accord with what is observed in the biological system being modeled. Such models are generally more computationally intractable than their process counterparts, and therefore require extensive computer simulation to fit and test. Even so, despite this cost, implementational models have many advantages over more traditional process models (e.g., see Ashby 2018). For example, whereas process models can generally be tested only against response time and accuracy data, computational cognitive neuroscience models can be tested against virtually any dependent measure between behavior at the highest level and single-unit recordings at the lowest level, including for example, response times, accuracies, single-neuron recordings, fMRI blood oxygen-level dependent (BOLD) responses, and EEG recordings. Another advantage is that if two computational cognitive neuroscience models are built and validated that each account for different types of behaviors, then because each should be faithful to the underlying neuroanatomy, it should be possible to link the two in a plug-and-play fashion to create a new composite model that is consistent with all the behavioral and neuroscience data that are consistent with either model alone (as done e.g., by Cantwell, Riesenhuber, Roeder, and Ashby 2017).

Implementational models attempt to model activity in the actual neural circuits that mediate the behavior under study. And rather than borrow learning algorithms from Bayesian statistics or machine learning, they directly model the types of synaptic plasticity thought to occur during LTP and LTD. Thus, whereas implementational models directly model the structures and processes thought to mediate learning, Bayesian models and models based on machine-learning notions of RL are examples of modeling by analogy – in the sense that they are based on algorithms developed for other purposes (statistics or machine learning). Modeling by analogy has a long history in psychology ('the brain is like a telephone switchboard'; 'the brain is like a computer'), and comparing human behavior to other devices can be a useful exercise because it can expose uniquely human characteristics. Implementational models should be the ultimate goal, but they require far more knowledge to build, and for many behavioral phenomena, this high threshold has not yet been reached.

Whereas it was always obvious that 'the brain is like a telephone switchboard' is an analogy, as the analogies became more sophisticated, they also became more difficult to recognize. This is especially true with models based on machine-learning RL algorithms. After all, RL has been a central focus of research within psychology for more than a century. Furthermore, it was quickly noted that synaptic plasticity in the striatum has properties that are similar to several popular machine-learning RL algorithms (e.g., Doya

2000; Houk, Adams, and Barto 1995). Because of this similarity, learning models based on machine-learning notions of RL can be especially useful. Ultimately though, we should expect that synaptic plasticity, and therefore learning, will have some uniquely human properties that require their own uniquely human models to capture completely.

4.4 Descriptive and Process Models of Human Learning

4.4.1 Reinforcement learning

In computer-science, RL is a general approach to building decision-making agents that learn to maximize rewards. The standard approach (Sutton & Barto, 1998) is to model the environment as a Markov decision process and to assume that the agent moves through a set of discrete states $S = \{s_1, s_2, \ldots, s_n\}$ by choosing among a set of possible actions $A = \{a_1, a_2, \ldots, a_m\}$. The decision rule that determines the probability of each possible action, given a particular state is called the action policy π.

A state can be almost anything. The one requirement is that since we assume a Markov decision process, the states, as defined by the model, must satisfy the Markov property, in the sense that knowledge of the current state alone should be enough to compute the predicted probability of reward and this probability should not depend on the path the agent took to reach the current state. So for example, a state could be the position of a rat in a maze. If the rat is in an arm that is not baited with reward, then the probability of imminent reward is low, whereas if the animal is in a baited arm, then the probability of imminent reward is high.

The action policy π determines the probability that the state will change from s_i to s_k, for any i and k. Since each state has some true probability of imminent or future reward that is determined by the environment, the actions selected by the agent therefore also determine current and future reward probabilities. Thus, the agent must learn to take actions that cause transitions to the most rewarding states. This requires that the agent learns the value of each state. Value is formalized in the state-value function $V_\pi(s)$, which equals the expected value of all rewards – both current and future – that the agent can expect if the state is s and future actions are selected according to policy π.[1] Let r_t denote the value of a reward received t time units in the future, and let R denote the total value of all current and future

[1] Sutton and Barto (1998) define the value function as the expected value of all future rewards. Therefore, in their formulation, the current reward does not contribute to the value function. This definition implies that the value to an animal of reaching a baited goal box when exploring a maze is zero. For this reason, we chose to define the value function as the expected value of all current and future rewards.

rewards. Then RL models assume

$$R = \sum_{t=0}^{\infty} \gamma^t r_t, \tag{4.1}$$

where $0 < \gamma \leq 1$ is a temporal discounting parameter that serves to reduce the value of rewards the more distant they occur in the future. The value function is then defined as

$$\begin{aligned} V_\pi(s) &= \mathrm{E}[R|\pi, s] \\ &= \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s\right]. \end{aligned} \tag{4.2}$$

Different methods use the estimated value function in different ways to select the best actions, and a complete description of all these methods is beyond the scope of this chapter. However, one major dimension on which different methods are classified is whether or not the agent directly estimates the state transition probabilities (i.e., the probability that the state will transition from s_i to s_k when action a_j is selected). Methods that estimate these transition probabilities are called *model-based*, whereas methods that do not are called *model-free*.

Model-free RL approaches

The iterative sample mean. As we have seen, the goal of many RL models and algorithms is to estimate a state-value function. For example, the Rescorla-Wagner model estimates the expected reward value of a cue in a classical conditioning paradigm, temporal-difference learning estimates the expected value of all future rewards given some fixed action policy, and Q learning estimates a similar value for different state-action pairs. The standard statistical approach to parameter estimation assumes a sample of fixed size. RL algorithms however, apply to an agent operating in real time through an environment that presents successive opportunities to receive rewards. Therefore, the agent must continually update value estimates when moving through the environment. For this reason, parameter estimation must be iterative (e.g., as in dynamic programming). This is a straightforward and well-known statistical problem. For example, a population mean can be estimated iteratively as follows.

Theorem 4.1 *Consider a set of successive samples $X_1, X_2, ..., X_n$ that are* *all drawn from some population. Then the sample mean equals*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_{n-1} + \frac{1}{n}(X_n - \overline{X}_{n-1}), \tag{4.3}$$

where $\overline{X}_0 = 0$. Furthermore, note that X_n is the current sample and $\overline{X}_{n-1}$ is the best guess of X_n after $n-1$ samples have been collected (i.e., in the sense of the law of large numbers). As a result $X_n - \overline{X}_{n-1}$ is the prediction error – call it PE. So Eq. 4.3 is equivalent to:

$$\overline{X}_n = \overline{X}_{n-1} + \frac{1}{n}PE. \tag{4.4}$$

Proof See for example, Ashby (2018). □

In other words, the standard, batch estimate of the population mean, $\overline{X}_n$, can be efficiently computed in real time by updating the old estimate by an amount that is proportional to the prediction error. If the newest sample is larger than expected (i.e., if $X_n > \overline{X}_{n-1}$) then the mean estimate is increased, and if the newest sample is smaller than expected (i.e., if $X_n < \overline{X}_{n-1}$) then the mean estimate is decreased.

As we will see, the most popular RL algorithms are all based on Eq. 4.4. They differ mainly in how they define X_n, although in all RL algorithms the goal is to estimate some reward-related value. In such cases, the prediction error in Eq. 4.4 becomes a *reward prediction error* (RPE), which in general, is defined as obtained reward minus expected reward.

Because the iterative estimate of the mean is mathematically equivalent to the standard, batch estimate, it possesses the same statistical properties. Therefore, note that this iterative estimate is the uniformly minimum variance unbiased estimator of the population mean if the X_i are independent, and identically distributed (iid) samples from some population, and the sample size n is known ahead of time. In many real-world environments of course, the samples are not iid, and if the sampling is done in real time, the final sample size is often unknown. The standard RL solution to these problems is to replace the constant $1/n$ with some constant α that can be adjusted or set in a way that depends on the environment. For example, a standard approach is to set α in a way that causes temporal discounting, so that recent samples are weighted more heavily than early samples.[2] In fact, this form of temporal discounting occurs whenever $\alpha > 1/n$. Therefore,

[2] Note that we have now introduced two different temporal discounting parameters. The parameter γ discounts future rewards and the parameter α discounts distant samples.

when applied to nonstationary data, the *iterative sample mean* equals:

$$\overline{X}_n = \overline{X}_{n-1} + \alpha(X_n - \overline{X}_{n-1}) = \overline{X}_{n-1} + \alpha PE. \tag{4.5}$$

The parameter α is commonly referred to as the learning rate because increases in α cause $\overline{X}_n$ to change more quickly.

Another advantage of the iterative sample mean, relative to the batch estimate, is that it is easier to incorporate prior beliefs into the estimate of the population mean. For example, consider a simple coin-tossing experiment in which the goal is to estimate the probability of a heads (i.e., where we assign a value of 1 to each heads and 0 to each tails, and then use Eq. 4.5 to estimate the true probability of a heads). A natural prior belief might be that the coin is fair, which is easily incorporated into Eq. 4.5 by setting $\overline{X}_0 = .5$.

Temporal-Difference Learning. Temporal-difference learning estimates the state-value function under the assumption that the action policy is fixed. A popular paradigm that satisfies this constraint is classical conditioning, in which some cue may or may not be followed some time later by a reward or perhaps by multiple rewards. The goal of the agent in this case, is to learn that the cue predicts a future reward. Note that there is no action to produce here and so in this special case, we can omit the subscript π in our notation. And although temporal-difference learning applications to classical conditioning are free to define the states in any way that satisfies the Markov property, the most common definition, by far, is to define the states as time points that begin with the cue and end with the last possible reward.

Therefore, define the state $s_t = t$, where t equals the number of time steps since cue presentation, and let $r_n(t)$ equal the value of the reward received at time t on trial n. Then the total value of all future rewards on trial n at time t equals

$$R_n(t) = \sum_{i=t}^{T} \gamma^{i-t} r_n(i), \tag{4.6}$$

where T equals the time of the last possible reward on each trial. As in other RL algorithms, the goal of temporal-difference learning is to estimate the state-value function – that is, the expected value of $R_n(t)$:

$$V_n(t) = \mathrm{E}[R_n(t)]. \tag{4.7}$$

Now because $V_n(t)$ is a population mean, our best estimate is the sample mean of the $R_i(t)$ across previous trials:

$$\hat{V}_n(t) = \frac{1}{n}\sum_{j=1}^{n} R_j(t). \tag{4.8}$$

This sample mean can be estimated efficiently via term-by-term substitution into the iterative sample mean defined in Eq. 4.5 to produce

$$\hat{V}_n(t) = \hat{V}_{n-1}(t) + \alpha[R_n(t) - \hat{V}_{n-1}(t)]. \tag{4.9}$$

The problem with this estimate is that $R_n(t)$ includes the immediate reward $r_n(t)$, plus all future rewards that will be obtained on trial n – that is,

$$\begin{aligned} R_n(t) &= r_n(t) + \sum_{i=t+1}^{T} \gamma^{i-t} r_n(i) \\ &= r_n(t) + \gamma \sum_{i=t+1}^{T} \gamma^{i-(t+1)} r_n(i) \end{aligned} \tag{4.10}$$

and unfortunately, all reward-related terms on the right except $r_n(t)$ are unknowable since they occur in the future. Temporal-difference learning estimates the unknowable part – that is, the expression defined by the summation sign – by using the iterative sample mean of all rewards that occurred after time t on previous trials [i.e., via $\hat{V}_{n-1}(t+1)$]. This results in the following estimate:

$$\hat{R}_n(t) = r_n(t) + \gamma\hat{V}_{n-1}(t+1). \tag{4.11}$$

Substituting this estimate into Eq. 4.9 for $R_n(t)$ produces the final form of temporal-difference learning:

$$\hat{V}_n(t) = \hat{V}_{n-1}(t) + \alpha[r_n(t) + \gamma\hat{V}_{n-1}(t+1) - \hat{V}_{n-1}(t)]. \tag{4.12}$$

Note that, despite initial appearances, the expression in square brackets equals the prediction error (or more specifically, the RPE), just as in Eq. 4.5. The first term in the square brackets is the immediately obtained reward and the second term is the best guess of the (discounted) value of all future rewards expected on trial n. The sum of the first two terms is therefore the agent's estimate of the total obtained rewards on trial n given that we are t time units into the trial. The last term is the predicted value of this quantity that was made before the trial began. Therefore, the sum of the first two terms represents obtained reward, whereas the last term represents predicted reward.

As an application of temporal-difference learning, consider a simple classical-conditioning task in which the same CS (e.g., a light or tone) is followed T time steps later by a reward. On the first presentation of the CS, the subsequent reward is unexpected, but as the CS-reward pairing is repeated, the agent will eventually learn that the CS is paired with future reward. The standard temporal-difference learning application to this task assumes that initially, all states have zero value [i.e., $V_0(t) = 0$, for all t] because the CS has never before been paired with reward. On trial 1, the CS is presented and then the agent unexpectedly receives a reward at time T. Suppose the value of this reward is r. Then temporal-difference learning predicts that

$$\hat{V}_1(T) = \hat{V}_0(T) + \alpha[r_1(T) + \gamma\hat{V}_0(T+1) - \hat{V}_0(T)]. \tag{4.13}$$

Note that, by our assumptions about initial conditions, all V_0 terms equal 0. However, $r_1(T) = r$ because the agent receives a reward on each trial at time T. Therefore,

$$\hat{V}_1(T) = \alpha r. \tag{4.14}$$

Now consider the value that temporal-difference learning assigns to the state that is one-time unit earlier than reward delivery on trial 2:

$$\hat{V}_2(T-1) = \hat{V}_1(T-1) + \alpha[r_2(T-1) + \gamma\hat{V}_1(T) - \hat{V}_1(T-1)]. \tag{4.15}$$

Note that $\hat{V}_1(T-1) = 0$ because at time $T-1$ of trial 1 the agent has not yet received any rewards. Furthermore $r_2(T-1) = 0$ because rewards are delivered at time T, not at time $T-1$. However, as we saw, $\hat{V}_1(T) = \alpha r$. Therefore,

$$\hat{V}_2(T-1) = \alpha^2\gamma r. \tag{4.16}$$

In other words, the RPE that occurred at time T on trial 1 has propagated back on trial 2 to the immediately preceding state (i.e., $T-1$). Similarly, on trial 3, the positive value associated with state $T-1$ will propagate back to state $T-2$. In this way, the value associated with earlier and earlier states will increase. This propagation will continue until it eventually reaches the time of cue presentation – that is, until $V_n(0) > 0$, for some value of n. It will not propagate to earlier times than this however, so long as cue presentation times are unpredictable.

Temporal-difference learning is popular, in part because it shares some properties with the firing properties of dopamine (DA) neurons. In particular, in this same classical-conditioning experiment, DA neurons will eventually begin to fire to any cue that predicts a future reward. We will consider temporal-difference learning as a model of DA neuron firing in more detail in a later section.

Q-learning. Q-learning is a model-free RL algorithm to learn actions that maximize current and future rewards. It is similar to temporal-difference learning, but it learns the value of state-action pairs, instead of states independently of the selected action. The resulting value function, denoted by $Q_n(s, a)$ (i.e., "Q" for quality), gives the value of taking action a from state s on trial n, under the assumption that all actions after a are optimal with respect to the estimated action-value function – that is, that all future actions are selected so as to maximize total reward. The policy that always chooses the action that maximizes reward is called the *greedy policy*. So Q learning updates the value function under the assumption that a greedy policy will be used, even when the agent follows some non-greedy policy. Algorithms that estimate the value function using a policy that is different from the one that is currently being followed are called *off-policy* algorithms.

Let s_t denote the state at time t and a_t denote the action taken at time t. Let $R_n(a_t|s_t)$ denote the total current and future rewards obtained on trial n if the state is s_t, action a_t is immediately taken, and all subsequent actions are greedy. Then term-by-term substitution into the iterative sample mean produces:

$$\hat{Q}_n(s_t, a_t) = \hat{Q}_{n-1}(s_t, a_t) + \alpha[R_n(a_t|s_t) - \hat{Q}_{n-1}(s_t, a_t)]. \tag{4.17}$$

As with temporal-difference learning, R_n is unknowable since it depends on future rewards. So Q learning estimates R_n via:

$$\hat{R}_n(a_t|s_t) = r_n(t) + \gamma \max_a \hat{Q}_{n-1}(s_{t+1}, a); \tag{4.18}$$

that is, by adding the immediate reward to the discounted (iterative sample) mean of the best rewards produced by any sequence of past actions that started from the state that results from taking action a_t from state s_t. Combining these two equations produces the final form of Q learning:

$$\hat{Q}_n(s_t, a_t) = \hat{Q}_{n-1}(s_t, a_t) + \alpha \left\{ r_n(t) + \gamma[\max_a \hat{Q}_{n-1}(s_{t+1}, a)] - \hat{Q}_{n-1}(s_t, a_t) \right\}. \tag{4.19}$$

A common assumption is that all initial Q values equal 0 (i.e., $Q_0(s, a) = 0$, for all s and a). Once these initial values are all set, Eq. 4.19 is used to update the Q estimates beginning on trial 1.

The Q values are often used to define action policies. For example, as we already saw, the greedy action policy is to always choose the action with the maximum Q value. Although at first glance, this policy sounds appealing, note that it fails to explore the set of all possible actions. Unless the optimal policy is discovered early on by chance, then the greedy policy is unlikely to ever discover this optimal policy. Therefore, many action policies trade off exploration and exploitation. One way to do this is via an ϵ-greedy algorithm that selects an action randomly with probability ϵ and uses a greedy policy with probability $1 - \epsilon$. Another popular choice is to compute the action selection probabilities by passing the Q values through a softmax function:

$$P(a_i|s) = \frac{e^{Q(s,a_i)}}{\sum_j e^{Q(s,a_j)}}. \tag{4.20}$$

Note that since this policy depends on the Q values, updating or changing the estimates of Q changes the policy.

To make the discussion concrete, consider an agent in a maze in which one or more arms are baited with reward. In this case, we can consider the states to be locations within the maze, and actions to be movements that carry the agent from one location to another. Before any learning has occurred, if state s_i is one step before a baited arm, then the action that moves the animal one step forward (i.e., towards the reward) will be rewarded and $Q(s_i, a_{\text{forward}})$ will gain positive value.

Model-based RL approaches

Model-based RL approaches build a model of the environment by estimating the value function [e.g., $V_\pi(s)$ or $Q(s,a)$], and the state-transition function $T(s_k|a_j, s_i)$, which specifies the probability that taking action a_j will transition the agent from state s_i to state s_k. Once accurate estimates of these functions are available, action selection is a matter of solving directly for the action sequence that maximizes reward.

Daw, Niv, and Dayan (2005) proposed a dual-controller model that assumes the brain includes both model-free and model-based RL algorithms, with behavior determined by the system that is most confident in its predictions (i.e., has the lowest uncertainty). The model includes a striatal-mediated, model-free cache system that implements habit learning and a prefrontal-cortex-mediated model-based tree-search system that implements goal-directed learning. Both systems use a form of Q learning. Traditional Q learning (i.e., Eq. 4.19) does not track uncertainty, so Daw et al. (2005) proposed a Bayesian version (i.e., based on Dearden, Friedman, and Russell 1998), in which both systems attempt to estimate a distribution of Q values across trials. If we assume that rewards are Bernoulli distributed, then a convenient prior distribution of Q values is the beta distribution (because the beta distribution is a conjugate prior for the Bernoulli distribution).

This prior is then updated through Bayes rule to obtain model-free and model-based posterior distributions of Q values.

The model-based system consists of a Bayesian tree-search algorithm in which the agent uses experience in its environment to estimate the distribution of reward values $R(a_j|s_i)$ and state-transition probabilities $T(s_k|a_j, s_i)$. The model assumes a beta distribution for the prior on rewards and a Dirichlet distribution for the prior on the state transitions. These distributions are then updated according to Bayes rule by counting up the obtained rewards and state transitions. Since both systems estimate distributions of Q values, the variances from these two estimates are compared and the system with the lowest uncertainty controls the response of the agent.

The model successfully accounts for a variety of instrumental conditioning phenomena, including for example, the effects of reward devaluation (Dickinson & Balleine, 2002). In these experiments, an animal is trained to lever press (for example) and at some point in training, the reward is devalued prior to the session, typically either by providing free access to food or administering a drug that causes ingestion of the reward to induce illness. Early in training, reward devaluation reduces the frequency of the instrumental behavior, but after extensive overtraining, the behavior becomes immune to the devaluation. Furthermore, the degree to which the animal is sensitive to devaluation is proportional to the complexity of the task and the temporal proximity of the action to the reward. The dual-controller model accounts for these phenomena by proposing that the model-based tree-search system controls responding early in training, but that control is passed to the model-free cache system after overtraining. Furthermore, the amount of training required for the transfer of control is assumed to increase with the complexity of the task.

Early in training, new information immediately influences action values at all states in the tree-search system. In contrast, the cache system takes significantly longer to propagate new information to other states. Additionally, in more complex tasks, the tree-search system takes control because it is more data efficient – in more complex tasks there is less data available for each state-action pair. Finally, the tree-search system performs better for actions more proximate to reward due to its superior data efficiency, whereas the cache system performs better for actions more distal to reward due its lower sensitivity to computational noise. Since the tree-search system has a model of the task and it has access to the long-term consequences of its actions at each time step, it can adapt its policy in response to reward devaluation. Alternatively, the cache system estimates the value of each action directly, and reward devaluation is insufficient to reverse the cumulative effects of the many positive rewards that were received earlier in training.

4.4.2 Bayesian modeling of human learning under uncertainty

The RL models described in the previous section are arguably more popular than Bayesian models of learning, at least partly because they are computationally simpler to implement. Traditional Bayesian approaches require numerical evaluation of complex multiple integrals. This section reviews the hierarchical Gaussian filter (Mathys, Daunizeau, Friston, & Stephan, 2011; Mathys et al., 2014), which attempts to overcome this limitation by deriving computationally simple, interpretable, and efficient update equations – similar to those used in RL models – except from normative Bayesian principles. Conveniently, these update equations also enable the estimation of agent-specific parameters that allow each individual to be modeled as subjectively optimal with respect to minimizing the agent's surprise (i.e., free energy) when unexpected events occur.[3] Furthermore, the form of these update equations is similar to a version of the iterative sample mean (Eq. 4.5) in which the learning rate, α, is modulated by various forms of uncertainty. Accordingly, the benefits of the hierarchical Gaussian filter extend past the Bayesian framework by providing a normative foundation for the sequential updating equations of heuristic RL algorithms.

As a context for describing the model, consider an (A, not A) categorization task in which an agent is asked to report whether or not a presented stimulus belongs to category A (e.g., by responding YES or NO). Suppose the stimuli in category A vary on one stimulus dimension, call it w, and are normally distributed on this dimension with mean μ_A and variance π_A^{-1}; that is

$$w \sim \mathcal{N}(\mu_A, \pi_A^{-1}), \tag{4.21}$$

where π_A is the precision of the category A samples (not to be confused with action policies as defined in the RL literature). Suppose that on "not A" trials, the stimuli are uniformly distributed on dimension w over all physically realizable values. Therefore, the optimal decision strategy is to respond YES if the presented stimulus is close to μ_A on dimension w and NO if it is far away. Consider the simplest possible case in which the agent

[3] See Ashby (2019), Friston, Mattout, Trujillo-Barreto, Ashburner, and Penny (2007), or Penny (2012) for a description of free energy minimization in the context of model selection, and Friston (2010) for a description of free energy minimization as a general principle of brain function.

knows π_A but not μ_A. Then the optimal strategy requires the agent to estimate μ_A.

An agent trying to estimate μ_A could do so by computing the iterative sample mean (Eq. 4.5). Instead, however, consider a Bayesian approach. Suppose the agent assumes that the prior distribution of μ_A is

$$\mu_A \sim \mathcal{N}(\mu_0, \pi_0^{-1}), \tag{4.22}$$

where π_0 is the precision of the agent's knowledge of the task. Suppose further that on trial n the stimulus has value w_n on the relevant dimension and the feedback informs the agent that this stimulus belonged to category A. Then a Bayesian approach indicates that the posterior likelihood that the true orientation is μ_A equals:

$$p(\mu_A|w_n) = \frac{p(w_n|\mu_A)p(\mu_A)}{\int p(w_n|\mu)p(\mu)d\mu} \sim \mathcal{N}(\mu_{\mu_{A|w_n}}, \pi^{-1}_{\mu_{A|w_n}}). \tag{4.23}$$

Equation 4.23 illustrates the traditional problem of Bayesian approaches – the integral in the denominator is often computationally intractable. As a model of human learning, Eq. 4.23 would be more attractive if it included a plausible hypothesis about how humans could approximate such integrals sequentially in real time. As a start, it turns out that the posterior mean and precision can be rewritten as (e.g., see Kruschke 2011, for a derivation):

$$\mu_{\mu_{A|w_n}} = \mu_{\mu_{A|w_{n-1}}} + \frac{\pi_A}{\pi_0 + \pi_A}(w_n - \mu_{\mu_{A|w_{n-1}}}) \tag{4.24}$$

and

$$\pi_{\mu_{A|w_n}} = \pi_0 + \pi_A. \tag{4.25}$$

Note that Eq. 4.24 is in the same form as the iterative sample mean (Eq. 4.5), except that α is replaced with the ratio of the precision of the category A samples, π_A, to the sum of the category A precision plus the agent's precision about the task, π_0. Therefore, if category A precision is low (relative to π_0), then the learning rate is small. This makes sense intuitively – if we trust our model of the environment (i.e., π_0 is large) then we should be conservative about updating that model on the basis of noisy observations. On the other hand, if we have poor knowledge about the environment (i.e., π_0 is small) and there is not much variation in the samples (i.e., π_A is large) then we should use those samples to rapidly update our model of the environment.

This Bayesian formulation is beneficial for ensuring that prediction errors are precision-weighted according to their informativeness in a stable environment. However, this formulation will perform poorly in a non-stationary environment because the learning rate will not adapt to the environmental changes. For example, suppose the experimenter periodically changes the category A mean. The hierarchical Gaussian filter provides an efficient method for adapting to such changes in the environment by iteratively adjusting its estimate of μ_A using a variational Bayesian procedure (Mathys et al., 2011, 2014).

The hierarchical Gaussian filter estimates μ_A in a hierarchical fashion. Let $x_1(n)$ denote the current estimate of μ_A on trial n [i.e., so on trial n, $\hat{\mu}_A = x_1(n)$]. Then the agent's model of category A on trial n is that

$$w_n \sim \mathcal{N}[x_1(n),\, \pi_A^{-1}], \tag{4.26}$$

since again, we are considering the simple case where π_A is known. This is the lowest level of the hierarchy. The next level up (i.e., level 2) estimates the distribution of the mean, $x_1(n)$. Specifically, The hierarchical Gaussian filter assumes that

$$x_1(n) \sim \mathcal{N}\left\{x_1(n-1), \quad \exp[\kappa_1 x_2(n-1) + \omega_1]\right\}, \tag{4.27}$$

where κ_1 and ω_1 are constants, with $\kappa_1 > 0$, and $x_2(n)$ is a new random variable. The exponential function was chosen as a mathematically convenient form via which to estimate the variance of $x_1(n)$ (see, e.g., Mathys et al. 2014). Because $\kappa_1 > 0$, note that the variance of $x_1(n)$ increases with $x_2(n)$. The standard deviation of $x_1(n)$ is often referred to as volatility (Behrens, Woolrich, Walton, & Rushworth, 2007; Bland & Schaefer, 2012; Nassar, Wilson, Heasly, & Gold, 2010; Payzan-LeNestour & Bossaerts, 2011; R. C. Wilson, Nassar, & Gold, 2013), so $x_2(n)$ increases with volatility.

Level 3 of the hierarchy estimates the variance of $x_1(n)$ by assuming that

$$x_2(n) \sim \mathcal{N}\left\{x_2(n-1), \quad \exp[\kappa_2 x_3(n-1) + \omega_2]\right\}, \tag{4.28}$$

where $x_3(n)$ is a new random variable that increases with the variance of volatility. In other words, $x_3(n)$ is measuring how much volatility is changing in the environment. In principle, this hierarchy can continue indefinitely. At each new level, the variance is defined in terms of a new random variable that is itself defined at the next higher level. So for example, level 4 would define $x_3(n)$ as normally distributed with a variance that depends on a new random variable $x_4(n-1)$.

Another critical feature of the hierarchical Gaussian filter is that it specifies trial-by-trial update equations for the mean and precision parameters at each level of the hierarchy. These updates, which were all derived using a variational Bayesian procedure, are in the same general form as Eq. 4.24, with the notable exception that the learning rates [i.e., the analogue of $\pi_A/(\pi_0 + \pi_A)$ in Eq. 4.24] are sensitive to changes in the environment, including for example, volatility and changes in volatility. For example, the update equations specify that when the environment becomes more volatile, the learning rate on μ_A increases. This makes sense intuitively because in a more volatile environment, deviations from our expectations may indicate that environmental events driving our sensory data have changed and learning should therefore proceed more rapidly.

The hierarchical Gaussian filter update equations enable real-time estimation of states and are optimal in the sense that they minimize variational free energy – an upper bound on an agent's surprise given its model of the world. The hierarchical Gaussian filter has a number of advantages over more traditional Bayesian models. First, it avoids the need to evaluate intractable integrals. Second, by placing different subject-specific priors on the κ_i and ω_i parameters, it provides a convenient method for modeling individual differences across agents. Third, it provides a foundation for RL-style update equations and firmly grounds RL models within the foundations of probability theory. Finally, the hierarchical Gaussian filter has also had considerable success at accounting for a wide variety of empirical phenomena, including impulsivity in healthy individuals (Paliwal, Petzschner, Schmitz, Tittgemeyer, & Stephan, 2014) and Parkinson's patients with deep brain implants (Paliwal et al., 2018), reward-based decision making in schizophrenia (Deserno et al., 2020), social learning (Diaconescu et al., 2017), perceptual learning (Weilnhammer, Stuke, Sterzer, & Schmack, 2018), and sensory learning (Iglesias et al., 2013).

4.4.3 Supervised-learning models of sensorimotor adaptation

Models based on supervised learning are also popular. As described above, supervised learning is a local learning rule that uniquely changes each modifiable weight or connection strength in the model. The most popular versions, which include backpropagation and the delta rule, implement a form of gradient descent (e.g., Rumelhart and McClelland 1986). Consider a general model in which some unit i projects to some unit j. Let x_i denote the output of unit i, y_j denote the output of unit j, and denote the connection strength between units i and j by the parameter $\omega_{i,j}$. Then supervised learning algorithms change each $\omega_{i,j}$ differently. The most common approach, which is followed for example by gradient descent algorithms, is to modify $\omega_{i,j}$ according to the error between the desired output y_j^* of unit j and the observed output y_j. This error is typically referred to as $\delta_j = y_j^* - y_j$.

Gradient descent algorithms modify $\omega_{i,j}$ in a way that causes δ_j to decrease as quickly as possible at each time step. Specifically, if we let F represent the mathematical transformation that unit j performs on its input, then $y_j = F(x_i \mid \omega_{i,j})$. Gradient descent algorithms modify $\omega_{i,j}$ according to

$$\Delta\omega_{i,j} \propto -\frac{\partial \delta_j}{\partial \omega_{i,j}}, \tag{4.29}$$

that is, in proportion to the negative of the gradient on the error surface. Here, we can see that the key feature of a supervised learning system is that (1) the system is provided a teaching signal in the form of a desired output, and (2) the error signal (i.e., the difference between actual and desired output) is differentiable with respect to the parameters of the model. Equation 4.29 describes a local learning rule because every output unit in the model has its own unique desired output. Because of this, in response to an error signal at time t, some parameters will be increased, and others will be decreased. This property also strongly distinguishes supervised learning from RL, in which all active weights are either strengthened or weakened in accord with the presence or absence of unexpected rewards.

One prominent class of supervised-learning models uses linear dynamical systems to model the sensorimotor learning that causes adaptive changes in motor outputs in response to changing sensory inputs (Baddeley, Ingram, & Miall, 2003; Cheng & Sabes, 2006; Donchin, Francis, & Shadmehr, 2003; Scheidt, Dingwell, & Mussa-Ivaldi, 2001; Thoroughman & Shadmehr, 2000). Such changes are essential for coordinated and efficient execution of action selection and motor control. For example, as muscles are fatigued they require greater neural impulses to be activated, and therefore the motor commands that achieve some goal before muscle fatigue need to be scaled up to achieve that same goal after fatigue has accumulated. Sensorimotor learning also allows agents to adjust for noisy and dynamic environments. For example, the brakes on a rental car only feel foreign and jerky for a short while before we adapt our motor commands to smoothly operate them.

In the lab, sensorimotor learning is commonly studied with visuomotor adaptation experiments (Cunningham, 1989; Krakauer, Pine, Ghilardi, & Ghez, 2000; T. A. Martin, Keating, Goodkin, Bastian, & Thach, 1996a, 1996b; Redding, Rossetti, & Wallace, 2005; Von Helmholtz, 1925). The agent's objective in such tasks is typically to reach from a start location to a target location as quickly, smoothly, and accurately as possible. After a baseline or familiarization phase, the visual feedback provided by the moving hand is perturbed to introduce a mismatch between the actual and perceived hand position. Early experiments of this nature used prism glasses to induce lateral shifts, but recently the most common approach has been to use crude virtual reality environments to impose visuomotor rotations such that movements beginning at the centre of a work space and travelling in a given direction generate on-screen cursor trajectories that match the radial distance from the reach origin but are rotated by some amount. People readily learn to compensate for a range of perturbations, quickly becoming proficient at moving to a target with relatively normal kinematics (Welch, 1986).

Since the early 2000s, linear dynamical systems endowed with supervised learning algorithms have provided a popular general model of sensorimotor learning, including behavior observed in visuomotor adaptation tasks (Baddeley et al., 2003; Cheng & Sabes, 2006; Donchin et al., 2003; Scheidt et al., 2001; Thoroughman & Shadmehr, 2000). This is usually done by defining the state of the dynamical system as a sensorimotor transformation – that is, as an intermediate mapping from sensory input to motor output. Sensorimotor learning is then modeled as adaptive changes that reduce the errors in each state, and for this reason, models that employ this method are often referred to as state-space models.

As an example, consider a simple reaching task in which participants make center-out reaches to a single target. After some baseline phase in which participants are afforded the opportunity to familiarize themselves with the apparatus, the visual feedback is perturbed by a rotation. Further suppose that feedback is only given at the end of each reach, so that any adaptive change in the sensorimotor mapping occurs exclusively between trials. A simple and common state-space model of this task is described on trial n by the following equations:

$$\begin{aligned} \delta_n &= y_n^* - y_n \\ x_{n+1} &= \beta x_n + \alpha \delta_n \\ y_n &= x_n + \theta_n \end{aligned} \tag{4.30}$$

where δ_n is the error (i.e., the angular distance between the reach endpoint and the target location), y_n^* is the desired output (e.g., the angular position of the reach target), y_n is the output and corresponds to the angle of the movement that will be generated when trying to reach to the target (i.e., it is a readout of the sensorimotor state), x_n is the state of the system (i.e., the sensorimotor transformation), β is a retention rate that describes how much is retained from the value of the state at the previous trial, α is a learning rate that describes how quickly states are updated in response to errors, and θ_n is the imposed rotation.

If we assume that in the absence of visuomotor rotations, the system is calibrated such that $\delta_n = 0$, and that this state corresponds to $x_n = 0$, then in the presence of a rotation $\theta_{n+1} \neq 0$, the system will experience the error $\delta_{n+1} = -\theta_{n+1}$, and will adjust x_{n+2} in a direction that would reduce the experienced error if the same rotation was applied on the next trial. For example, if θ_{n+1} is in a clockwise direction, then Eq. 4.30 leads to $x_{n+2} = \beta x_{n+1} - \alpha\theta_{n+1}$. This means that adaptation to a clockwise rotation occurs by adjusting the sensorimotor state to generate more counterclockwise movements. If $\beta < 1$, then on each trial the state will respond to errors in the way just described, but will also return to baseline by some increment. Thus, in the absence of reach errors, the system has a tendency to reset itself. Because the goal of learning is to reduce the state error – that is, δ_n – this model is based on a form of supervised learning known as the delta rule or the Widrow-Hoff rule (Widrow & Hoff, 1960).

Another key feature of linear dynamical systems as models of sensorimotor learning is that they are easily modified to accommodate considerably more complexity than the simple version described above. For example, Cheng and Sabes (2006) outlined a more general form for these models governed by the following equations:

$$\begin{aligned}\mathbf{x}_n &= \mathrm{A}\mathbf{x}_{n-1} + \mathrm{B}\delta_{n-1} + \eta_{n-1}\\ \mathbf{y}_n &= \mathrm{C}\mathbf{x}_n + \mathrm{D}\omega_n + \gamma_n,\end{aligned} \tag{4.31}$$

where $\mathbf{x}_n$ is a state vector of sensory transformations, δ_n is the vector of errors – that is, the differences between the desired and actual states, η_n is a random vector that models noise in the learning process and is typically assumed to have a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix Σ, $\mathbf{y}_n$ is a vector of motor outputs (e.g., angle and distance of movement), ω_n is a vector of inputs to the system (e.g., θ_n in the simple example above), γ_n is a random vector that models noise in the output process (again typically assumed to have a multivariate normal distribution), and A, B, C, and D are all matrices of constant values. Note that this model modifies each state according to its own unique error, which is a hallmark of supervised learning (and of the delta rule).

In this form, it is easy to see that linear dynamical systems can be flexibly applied to a variety of sensorimotor learning scenarios in which the factors relevant to sensorimotor learning (stored in δ_n) can be stated independently of the factors relevant to sensorimotor output (stored in ω_n). The result is a convenient yet powerful framework that can be used to generate predictions about sensorimotor learning on a trial-by-trial basis, or even on a moment-to-moment basis if adaptive changes are thought to occur on that timescale.

This approach is therefore well suited to modeling behavioral learning phenomena that appreciably change on these fast timescales.

4.5 Implementational Models of Human Learning

Implementational-level models explicitly state how neural circuits drive behavior, and how changes in connection weights within these circuits drive learning. Thus, at the core of these models are clear statements about the brain regions and networks that drive a behavior, and the forms of synaptic plasticity that govern changes in connection weights between neurons in constituent regions. We now know that synaptic plasticity comes in many different forms. For instance, it operates by different computational principles in different brain regions and between different cell types (Doya, 2000; Feldman, 2009), and it is governed physiologically by different molecular mechanisms and intracellular signaling cascades. A complete review of both the physiological and computational underpinnings of every form of synaptic plasticity is well beyond the scope of this chapter. Instead, we focus on three forms of synaptic plasticity that are deeply understood from a physiological perspective, and are at the core of both classic and contemporary computational models of learning. In particular, we will discuss two-factor synaptic plasticity in cerebral cortex and the hippocampus that is similar to Hebbian learning, three-factor DA-dependent synaptic plasticity in the basal ganglia that is similar to RL (Doya, 2000; Houk et al., 1995), and a form of synaptic plasticity in the cerebellum that resembles supervised learning.

4.5.1 Physiology of DA-dependent two- and three-factor synaptic plasticity

The most common excitatory neurotransmitter in the brain is glutamate, and LTP at glutamatergic synapses is well understood. Glutamate binds to a number of different receptors, but the most important for LTP is NMDA. The biochemical details are not important for our purposes, except to note that NMDA requires partial depolarization to become activated, and so it has a higher threshold for activation than non-NMDA glutamate receptors. NMDA-receptor activation initiates a number of chemical cascades that can increase synaptic efficacy. Because of its high threshold, however, activation of NMDA receptors on the post-synaptic membrane requires strong presynaptic activation. If presynaptic activation either fails to activate or only weakly activates NMDA receptors, then a variety of evidence suggests that the long-term efficacy of the synapse is weakened (i.e., LTD occurs; Bear and Linden 2001; Kemp and Bashir 2001).

DA plays a critical modulatory role in these processes because it can potentiate synaptic efficacy if it is above baseline when NMDA receptors are activated, but synaptic weakening occurs if DA is below baseline during NMDA receptor activation (Calabresi, Pisani, Mercuri, & Bernardi, 1996; Reynolds & Wickens, 2002; Yagishita et al., 2014). A large literature shows that DA neurons in the ventral tegmental area and substantia nigra pars compacta increase their firing above baseline following unexpected rewards, and decrease their firing below baseline following the failure to receive an expected reward (e.g., Hollerman and Schultz 1998; Mirenowicz and Schultz 1994; Schultz 1998). Thus, this form of DA-enhanced LTP should be in effect following an unexpected reward in any brain region that is a target of DA neurons. This includes the basal ganglia, the hippocampus, the amygdala, and all of frontal cortex. In contrast, there is virtually no DA projection to visual or auditory cortex. In these regions however, there is evidence that acetylcholine may play a modulatory role similar to DA in LTP and LTD (e.g., Gu 2003; McCoy, Huang, and Philpot 2009).

Although the biochemistry that mediates the modulatory role that DA plays in synaptic plasticity is similar in all DA target regions, the functional role of this plasticity is qualitatively different in the striatum and frontal cortex. Within the striatum, DA is quickly cleared from synapses by DA active transporter and, as a result, the temporal resolution of DA in the striatum is high enough for DA to serve as an effective trial-by-trial reinforcement-learning signal. For example, if the first response in a training session receives positive feedback and the second response receives negative feedback, then the elevated DA levels in the striatum that result from the positive feedback on trial 1 should have decayed back to baseline levels by the time of the response on trial 2. Unlike the striatum however, the concentration of DA active transporter in frontal cortex is low (e.g., Seamans and Robbins 2010). As a result, cortical DA levels change slowly. For example, the delivery of a single food pellet to a hungry rat increases DA levels in prefrontal cortex above baseline for approximately 30 min (Feenstra & Botterblom, 1996). Thus, the first rewarded behavior in a training session is likely to cause frontal cortical DA levels to rise, and the absence of DA active transporter will cause DA levels in frontal cortex to remain high throughout the training session. As a result, all synapses that are activated during the session are likely to be strengthened, regardless of whether the associated behavior is appropriate or not. Thus, although DA may facilitate LTP in frontal cortex, it appears to operate too slowly to serve as a frontal-cortical trial-by-trial reinforcement training signal (Lapish, Kroener, Durstewitz, Lavin, & Seamans, 2007).

From a computational perspective, the high temporal resolution of the striatal DA signal means that whether a synapse is strengthened or weakened depends on three factors: the amount of presynaptic activation, the amount of postsynaptic activation, and whether DA is above or below baseline. As a result, synaptic plasticity in the striatum is said to follow the *three-factor* learning rule (Wickens, 1993). In contrast, in cortex, DA levels will change only slowly over time, so only two factors are needed to predict whether a synapse will be strengthened or weakened – the amount of pre- and postsynaptic activation. As a result, plasticity in cortex follows the *two-factor* learning rule.

4.5.2 Models based on two-factor plasticity

Models of two-factor plasticity

The structural changes at the synapse that accompany LTP and LTD are complex and highly diverse. For example, changes in synaptic plasticity might be mediated by changes in the number of receptors, their distribution, the type of receptors, or their sensitivity. But plasticity changes could also occur because of changes in the size and/or shape of dendritic spines. If our goal is to model learning-related changes in human behavior, then the molecular and cellular mechanisms that mediate changes in synaptic plasticity are irrelevant. We only need an accurate model of how *much* the efficacy of the synapse changes from one behavioral measurement to the next.

The structural changes at the synapse unfold continuously in time, but unless the behavioral measurements are continuous, there is no need to build a continuous model. In particular, if the data have a discrete trial-by-trial structure, as is common in many cognitive-behavioral experiments, then a discrete-time model of changes in synaptic efficacy is often sufficient. Typically, such a model would be constructed from difference equations, where the index is trial number, so the implicit time interval is the duration of one trial. A continuous-time learning model (e.g., that uses differential equations) is typically required only when modeling a continuous-time behavioral task.

The simplest and original form of Hebbian learning predicts that between trials n and $n+1$, the strength of the synapse between units i and j, denoted by $w_{ij}(n+1)$, equals

$$w_{ij}(n+1) = w_{ij}(n) + \alpha A_i(n) A_j(n), \tag{4.32}$$

where $A_i(n)$ and $A_j(n)$ are the total activations in units i and j on trial n and α is the learning rate. This model has two significant weaknesses. First, all terms in Eq. 4.32 are positive, so this model includes no mechanism to weaken a synapse, and as a result, it cannot account for LTD. Second, note that it predicts that all synaptic strengths will eventually increase to infinity. For these reasons, a variety of alternative models of Hebbian learning have been proposed.

One model of two-factor plasticity, which can be seen as a generalization of classical Hebbian learning, assumes that (Ashby, 2018)

$$\begin{aligned} w_{ij}(n+1) &= w_{ij}(n) \\ &\quad + \alpha\ \Delta\ H\left[A_j(n) - \theta_{\text{NMDA}}\right] A_i(n) \left\{1 - \mathrm{e}^{-\lambda[A_j(n)-\theta_{\text{NMDA}}]}\right\} [1 - w_{ij}(n)] \\ &\quad - \beta\ H\left[\theta_{\text{NMDA}} - A_j(n)\right] A_i(n)\ \mathrm{e}^{-\lambda[\theta_{\text{NMDA}}-A_j(n)]} w_{ij}(n). \end{aligned} \tag{4.33}$$

The positive term describes conditions that strengthen the synapse and the negative term describes conditions that cause the synapse to be weakened. Ignore the constant Δ for now (i.e., assume $\Delta = 1$). The function $H[g(x)]$ is the Heaviside function that equals 1 when $g(x) > 0$ and 0 when $g(x) \leq 0$. The constant θ_{NMDA} represents the threshold for NMDA-receptor activation. Note that the synaptic strengthening term is positive only on trials when the postsynaptic activation exceeds the threshold for NMDA-receptor activation, and that the amount of strengthening depends on the product of the presynaptic activation and an exponentially increasing function of the postsynaptic activation. The $[1 - w_{ij}(n)]$ term is a rate-limiting term that prevents $w_{ij}(n+1)$ from exceeding 1.0, and the constant λ scales the postsynaptic activation.

Note that the synapse is weakened only when the postsynaptic activation is below the NMDA threshold. Also note that the exponential term reaches its maximum when postsynaptic activation is near the NMDA threshold and decreases as the postsynaptic activation gets smaller and smaller. This is consistent with the neurobiology. For example, in the absence of any postsynaptic activation, we do not expect any synaptic plasticity. The $w_{ij}(n)$ at the end prevents $w_{ij}(n+1)$ from dropping below 0. Figure 4.1 shows predicted changes in synaptic strength [i.e., $w_{ij}(n+1) - w_{ij}(n)$] for this model as a function of the magnitude of postsynaptic activation during both early [when $w_{ij}(n) = 0.2$] and late [when $w_{ij}(n) = 0.8$] learning.

The Eq. 4.33 model of two-factor learning assumes that any activation in postsynaptic unit j was caused by activation in presynaptic unit i. This assumption is really only plausible in simple feedforward models. If unit j receives input from many other units, then Eq. 4.33 could strengthen inap-

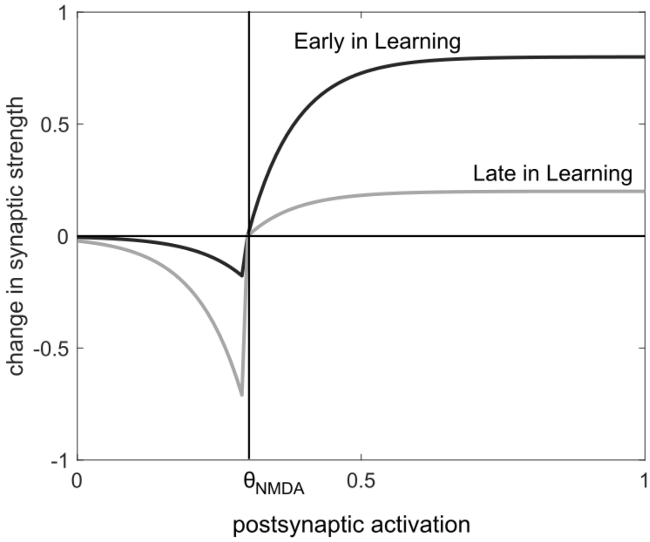

Figure 4.1 Change in synaptic strength predicted by the two-factor learning model described in Eq. 4.33 as a function of amount of postsynaptic activation (here scaled from 0 to 1). Predictions are shown for early in learning [i.e., when $w_{ij}(n) = 0.2$] and late in learning [i.e., when $w_{ij}(n) = 0.8$].

propriate synapses. In the mammalian brain, the magnitude and even the direction of plasticity at a synapse depends not only on the magnitude of the pre- and postsynaptic activations, but also on their timing – a phenomenon known as *spike-timing-dependent plasticity* . Considerable data show that if the postsynaptic neuron fires just after the presynaptic neuron then synaptic strengthening (i.e., LTP) occurs, whereas if the postsynaptic neuron fires first then the synapse is weakened (e.g., Bi and Poo 2001; Sjöström, Rancz, Roth, and Häusser 2008). Furthermore, the magnitude of both effects seems to fall off exponentially as the delay between the spikes in the pre- and postsynaptic neurons increases. Let T_{pre} and T_{post} denote the time at which the pre- and postsynaptic units fire, respectively. Then a popular model of spike-timing-dependent plasticity (e.g., Zhang, Tao, Holt, Harris, and Poo 1998) assumes that the amount of change in the synaptic strength equals

$$\Delta = \begin{cases} \mathrm{e}^{-\theta_+(T_{\text{post}}-T_{\text{pre}})}, & \text{if } T_{\text{post}} > T_{\text{pre}} \\ \mathrm{e}^{\theta_-(T_{\text{post}}-T_{\text{pre}})}, & \text{if } T_{\text{post}} < T_{\text{pre}}, \end{cases} \tag{4.34}$$

where θ_+ and θ_- are parameters that determine the decay rates of synaptic strengthening and weakening, respectively. Figure 4.2 shows an example of this function.

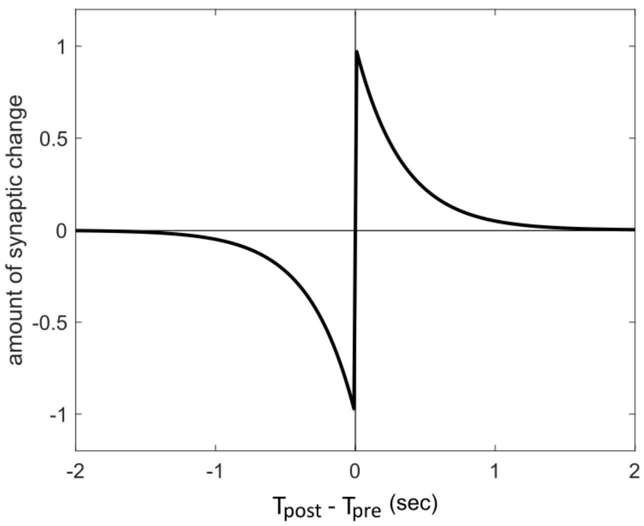

Figure 4.2 Amount of change in synaptic strength predicted by spike-timing-dependent plasticity as a function of the difference in time between firing in the postsynaptic neuron (i.e., $\mathrm{T}_{\mathrm{post}}$) and the presynaptic neuron (i.e., $\mathrm{T}_{\mathrm{pre}}$).

To incorporate spike-timing-dependent plasticity into two-factor learning, the first step is to compute Δ from Eq. 4.34 anytime the pre- and postsynaptic units both fire. Next this value is inserted into Eq. 4.33 to compute $w(n+1)$.

Models of human learning that incorporate two-factor plasticity

Hasselmo and Wyble (1997) proposed a model that includes two-factor plasticity in the hippocampus to account for the effects of scopolamine, an acetylcholine anatagonist, on free recall and recognition. They tested this model against data from an experiment reported by Ghoneim and Mewaldt (1975), in which participants studied lists of 16 words each and were then tested on their ability to recall and recognize the studied words. Recall and recognition were both intact when scopolamine was administered between study and test. In contrast, the administration of scopolamine before study impaired recall, but not recognition.

Figure 4.3 shows the neural architecture of the Hasselmo and Wyble (1997) model. Neural activation in each region was modeled by firing-rate models (e.g., see Ashby 2018). The hippocampus contains two subfields, the cornus ammonis and the dentate gyrus, each of which receives input from

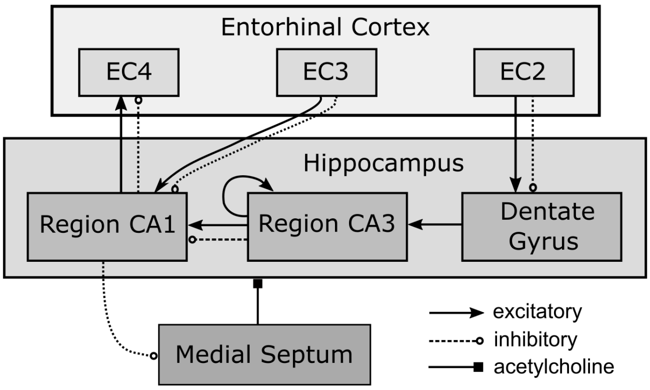

Figure 4.3 The neural architecture of the Hasselmo and Wyble (1997) hippocampal model. EC2, EC3, and EC4 denote different subregions in entorhinal cortex, whereas CA1 and CA3 denote different subregions in the cornus ammonis. Two-factor learning occurs at virtually all synapses, except at the synapses between dentate gyrus and CA3 and between CA1 and medial septum.

entorhinal cortex, which in turn is driven by widespread input from neocortex. The network is characterized by sparse encoding and many feedback loops, and the behavior of the model is governed largely by how the resulting network dynamics approach attractor states.

The network has two global states (encoding and retrieval) that are controlled by the concentration of acetylcholine. The encoding mode is triggered by elevated acetylcholine and is characterized by potentiated two-factor learning at all plastic synapses (hence encoding), and also by inhibited output from EC4 back to neocortex (hence no retrieval). Acetylcholine can also reduce excitatory transmission, limiting the effects of recurrent collaterals and making the network primarily sensitive to external inputs. This is good for learning because it helps reduce interference between new items and previously stored items. The retrieval mode is triggered by depressed acetylcholine and is characterized by reduced two-factor learning at plastic synapses (hence no encoding) and also by potentiated output from EC4 to neocortex (hence retrieval). Low acetylcholine also allows excitatory transmission via the networks recurrent collaterals, making the network sensitive to stored representations.

The form of two-factor learning used in the model is essentially the same as in Eq. 4.33, but with the addition of providing a model of how the α and β parameters in Eq. 4.33 change with concentrations of acetylcholine. The model successfully simulates recall when context (i.e., cues associated with the word list) is presented to the network and it outputs words associated with that context. Additionally, the model successfully simulates recognition when it is presented with words and it outputs the context associated with the words. Hasselmo and Wyble (1997) showed that in the presence of scopolamine, the network has no difficulty retrieving inputs learned prior to scopolamine administration, whereas recall of inputs encoded in the presence of scopolamine is disrupted and recognition of these inputs is spared. For a full explanation of the network dynamics that enable the model to account for these phenomena, see Hasselmo and Wyble (1997).

Here we only focused on the synaptic effects of acetylcholine on the hippocampus. However, Hasselmo and Wyble (1997) also explored the effects on depolarization and adaptation of neurons. Furthermore, the model was also shown to account for the list length and list strength effects (Murdock & Kahana, 1993; Murdock Jr, 1962; Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Roberts, 1972) in addition to making predictions about the effects of scopolamine on paired-associate tasks (Caine, Weingartner, Ludlow, Cudahy, & Wehry, 1981; Crow & Grove-White, 1973; Ostfeld & Aruguete, 1962). The Hasselmo and Wyble (1997) model provides a good illustration of how relatively simple two-factor plasticity rules can be incorporated into sophisticated implementational-level models that account for neuropharmacological and behavioral phenomena.

4.5.3 Models based on DA-dependent three-factor plasticity

Models of DA-dependent three-factor plasticity

In the striatum, DA reuptake is fast, so plasticity follows the three-factor rule. In other words, three factors are needed to strengthen a synapse: strong presynaptic activation, strong postsynaptic activation, and DA above baseline. If any of these factors are missing, then the synapse is weakened. A discrete-time model of three-factor learning is as follows:

$$\begin{aligned}
w_{ij}(n+1) &= w_{ij}(n) \\
&\quad + \alpha\ H\left[A_j(n) - \theta_{\text{NMDA}}\right] H[D(n) - D_{\text{base}}] \\
&\qquad \times A_i(n) \left\{1 - \mathrm{e}^{-\lambda[A_j(n) - \theta_{\text{NMDA}}]}\right\} [D(n) - D_{\text{base}}][1 - w_{ij}(n)] \\
&\quad - \beta\ H\left[A_j(n) - \theta_{\text{NMDA}}\right] H[D_{\text{base}} - D(n)] \\
&\qquad \times A_i(t)\ \left\{1 - \mathrm{e}^{-\lambda[A_j(n) - \theta_{\text{NMDA}}]}\right\} [D_{\text{base}} - D(n)] w_{ij}(n) \\
&\quad - \gamma\ H\left[\theta_{\text{NMDA}} - A_j(n)\right] A_i(n)\ \mathrm{e}^{-[\theta_{\text{NMDA}} - A_j(n)]} w_{ij}(n), \qquad (4.35)
\end{aligned}$$

where $D(n)$ is the amount of DA released on trial n and D_{base} is the baseline DA level (Ashby, 2018).

Recall that $H(x)$ is the Heaviside function, which equals 0 if $x \leq 0$ and 1 if $x > 0$. Therefore, the positive LTP term equals 0 except when presynaptic activation exceeds the postsynaptic NMDA threshold (i.e., $A_j(n) > \theta_{\text{NMDA}}$) and DA exceeds baseline (i.e., $D(n) > D_{\text{base}}$). Thus, synaptic strengthening requires three conditions – strong presynaptic activation, postsynaptic activation above the threshold for NMDA-receptor activation, and DA above baseline. Once these conditions are met, synaptic strengthening is the same as in the Eq. 4.33 two-factor learning model. Two different conditions cause the synapse to be weakened. The second (the last γ term in Eq. 4.35) is the same as in the two-factor model. The first (i.e., the β term) however, is unique to striatal-mediated three-factor plasticity. Cortical-striatal synapses are weakened if postsynaptic activation is strong and DA is below baseline – a condition that would occur for example, on trials when feedback indicates the trial n response was incorrect.

The Eq. 4.35 model of three-factor plasticity requires that we specify the amount of DA released on every trial in response to the feedback signal [the $D(n)$ term]. The more that DA increases above baseline (D_{base}), the greater the increase in synaptic strength, and the more it falls below baseline, the greater the decrease.

Although there are a number of powerful models of DA release, Eq. 4.35 requires only that we specify the amount of DA released to the feedback signal on each trial. The key empirical results are (e.g., Schultz, Dayan, and Montague 1997; Tobler, Dickinson, and Schultz 2003): (1) midbrain DA neurons fire tonically, and therefore have a nonzero baseline (i.e., spontaneous firing rate); (2) DA release increases above baseline following unexpected reward, and the more unexpected the reward the greater the release, and (3) DA release decreases below baseline following unexpected absence of reward, and the more unexpected the absence, the greater the decrease. One common interpretation of these results is that over a wide range, DA firing is proportional to the reward prediction error (RPE) – that is, to the difference between obtained reward and predicted reward. If we denote the obtained reward on trial n by r_n and the predicted reward by P_n, then the RPE on trial n is defined as:

$$RPE_n = r_n - P_n. \tag{4.36}$$

So positive prediction errors occur when the reward is better than expected, and negative prediction errors when the reward is worse than expected. Note that either a positive or negative prediction error is a signal that learning is incomplete.

A simple model of DA release can be built by specifying how to compute 1) obtained reward, 2) predicted reward, and 3) exactly how the amount of DA release is related to the RPE. A straightforward solution to these three problems is as follows (Ashby & Crossley, 2011). First, in tasks that provide positive feedback, negative feedback, or no feedback on every trial and where reward magnitude never varies, then a simple model can be used to compute obtained reward. Specifically, define the obtained reward r_n on trial n as +1 if correct or reward feedback is received, 0 in the absence of feedback, and -1 if error feedback is received.

Second, following an old tradition (Bush & Mosteller, 1951), predicted reward can be computed using the iterative sample mean (i.e., Eq. 4.5):

$$P_{n+1} = P_n + \alpha_p(r_n - P_n), \tag{4.37}$$

where α_p is the learning rate. [4]

The final step is to compute the amount of DA released for any specific value of RPE. A simple model, which is consistent with the single-unit recording data reported by Bayer and Glimcher (2005) assumes that

$$D(n) = \begin{cases} 1 & \text{if } RPE > 1; \\ .8\ RPE + .2 & \text{if } -.25 < RPE \leq 1; \\ 0 & \text{if } RPE < .25. \end{cases} \tag{4.38}$$

Note that this model assumes a baseline DA level of 0.2 [i.e., $D(n)$ on trials when $RPE = 0$]. Positive RPEs increase DA release above this baseline, and negative RPEs depress it below baseline.

Figure 4.4 shows predicted changes in synaptic strength [i.e., $w_{ij}(n+1) - w_{ij}(n)$] for this model as a function of the magnitude of postsynaptic activation, separately for early [when $w_{ij}(n) = 0.2$] and late learning [when $w_{ij}(n) = 0.8$], and following correct and incorrect responses. Note that synaptic plasticity following correct (rewarded) responses is similar to plasticity in the two-factor model (compare the top panel of Figure 4.4 with Figure 4.1). The only real difference is that plasticity is attenuated more during late learning in the three-factor model. This is because DA fluctuations decrease as rewards become more predictable. Note also that errors have a greater effect on synaptic plasticity late in learning. This is because errors are expected early in learning, so DA fluctuations are small. Late in learning

[4] The subscript p is to distinguish this learning rate parameter from the learning rate α in Eq. 4.35.

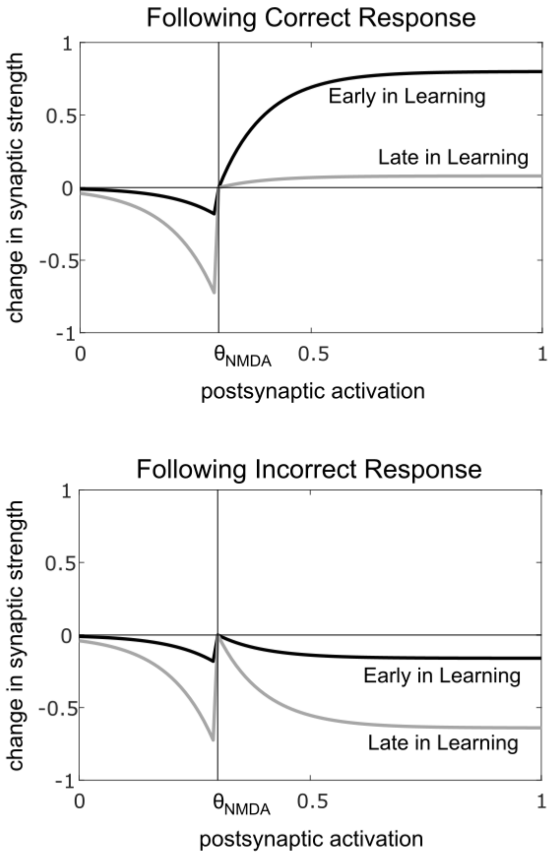

Figure 4.4 Change in synaptic strength predicted by the model of three-factor plasticity described in Eq. 4.35 as a function of amount of post-synaptic activation (here scaled from 0 to 1). Predictions are shown for early in learning [i.e., when $w_{A,B}(n) = 0.2$] and late in learning [i.e., when $w_{A,B}(n) = 0.8$], and following feedback that the response was correct response or incorrect. ($\alpha = 2$, $\beta = 4$, $\gamma = 1$).

however, when accuracy is high, errors are unexpected, which causes a large DA depression and therefore a large decrease in synaptic efficacy.

Relationship of 3-factor plasticity to psychological constructs of RL

Three-factor plasticity may – in some respects – be seen as a possible neural implementation of the many SR association learning models that were inspired by Thorndike's (1927) law of effect. The obvious analogy maps presynaptic activity onto the stimulus component, postsynaptic activity onto the response component, and DA onto the reinforcement signal. A step further, and we might expect the stimulus component to be encoded by a primary sensory neuron, the response unit to be encoded by a primary motor neuron, and the reinforcement signal to strengthen or weaken the synapse between these two neurons. Although human neuroanatomy supports the existence of direct projections from sensory to motor areas, the evidence suggests that these synapses are not strengthened via a DA-mediated reinforcement signal, because DA reuptake in cortex is too slow. Rather, the available evidence suggests that sensory and motor neurons are indirectly wired together via a DA-mediated reinforcement signal in the basal ganglia. Here, stimulus-response associations can be learned at cortical-striatal synapses, with the striatum projecting via a multi-synaptic pathway to the motor neurons representing the response component of the association. From this perspective, the anatomy and physiology of cortical–basal ganglia–DA interactions may provide a plausible neural substrate for the classic psychological constructs of stimulus-response learning originally posed by Thorndike. However, the anatomy also suggests that the association mechanism is more indirect and complex than in the original proposals of direct reinforcement of stimulus/response components.

Relationship of 3-factor plasticity to machine-learning constructs of RL

Three-factor plasticity in the basal ganglia may also offer a plausible biological substrate for various machine-learning constructs of RL. In this view, cortical-striatal synaptic weights implement a value function, and DA neurons provide the reinforcement signal – a role motivated by the finding that DA neuron firing reflects an RPE (Glimcher, 2011; Schultz et al., 1997). This arrangement could be seen as compatible with a range of specific RL algorithms, including temporal-difference learning, Q learning, and actor-critic architectures, although the mapping does not seem perfect for any of these.

To be compatible with temporal-difference learning, cortical-striatal synaptic weights would need to encode a value function that depends exclusively on sensory states (i.e., is independent of action). This sort of value function encoding may be characteristic of the ventral striatum (e.g., nucleus accumbens). The value function would also need to be used to generate prediction errors, which is consistent with one of the roles sometimes ascribed to the ventral striatum. However, the value function would also need to operate under the assumption of a fixed action policy, and at present, it is unclear whether the ventral striatum learns different value functions for different policies. Another feature of temporal-difference learning, which makes it a problematic model of DA neuron firing, is that, as we saw earlier, the temporal-difference signal propagates back one time step every trial, until it reaches the cue, at which point the propagation ends. DA neurons initially fire to the reward, and eventually, after learning occurs, they begin to fire to the cue. But there is no evidence that the propagation backwards is incremental – that is, there is never a DA response to an intermediate time point between cue and reward [5].

To be compatible with Q-learning, cortical-striatal synaptic weights would need to encode a value function that combines both sensory states and actions. This sort of value function encoding may be characteristic of the dorsal striatum. Parts of the dorsal striatum have quite direct access to motor areas of cortex, so it is plausible that they could also directly implement the action selection components of Q learning. However, DA-encoded RPEs would also need to be derived from the value estimates provided by the dorsal striatum. At present, it is unclear to what degree such prediction errors factor in information about action.

In actor-critic RL models, an actor system implements an action selection policy, and a critic system estimates the value of different states and uses these estimates to generate prediction errors, which are then used to update the critic's value estimates and the actor's selection policy. Of all the machine-learning RL algorithms, these models may most easily map onto three-factor plasticity in the basal ganglia (Houk et al., 1995; Joel, Niv, & Ruppin, 2002; Sutton & Barto, 1998). In this view, the critic is implemented by the DA system and the actor is implemented by cortical-striatal projections through the dorsal striatum. Since the critic is a separate module from the actor, there is no need for cortical-striatal synaptic weights (part of the actor) to be used to compute prediction errors. However, this view does not say where and how the value function is implemented. One possibility is the ventral striatum (Takahashi, Schoenbaum, & Niv, 2008).

Models of human learning that incorporate three-factor plasticity

The COVIS procedural-learning model incrementally learns arbitrary stimulus-response associations via a model of three-factor plasticity that is essentially identical to Eq. 4.35. Figure 4.5 shows the architecture of the model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Crossley, 2011; Ashby & Waldron, 1999; Cantwell, Crossley, & Ashby, 2015). The key structure is the striatum, a major input region within the basal ganglia that includes the caudate nucleus and the putamen. In primates, all of extrastriate visual cortex projects directly to the striatum, with a cortical-striatal convergence

[5] This problem can be solved by replacing the temporal-difference learning algorithm with a version that includes an eligibility trace, which allows the error to propagate backwards by more than a single state per step (Sutton & Barto, 1998).

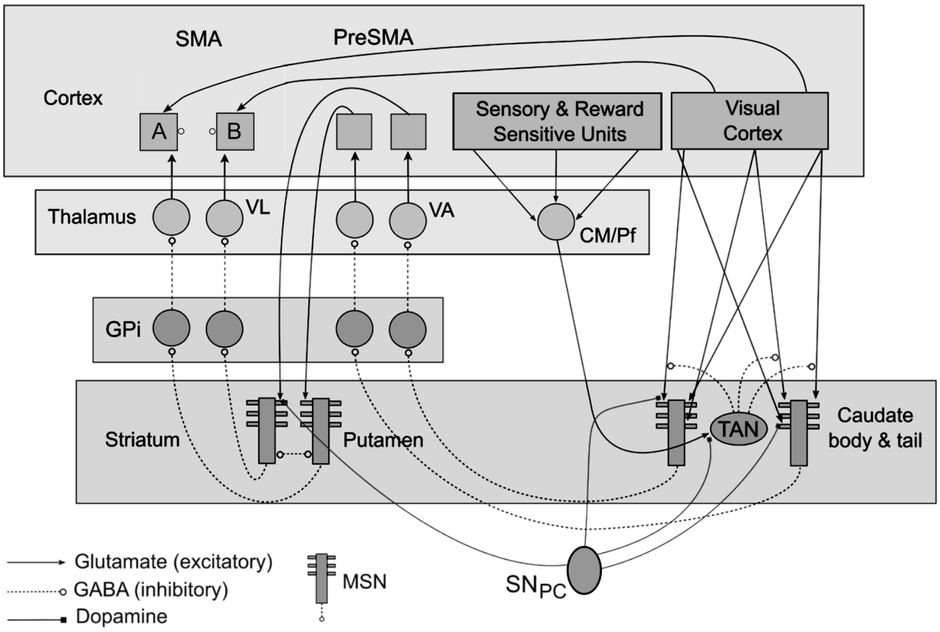

Figure 4.5 The neural architecture of the COVIS model of procedural learning for a two-alternative forced-choice task with responses A and B (SMA = supplementary motor area, PreSMA = presupplementary motor area, VL = ventral lateral nucleus of the thalamus, VA = ventral anterior nucleus of the thalamus, CM/Pf = centromedian and parafascicular nuclei of the thalamus, GPi = internal segment of the globus pallidus, TAN = tonically active neuron, SN$_{\text{PC}}$ = substantia nigra pars compacta, MSN = medium spiny neuron of the striatum).

ratio of approximately 10,000 to 1 (e.g., C. J. Wilson 1995). The model assumes that, through a procedural-learning process, each striatal medium spiny neuron associates a motor goal (e.g., press the button on the left) with a large group of visual cortical neurons (i.e., all that project to it). Much evidence supports the hypothesis that procedural learning is mediated within the basal ganglia, and especially at cortical-striatal synapses, which exhibit three-factor plasticity (Ashby & Ennis, 2006; Houk et al., 1995; Mishkin, Malamut, & Bachevalier, 1984; Willingham, 1998). The COVIS procedural-learning model is a formal instantiation of these ideas.

Note that the model includes two loops through the basal ganglia (Cantwell et al., 2015). One loop projects from visual cortex through the body and tail of the caudate nucleus and terminates in pre-supplementary motor area, and the second loop projects from pre-supplementary motor area through the putamen and terminates in supplementary motor area. Because this second loop terminates in premotor cortex, COVIS predicts that the associations that are learned are between stimuli and motor goals. Both loops rely on three-factor learning at cortical-striatal synapses. The first loop learns which stimuli are associated with the same response and the second loop learns what motor response is associated with each of these stimulus clusters. In a novel task, both types of learning are required. However, note that if we train agents to make accurate classification responses and then switch the responses associated with the two stimulus classes, then the classes remain unchanged – only the response mappings must be relearned. So COVIS predicts that reversing the locations of the response keys will interfere with procedural classification performance, but that recovery from such a reversal should be easier than novel classification learning – a prediction that has been supported in several studies (Cantwell et al., 2015; Kruschke, 1996; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010; Sanders, 1971; Wills, Noury, Moberly, & Newport, 2006).

COVIS uses a biologically accurate model of spiking in individual neurons proposed by Izhikevich (2003). Let $V_i(t)$ and $V_j(t)$ denote the intracellular voltages of the pre- and postsynaptic neurons, respectively, at time t. Then the Izhikevich (2003) model assumes that the intracellular voltage of the postsynaptic neuron on trial n is described by the following differential equations:

$$\begin{aligned} \frac{dV_j(t)}{dt} &= w_{ij}(n) f\left[V_i(t)\right] + \beta + \gamma\left[V_j(t) - V_r\right]\left[V_j(t) - V_t\right] - \theta U_j(t), \\ \frac{dU_j(t)}{dt} &= \lambda\left[V_j(t) - V_r\right] - \omega U_j(t), \end{aligned} \tag{4.39}$$

where β, γ, V_r, V_t, θ, λ, and ω are constants that are adjusted to produce dynamical behavior that matches the neural population being modeled. $U_j(t)$ is an abstract regulatory term that is meant to describe slow recovery in the postsynaptic neuron after an action potential is generated. Equation 4.39 produces the upstroke of an action potential via its own dynamics. To produce the downstroke, $V_j(t)$ is reset to V_{reset} when it reaches V_{peak}, and at the same time, $U_j(t)$ is reset to $U_j(t) + U_{\text{reset}}$, where V_{reset}, V_{peak}, and U_{reset} are free parameters.

The model has many free parameters and therefore can fit a wide variety of dynamical behavior. Izhikevich (2003) identified different sets of parameter values that allow the model to mimic the spiking behavior of approximately 20 different types of neurons. For example, one set of parameter values allows the model to mimic the firing properties of the striatal medium spiny neurons shown in Figure 4.5 (including, e.g., their up and down states), and another set of values allows the model to mimic the regular spiking neurons that are common in cortex. Furthermore, Ashby and Crossley (2011) modified the Izhikevich model to account for the unusual dynamics of the striatal cholinergic interneurons (which produce a pronounced pause in their high tonic firing rate following excitatory input). In all these cases, the parameters are fixed by fitting the model to single-unit recording data from the neural population being modeled. Once set, the parameter values that define the models of each individual neuron type then remain fixed throughout all applications. Therefore, when testing the model against behavioral or neuroimaging data, the models of each neuron type have zero free parameters.

The function $f[V_i(t)]$ in Eq. 4.39 models the input from the presynaptic neuron i. In particular, it uses a simple model called the alpha function to mimic the temporal delays of spike propagation and the temporal smearing that occurs at the synapse (Rall, 1967). Specifically, the alpha function assumes that every time the presynaptic neuron spikes, the following input is delivered to the postsynaptic neuron (with spiking time $t = 0$):

$$\alpha(t) = \frac{t}{\delta} \exp\left(\frac{\delta - t}{\delta}\right), \tag{4.40}$$

where δ is a constant. This function has a maximum value of 1.0 and it decays to .01 at $t = 7.64\delta$. Thus, δ can be chosen to model any desired temporal delay. Suppose the presynaptic neuron i produces N spikes that occur at times $t_1, t_2, ..., t_N$. Then the function f in Eq. 4.39 equals

$$f\left[V_i(t)\right] = \sum_{k=1}^{N} \left[\alpha(t - t_k)\right]^+ , \tag{4.41}$$

where

$$[\alpha(t - t_k)]^+ = \begin{cases} \alpha(t - t_k) & \text{if } t > t_k; \\ 0 & \text{if } t \leq t_k. \end{cases} \tag{4.42}$$

Finally, synaptic plasticity, and therefore learning, is modeled by the $w_{ij}(n)$ multiplier on $f\left[V_i(t)\right]$ in Eq. 4.39. The value of this term is adjusted trial-by-trial, either via the two-factor (Eq. 4.33) or three-factor (Eq. 4.35) models of synaptic plasticity. COVIS assumes that the procedural learning in the striatum is mediated by three-factor plasticity at cortical-striatal synapses. Therefore, the presynaptic neuron i in Eq. 4.39 would be in cortex (either visual cortex or pre-supplementary motor area), the postsynaptic neuron j would be a medium spiny neuron in the striatum, and $w_{ij}(n)$ would be adjusted trial-by-trial by Eq. 4.35. For a complete description of this type of mathematical modeling, called computational cognitive neuroscience, see Ashby (2018).

COVIS uses the Izhikevich (2003) model (i.e., Eq. 4.39) to model spiking in all neuron types shown in all brain regions illustrated in Figure 4.5, and it uses the alpha function (Eq. 4.41) to model synaptic transmission between all connected neurons. The supplementary motor area in the model includes as many simulated neurons as there are response alternatives in the task under study. Figure 4.5 shows the architecture of the model when applied to a two-alternative forced-choice task with responses A and B. To generate a motor behavior, a response threshold is set on the integrated alpha function of each supplementary motor area unit (i.e., the integral of Eq. 4.41). The first unit to exceed its threshold initiates its associated motor response. The lateral inhibition between competing supplementary motor area units causes the units to display the type of push-pull activity identified in many premotor regions of cortex (e.g., as in Shadlen and Newsome 2001). Formally, this architecture – that is, separate accumulators with lateral inhibition – mimics a drift diffusion process, but of course, is more easily extended to tasks with more than two response alternatives (Bogacz, Usher, Zhang, & McClelland, 2007; P. L. Smith & Ratcliff, 2004; Usher & McClelland, 2001).

Note that COVIS predicts that synaptic strengthening can only occur when the visual trace of the stimulus and the post-synaptic effects of DA overlap in time. More specifically, synaptic plasticity in the striatum is strongest when the intracellular signaling cascades driven by NMDA receptor activation and DA D1 receptor activation coincide (Lisman, Schulman, & Cline, 2002; **?**). The further apart in time these two cascades peak, the less effect DA will have on synaptic plasticity. For example, Yagishita et al. (2014) reported that synaptic plasticity was best (i.e., greatest increase in spine volume on striatal medium spiny neurons) when DA neurons were stimulated 600 ms after medium spiny neurons. When the DA neurons were stimulated before or 5 s after the medium spiny neurons, then no evidence of any plasticity was observed. In a task mediated by procedural learning, activation of the medium spiny neurons should occur just before the motor response, and activation of the DA neurons should occur just after the feedback. So COVIS predicts that feedback delays during procedural learning should have effects that are similar to those observed by Yagishita et al. (2014). In fact, many studies have confirmed this prediction in a form of category learning thought to depend on procedural learning (i.e., the information-integration categorization task; Dunn, Newell, and Kalish 2012; Maddox, Ashby, and Bohil 2003; Maddox and Ing 2005; Worthy, Markman, and Maddox 2013). Valentin, Maddox, and Ashby (2014) showed that the

COVIS procedural-learning model can accurately account for the effects of all these feedback delays. In contrast, the same studies showed that delays up to 10 s have no effect on rule-based category learning that is thought to be mediated primarily in prefrontal cortex.

Ashby and Crossley (2011) proposed that the striatal cholinergic interneurons serve as a context-sensitive gate between cortex and the striatum (see also Crossley, Ashby, and Maddox 2013, 2014; Crossley, Horvitz, Balsam, and Ashby 2016). The idea, which is supported by a wide variety of neuroscience evidence, is that the striatal cholinergic interneurons tonically inhibit cortical input to striatal medium spiny neurons (e.g., Apicella, Legallet, and Trouche 1997; Pakhotin and Bracci 2007). The striatal cholinergic interneurons are driven by neurons in the centremedian–parafascicular nuclei of the thalamus, which in turn are broadly tuned to features of the environment. In rewarding environments, the cholinergic interneurons learn to pause to stimuli that predict reward, which releases the cortical input to the striatum from inhibition. This allows striatal output neurons to respond to excitatory cortical input, thereby facilitating cortical-striatal plasticity. In this way, cholinergic interneuron pauses facilitate the learning and expression of striatal-dependent behaviors. When rewards are no longer available, the cholinergic interneurons cease to pause, which prevents striatal-dependent responding and protects striatal learning from decay.

Extending the COVIS procedural-learning system to include striatal cholinergic interneurons allows the model to account for many new phenomena – some of which have posed difficult challenges for previous learning theories. One of these is that the reacquisition of an instrumental behavior after it has been extinguished is considerably faster than during original acquisition (Ashby & Crossley, 2011). The model accounts for this ubiquitous phenomenon because the withholding of rewards during the extinction period causes the cholinergic interneurons to stop pausing to sensory cues in the conditioning environment (since they are no longer associated with reward). This closes the gate between cortex and the striatum, which prevents further weakening of the cortical-striatal synapses. When the rewards are reintroduced, the cholinergic interneurons relearn to pause, and the behavior immediately reappears because of the preserved synaptic strengths.

4.5.4 Models based on plasticity that mimics supervised learning

The cerebellum is commonly thought to provide a neural substrate for supervised learning (Doya, 1999) and there is a rich basis of implementational-level models in support of this view, beginning with the seminal work of

Marr (1969). For this reason, the following sections are focused on learning in the cerebellum.

Learning in the cerebellum

The cerebellum is anatomically arranged into multisynaptic loops with the cerebral cortex (Ramnani, 2006). Influence over the cerebellum is orchestrated through the pons, which receives widespread inputs from cortical and peripheral sites – including those associated with proprioception (Sawtell, 2010), haptics (Ebner & Pasalar, 2008; Shadmehr & Krakauer, 2008; Weiss & Flanders, 2011), and ongoing motor commands (Schweighofer, Spoelstra, Arbib, & Kawato, 1998) – and gives rise to the mossy fiber inputs to cerebellar granule cells. Granule cells give rise to parallel fibers, which provide one of two major inputs to the Purkinje cells of the cerebellar cortex, which are the only projection neurons in the cerebellar cortex. The second input to Purkinje cells comes from climbing fibers, which originate in the inferior olive. Purkinje cells project to the cerebellar deep nuclei, which in turn are relayed to the thalamus, and ultimately back to cortex, thereby closing the anatomical loop.

Classic theories proposed that the cerebellum uses a form of supervised learning to control and coordinate motor function (Albus, 1971; Ito, 1984; Marr, 1969). In essence, these theories viewed the cerebellum as a biological implementation of a perceptron (Rosenblatt 1958; see Figure 4.6), with distributed inputs provided by the mossy fibers, error signals communicated by the climbing fibers, and supervised learning carried out by synaptic plasticity at the synapses between parallel fibers and Purkinje cells (either LTP as originally proposed by Marr 1969 or LTD as originally proposed by Ito 1984). Ito and colleagues played pivotal roles in establishing the biological plausibility of this synaptic plasticity (e.g., Ito 1984).

The anatomy of the cerebellum is unique in a few ways that probably played a large role in the development of these models. First, granule cells constitute more than half the neurons in the mammalian cerebellum (Eccles, Ito, & Szentágothai, 1967; Palay & Chan-Palay, 2012), so mossy fiber input seems like a plausible biological substrate for the distributed input representations commonly used with perceptrons. Second, each Purkinje neuron receives input from exactly one climbing fiber, and each fiber makes extensive synaptic contact with the dendritic tree of its target Purkinje neuron (Eccles et al., 1967; Palay & Chan-Palay, 2012). The most effective training methods for artificial neural networks rely on supervised learning algorithms that implement some form of gradient descent (e.g., backpropagation), which require the system to have fine-grained access to errors that occur at ev-

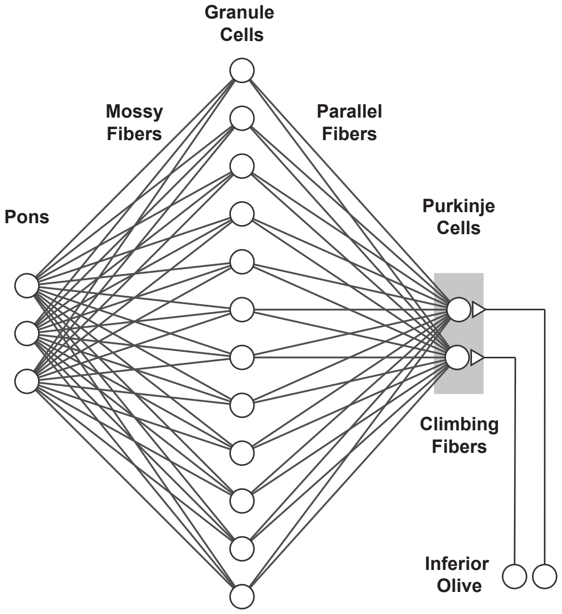

Figure 4.6 Simplified neuroanatomy of the cerebellum when viewed as a three layer perception. Purkinje cell output is inhibitory. All other illustrated projections are excitatory. See text for further details.

ery synapse. The one-to-one correspondence between Purkinje neurons and climbing fibers may be a biologically plausible way of projecting these errors into the cerebellum.

Later physiological discoveries also fall roughly in line with this classic view of the cerebellum. For instance, in Purkinje neurons, the shape of the spike evoked by activation of parallel fibers (i.e., "simple spike") is different from the shape of the spike evoked by inferior olive activation (i.e., "complex spike"). Simple spikes encode parameters of movement such as trajectory, velocity, and acceleration (Gomi et al., 1998; Shidara, Kawano, Gomi, & Kawato, 1993), whereas complex spikes encode errors in movement (Kitazawa, Kimura, & Yin, 1998; Kobayashi et al., 1998), which is compatible with their involvement in a learning process. Furthermore, the granule cell/Purkinje cell synapse is highly plastic (e.g., it exhibits LTP and LTD both presynaptically and postsynaptically), and climbing fiber signals can control the direction of plasticity (e.g., LTP versus LTD) at granule cell/Purkinje cell synapses (Coesmans, Weber, De Zeeuw, & Hansel, 2004; Lev-Ram, Mehta, Kleinfeld, & Tsien, 2003). Much is known about the intracellular signalling cascades that drive this plasticity (van Woerden et al., 2009), but the details are beyond the scope of this chapter.

The mechanisms of synaptic plasticity at parallel fiber/Purkinje cell synapses do not fall neatly into the network architectures assumed by two-factor and three-factor learning rules. The two-factor learning rule describes synaptic plasticity when only two neurons are connected (i.e., a presynaptic neuron and a postsynaptic neuron), and the three-factor learning rule describes plasticity when a presynaptic neuron and a dopaminergic input converge on a postsynaptic neuron. In contrast, plasticity at parallel fiber/Purkinje neuron synapses is determined by the convergence of parallel fibers and climbing fibers – both of which are excitatory glutamatergic projections – onto Purkinje neurons. Thus, synaptic plasticity at parallel fiber/Purkinje cell synapses follows its own unique learning rule. In particular, LTD is induced with (1) strong presynaptic activation from input 1, (2) strong presynaptic activation from input 2, and (3) strong postsynaptic activation. In contrast, LTP is induced with (1) weak presynaptic activation from input 1, (2) weak or absent activation from presynaptic input 2, and (3) weak postsynaptic activation. A further difference is that, in the two-factor learning rule, strong presynaptic activation (i.e., above the threshold for NMDA receptor activation) leads to LTP, and weak presynaptic activation leads to LTD. At parallel-fiber/Purkinje neuron synapses, these roles are reversed: weak activation of presynaptic Purkinje neurons leads to LTP, and strong activation leads to LTD.

Finally, we now know that there is synaptic plasticity at a multitude of synapses within the cerebellar circuit beyond those postulated by the classic model (e.g., between mossy fibers, between Purkinje cells and deep cerebellar nuclei, between various interneuron types, etc.), and we understand much of the cellular and molecular mechanisms at play. A complete review of these forms of plasticity and their mechanisms is outside the scope of this chapter, but see D'Angelo (2014) for a review.

Example models of supervised learning in the cerebellum

Classic models view the cerebellum as a neural implementation of a supervised-learning machine (Albus, 1971; Ito, 1984; Marr, 1969). In this conception, sensory input signals are carried by the mossy fibers, transformed into a more expansive basis set by the greatly divergent projections to the granule neurons, and ultimately transformed into the output signal by the granule neuron projections to Purkinje neurons. The ω parameters of Eq. 4.29 denote the synaptic strengths of the connections between granule and Purkinje neurons in this system. Climbing fibers from the inferior olive are thought to provide a supervised error or teaching signal that dictates plasticity at the granule neuron/Purkinje neuron synapse.

Owing largely to the homogeneity of anatomical circuitry across the cerebellum, this basic model has been proposed to apply to essentially every domain of cognition and action (Schmahmann, Guell, Stoodley, & Halko, 2019). However, likely because of the cerebellum's early association with motor function, the most clearly developed class of cerebellar-based supervised-learning models include models of motor planning and motor control – especially for arm-reaching movements (Schweighofer, Arbib, & Kawato, 1998; Schweighofer, Spoelstra, et al., 1998; Wolpert, Miall, & Kawato, 1998). In this case, all signals in Eq. 4.29 are considered to vary continuously in time, with output signals $y_j(t)$ conceived of as motor commands (i.e., muscle activation or joint torques), and input signals $x_i(t)$ conceived of as desired trajectories (i.e., position, velocity, and acceleration). In addition, the $\omega_{i,j}$ parameters represent synaptic weights between the granule cell and Purkinje cell layer, and the inferior olive is hypothesized to transmit a supervised error signal (actual trajectory minus desired trajectory).

4.5.5 Models of human learning that include multiple forms of plasticity

After long periods of practice, almost any behavior can be executed quickly, accurately, and with little or no conscious deliberation. At this point, we say that the behavior has become automatic. A strong case can be made that most behaviors performed by adults are automatic. When we sit in a chair, pick up a cup of coffee, or swerve to avoid a pothole, our actions are almost always automatic.

Automaticity could be viewed as the asymptotic state of learning. Ashby, Ennis, and Spiering (2007) proposed that skills learned procedurally are mediated entirely within cortex after they become automatized, and that the development of automaticity is associated with a gradual transfer of control from the striatum to cortical-cortical projections from the relevant sensory areas directly to the premotor areas that initiate the behavior. So in Figure 4.5, the cortical-cortical projections from visual cortex to the supplementary motor area eventually mediate the expression of automatic behaviors without any assistance from the subcortical loops through the basal ganglia. Therefore, according to this account, a critical function of the basal ganglia is to train purely cortical representations of automatic behaviors. Kovacs, Hélie, Tran, and Ashby (2021) proposed a similar account of how rule-guided behaviors are automatized in which the prefrontal cortex trains the cortical circuits that implement the automatic behaviors.

The Ashby et al. (2007) model was motivated by the observation that because cortical synaptic plasticity follows two-factor learning rules, the purely cortical circuits are incapable of learning any behavior that requires trial-by-trial feedback. Such behaviors require the three-factor plasticity of the basal ganglia. Ashby et al. (2007) proposed that the basal ganglia use DA-mediated three-factor learning (i.e., at cortical-striatal synapses) to gradually activate the correct postsynaptic targets in supplementary motor area, which thereby enables two-factor plasticity at cortical-cortical synapses to learn the correct associations (i.e., because there will be more postsynaptic activation at the correct synapses than at synapses leading to incorrect responses). As a result, in the full version of the Figure 4.5 model, plasticity at cortical-striatal synapses is modeled via three-factor learning rules (as in Eq. 4.35), whereas plasticity at cortical-cortical synapses is modeled via two-factor learning rules (as in Eq. 4.33).

This model accounts for many results that are problematic for other theories of automaticity. For example, it correctly predicts that people with Parkinson's disease, who have DA reductions and striatal dysfunction, are impaired in initial procedural learning (Soliveri, Brown, Jahanshahi, Caraceni, & Marsden, 1997; Thomas-Ollivier et al., 1999), but relatively normal in producing automatic skills (Asmus, Huber, Gasser, & Schöls, 2008). It also correctly predicts that blocking all striatal output to cortical motor and premotor targets does not disrupt the ability of monkeys to fluidly produce an overlearned motor sequence (Desmurget & Turner, 2010). Similarly, a neuroimaging study reported that activation in the putamen was correlated with performance of a procedural skill early in training but not after automaticity developed (Waldschmidt & Ashby, 2011). Instead, automatic performance was only correlated with activity in cortical areas (i.e., pre-supplementary motor area and supplementary motor area).

4.6 Empirical Testing

Of course, any psychological theory or model must eventually be tested against empirical data. In the case of learning models, this is especially challenging because, by definition, learning data are non-stationary. In fact, in some cases, the human learner could be in a different state on every trial of the experimental task. If so, then accurate estimation of that state is virtually impossible. In other words, learning data often provide, at best, a highly noisy sample of the learner's true state. As a result, model mimicry is perhaps a greater problem with models of learning than with models of other types of psychological phenomena – that is, learning data are often noisy enough that a less valid model could be statistically indistinguishable from a more valid model, based on goodness-of-fit alone. For these reasons, some extra steps are often needed to test models of learning.

One advantage of building models in which learning is mediated by the synaptic plasticity algorithms described in the previous sections, is that because of their biological constraints, such models tend to be mathematically rigid (Ashby, 2018). In other words, they tend to make a narrow set of predictions, regardless of how their free parameters are set. Because of this, in many cases, parameter-free *a priori* predictions are possible. For example, any model that assumes learning is based on DA-mediated synaptic plasticity that mimics reinforcement-learning algorithms must predict that omitting trial-by-trial feedback or even delaying feedback by just a few seconds should have devastating effects on learning.

Even if a model does not make *a priori* predictions in a given task, it may predict only a limited set of possible outcomes. If one of those outcomes is observed in an experiment, then a model predicting that this is one of the few outcomes possible should be favored over a model that can account for a wider variety of possible outcomes by manipulating free parameters in a *post hoc* manner. The method of parameter-space partitioning was designed to address this issue (Pitt, Kim, Navarro, & Myung, 2006). In particular, parameter-space partitioning estimates the volume of parameter space throughout which a model is consistent with a certain qualitative pattern of data. A parameter-space partitioning analysis is valuable with all kinds of modeling, but especially so with learning models because of the challenges their non-stationary nature presents to standard goodness-of-fit testing.

Other good model fitting practices are also recommended. For example, the models should be validated by simulating data under a variety of different parameter settings and then investigating under what conditions the generating parameter values can be recovered during the parameter estimation process.

When learning models are fit to behavioral data, the most common choice is to fit them to some form of empirical learning curve – most often a forward-learning curve, which plots proportion correct against trial or block number. As with all modeling, the most effective tests compare the fit of the model under investigation to some other established model from the literature. In the case of forward learning curves, a good choice for comparison is the exponential learning curve

$$P_n = P_\infty - (P_\infty - P_0)\,\mathrm{e}^{-\lambda n}, \tag{4.43}$$

where P_n is the probability correct on trial n, P_∞ and P_0 are asymptotic and initial accuracy, respectively, and λ is the learning rate. This model was proposed more than one hundred years ago (Thurstone, 1919), and remains popular today (e.g., Heathcote, Brown, and Mewhort 2000; Leibowitz, Baum, Enden, and Karniel 2010). As an example of how this model might be used, Cantwell et al. (2017) compared the fits of the exponential model and a biologically detailed model that assumes learning in procedural-memory-mediated tasks depends on three-factor plasticity (i.e., the model described in Figure 4.5) to learning curves from two separate experiments. In both cases, the biologically detailed model fit better than the exponential model.

Different learning strategies can produce qualitatively different learning curves. Procedural learning and instrumental conditioning predict incremental learning and gradual learning curves. In contrast, rule-guided learning predicts discrete and abrupt jumps in accuracy as the learner switches rules trial-by-trial. In many tasks, incorrect rules cause accuracy to be near chance, whereas the correct rule predicts perfect accuracy. In these cases, rule-learning strategies predict all-or-none learning curves.

Although incremental and all-or-none learning curves might seem easy to distinguish empirically, it has long been known that these differences can be obscured if the data are averaged across learners (Estes, 1956, 1964). In fact, it is well documented that averaging can change the psychological structure of many different types of data (Ashby, Maddox, & Lee, 1994; Maddox, 1999). As a result, averaging is typically inappropriate when testing models of how individuals learn. For example, if every learner's accuracy jumps from 50% to 100% correct on one trial, but the trial on which this jump occurs varies across participants, then the resulting averaged learning curve will be incremental – not all-or-none (Estes, 1956). The top panel of Figure 4.7 illustrates this phenomenon. This panel shows the traditional (forward) learning curve (i.e., mean accuracy across all participants on every trial) for 1,000 simulated participants who each display all-or-none learning. Specifically, each participant responds randomly with a probability correct of .5 until the correct strategy is discovered on some random trial (between 5 and 85), after which they respond perfectly. Note that the all-or-none nature of learning is completely obscured by the averaging process.

Hayes (1953) proposed the backward-learning curve as a solution to this problem. Backward-learning curves are most effective at discriminating between incremental and all-or-none learning in experiments where perfect

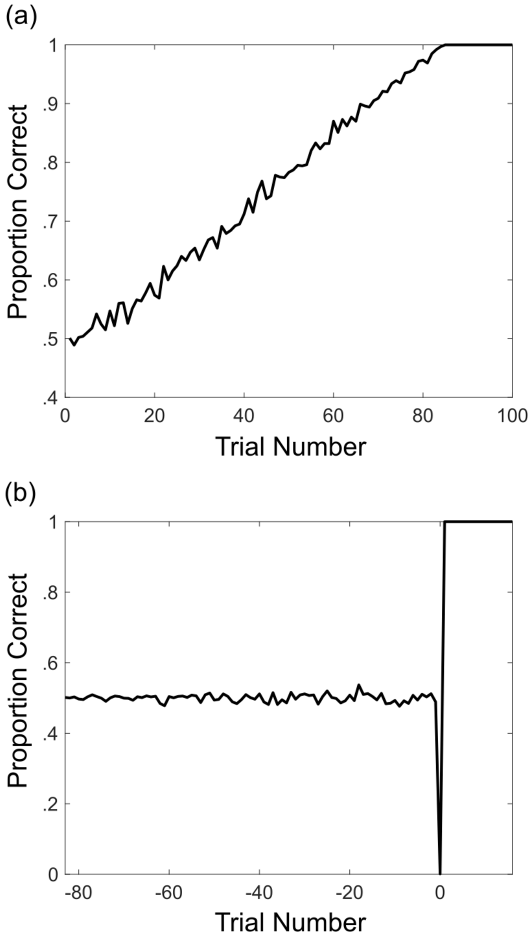

Figure 4.7 (a) Forward learning curve, which plots mean proportion correct on each trial for 1,000 simulated participants who are all characterized by one-trial learning in which accuracy jumps from .5 to 1 on one trial, but who all make this jump on a different random trial. (b) Backward learning curve of the same data.

accuracy is possible. The first step is to define a learning criterion, which is conservative enough to rule out guessing or partial learning. For example, consider a two-alternative task, like the one illustrated in Figure 4.7, in which the probability correct by guessing is .5 on each trial. Then a criterion of 10 consecutive correct responses is possible by guessing with a probability of less than .001. A backward-learning curve can only be estimated for participants who reach criterion, so the second step is to separate participants who reached criterion from those who did not. The most common analysis for nonlearners is to compare the proportion of nonlearners across conditions. The remaining steps proceed for all participants who reached criterion. Step 3 is to identify for each learner the trial number of the first correct response in the sequence of 10 correct responses that ended the learning phase. Let N_i denote this trial number for learner i. Then note that the response on trial N_i and the ensuing 9 trials were all correct. But also note that the response on the immediately preceding trial (i.e., trial $N_i - 1$) was necessarily an error. Step 4 is to renumber all the trial numbers so that trial N_i becomes trial 1 for every participant. Thus, for every participant, trials 1 – 10 are all correct responses and trial 0 is an error. The final step is to estimate a learning curve by averaging across learners. The bottom panel of Figure 4.7 shows the backward learning curve that results from this re-analysis of the data plotted in Figure 4.7a.

Because of our renumbering system, the mean accuracy for trials 1-10 will be 100% correct, and the mean accuracy for trial 0 will be 0% correct. Thus, if every learner shows a dramatic one-trial jump in accuracy, then the averaged accuracy on trial -1 should be low, even if the jump occurred on a different trial number for every participant (according to the original numbering system). In the Figure 4.7 example, all participants had perfect all-or-none, one-trial learning and note that the mean accuracy for all trials preceding trial 0 is at chance (i.e., .5). In contrast, if participants incrementally improve their accuracy then the averaged accuracy on trial -1 should be significantly higher than chance. So if one is interested in discriminating between strategies that predict incremental learning and strategies that predict all-or-none learning, then backward learning curves should be used rather than the more traditional forward learning curves.

Backward-learning curves are more problematic in tasks where most participants do not achieve perfect accuracy, because in these cases, it is usually impossible to define a learning criterion that ensures learning has terminated. Even so, if estimated with care, backward learning curves can be useful even in these more ambiguous cases (J. D. Smith & Ell, 2015).

4.7 Conclusions

Mathematical models of human learning have progressed enormously during the last century. After an initial period of intense activity that dominated experimental psychology during the first half of the 20th century, the field entered a lull that lasted for several decades. As we have described, several neuroscience breakthroughs reinvigorated the study of learning and the subsequent progress has been dramatic. Even so, the study of learning has not recaptured its formally prominent place within experimental psychology. For example, none of the leading textbooks on cognitive neuroscience currently include any chapters on learning. Learning is a fundamental component of the human experience, and we believe that the recent progress described in this chapter should re-establish the foundational role of learning, not only in mathematical psychology, but more generally within the cognitive sciences.

4.8 Related Literature

Many articles and texts review mathematical learning theory as it existed during the early years of mathematical psychology, including Atkinson et al. (1965), Bush and Estes (1959), and Laming (1973). No recent texts provide a similar comprehensive coverage. Even so, there are a variety of more specialized recent reviews. In the case of machine-learning, the classic text on reinforcement learning is Sutton and Barto (1998), whereas Neal (2012) covers Bayesian approaches. A number of computational neuroscience reviews include sections on learning, including Dayan and Abbott (2001) and Ashby (2018). For a review of the neurobiological foundations of learning (e.g., synaptic plasticity), see Rudy (2020).

4.9 Acknowledgments

We thank Sebastien Hélie and Michael Wenger for their helpful comments on this manuscript.

References

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, *10*(1-2), 25–61.

Alpaydin, E. (2020). *Introduction to machine learning.* Cambridge, MA: MIT press.

Apicella, P., Legallet, E., & Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research*, *116*(3), 456–466.

Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology, Volume 2* (pp. 223–270). New York: Cambridge University Press.

Ashby, F. G. (2019). *Statistical analysis of fMRI data, Second Edition.* Cambridge, MA: MIT Press.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442-481.

Ashby, F. G., & Crossley, M. J. (2011). A computational model of how cholinergic interneurons protect striatal-dependent learning. *Journal of Cognitive Neuroscience*, *23*(6), 1549-1566.

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1-36.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632-656.

Ashby, F. G., & Maddox, W. T. (1998). Stimulus categorization. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 251–301). San Diego, CA: Academic Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Science*, *2*, 83-89.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science, Second Edition* (pp. 157–188). New York: Elsevier.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363-378.

Asmus, F., Huber, H., Gasser, T., & Schöls, L. (2008). Kick and rush: Paradoxical kinesia in Parkinson disease. *Neurology*, *71*(9), 695.

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *Introduction to mathematical learning theory.* New York: Wiley.

Baddeley, R. J., Ingram, H. A., & Miall, R. C. (2003). System identification applied to a visuomotor task: Near-optimal human performance in a noisy changing task. *The Journal of Neuroscience*, *23*(7), 3066–3075.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129-141.

Bear, M., & Linden, D. (2001). The mechanisms and meaning of long-term synaptic depression in the mammalian brain. In W. Cowan, T. Sudhof, & C. Stevens (Eds.), *Synapses* (pp. 455–517). Baltimore, MD: Johns Hopkins University Press.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bi, G.-Q., & Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*(1), 139–166.

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in Neuroscience*, *6*, 85.

Bliss, T. V., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of Physiology*, *232*(2), 331–356.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1485), 1655–1670.

Bush, R. R., & Estes, W. K. (1959). *Studies in mathematical learning theory.* Stanford University Press.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413-423.

Caine, E. D., Weingartner, H., Ludlow, C. L., Cudahy, E. A., & Wehry, S. (1981). Qualitative analysis of scopolamine-induced amnesia. *Psychopharmacology*, *74*(1), 74–80.

Calabresi, P., Pisani, A., Mercuri, N. B., & Bernardi, G. (1996). The corticostriatal projection: From synaptic plasticity to dysfunctions of the basal ganglia. *Trends in Neurosciences*, *19*, 19–24.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*(6), 1598–1613.

Cantwell, G., Riesenhuber, M., Roeder, J. L., & Ashby, F. G. (2017). Perceptual category learning and visual processing: An exercise in computational cognitive neuroscience. *Neural Networks*, *89*, 31–38.

Cheng, S., & Sabes, P. N. (2006). Modeling sensorimotor learning with linear dynamical systems. *Neural Computation*, *18*, 760–793.

Coesmans, M., Weber, J. T., De Zeeuw, C. I., & Hansel, C. (2004). Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control. *Neuron*, *44*(4), 691–700.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2013). Erasing the engram: The unlearning of procedural skills. *Journal of Experimental Psychology: General*, *142*(3), 710-741.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2014). Context-dependent savings in procedural category learning. *Brain & Cognition*, *92*, 1-10.

Crossley, M. J., Horvitz, J. C., Balsam, P. D., & Ashby, F. G. (2016). Expanding the role of striatal cholinergic interneurons and the midbrain dopamine system in appetitive instrumental conditioning. *Journal of Neurophysiology*, *115*, 240-254.

Crow, T. J., & Grove-White, I. G. (1973). An analysis of the learning deficit following hyoscine administration to man. *British Journal of Pharmacology*, *49*(2), 322–327.

Cunningham, H. A. (1989). Aiming error under transformed spatial mappings suggests a structure for visual-motor maps. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 493–506.

D'Angelo, E. (2014). The organization of plasticity in the cerebellar cortex: From synapses to control. In *Progress in Brain Research* (Vol. 210, pp. 31–58). Elsevier.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems.* Cambridge, MA: MIT Press.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth national/tenth conference on artificial intelligence/innovative applications of artificial intelligence* (pp. 761–768). Menlo Park, CA: AAAI Press.

Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K. E., . . . Schlagenhauf, F. (2020). Volatility estimates increase choice switching and relate to prefrontal activity in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(2), 173–183.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: Kinematics, not habits. *Journal of Neuroscience*, *30*(22), 7685–7690.

Diaconescu, A. O., Mathys, C., Weber, L. A., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634.

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. *Stevens' handbook of experimental psychology*.

Donchin, O., Francis, J. T., & Shadmehr, R. (2003). Quantifying generalization from trial-by-trial behavior of adaptive systems that learn with basis functions: Theory and experiments in human motor control. *Journal of Neuroscience*, *23*(27), 9032–9045.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7-8), 961–974.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840-859.

Ebner, T. J., & Pasalar, S. (2008). Cerebellum predicts the future motor state. *The Cerebellum*, *7*(4), 583–588.

Eccles, J. C., Ito, M., & Szentágothai, J. (1967). *The cerebellum as a neuronal machine.* New York: Springer.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain.* Oxford University Press.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94–107.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Estes, W. K. (1964). All-or-none processes in learning and retention. *American Psychologist*, *19*(1), 16–25.

Fanselow, M. S., Zelikowsky, M., Perusini, J., Barrera, V. R., & Hersman, S. (2014). Isomorphisms between psychological processes and neural mechanisms: From stimulus elements to genetic markers of activity. *Neurobiology of Learning and Memory*, *108*, 5–13.

Feenstra, M. G., & Botterblom, M. H. (1996). Rapid sampling of extracellular dopamine in the rat prefrontal cortex during food consumption, handling and exposure to novelty. *Brain Research*, *742*(1), 17–24.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234.

Ghoneim, M., & Mewaldt, S. (1975). Effects of diazepam and scopolamine on storage, retrieval and organizational processes in memory. *Psychopharmacologia*, *44*(3), 257–262.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15647–15654.

Gomi, H., Shidara, M., Takemura, A., Inoue, Y., Kawano, K., & Kawato, M. (1998). Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkey I. Simple spikes. *Journal of Neurophysiology*, *80*(2), 818–831.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greeno, J. G., & Bjork, R. A. (1973). Mathematical learning theory and the new "mental forestry". *Annual Review of Psychology*, *24*(1), 81–116.

Grossberg, S. (1972). Neural expectation: Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, *10*(1), 49–57.

Gu, Q. (2003). Contribution of acetylcholine to visual cortex plasticity. *Neurobiology of Learning and Memory*, *80*(3), 291–301.

Gulliksen, H. (1934). A rational equation of the learning curve based on Thorndike's law of effect. *The Journal of General Psychology*, *11*(2), 395–434.

Guthrie, E. R. (1935). *Psychology of learning.* New York: Harper & Row.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1-2), 1–34.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review*, *60*(4), 269–275.

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

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.

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

Howard, M. W. (this volume). Formal models of memory based on temporally-varying representations. In F. G. Ashby, H. Colonius, & E. N. Dzhafarov (Eds.), *New handbook of mathematical psychology, volume 3: Perceptual and cognitive processes.* Cambridge University Press.

Hull, C. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519–530.

Ito, M. (1984). *The cerebellum and neural control.* New York: Raven Press Books.

Ito, M., Sakurai, M., & Tongroach, P. (1982). Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *The Journal of Physiology*, *324*(1), 113–134.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569-1572.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4-6), 535–547.

Kemp, N., & Bashir, Z. I. (2001). Long-term depression: A cascade of induction and expression mechanisms. *Progress in Neurobiology*, *65*(4), 339–365.

Kennedy, A. (2019). Learning with naturalistic odor representations in a dynamic model of the Drosophila olfactory system. *bioRxiv*, 783191.

Kitazawa, S., Kimura, T., & Yin, P.-B. (1998). Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature*, *392*(6675), 494–497.

Kobayashi, Y., Kawano, K., Takemura, A., Inoue, Y., Kitama, T., Gomi, H., & Kawato, M. (1998). Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys II. Complex spikes. *Journal of Neurophysiology*, *80*(2), 832–848.

Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological Review*, *128*(3), 488–508.

Krakauer, J. W., Pine, Z. M., Ghilardi, M.-F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *The*

Journal of Neuroscience, *20*(23), 8916–8924.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–247.

Kruschke, J. K. (2011). *Doing Bayesian data analysis.* Burlinton, MA: Academic Press.

Laming, D. R. J. (1973). *Mathematical psychology.* Academic Press.

Lapish, C. C., Kroener, S., Durstewitz, D., Lavin, A., & Seamans, J. K. (2007). The ability of the mesocortical dopamine system to operate in distinct temporal modes. *Psychopharmacology*, *191*(3), 609–625.

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*(3), 338–340.

Lev-Ram, V., Mehta, S. B., Kleinfeld, D., & Tsien, R. Y. (2003). Reversing cerebellar long-term depression. *Proceedings of the National Academy of Sciences*, *100*(26), 15989–15993.

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of CaMKII function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, *3*(3), 175–190.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*(2), 354–374.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 650-662.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, *74*(2), 219-236.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100-107.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, *202*, 437–470.

Martin, S., Grimwood, P., & Morris, R. (2000). Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience*, *23*(1), 649–711.

Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996a). Throwing while looking through prisms: I. Focal olivocerebellar lesions impair adaptation. *Brain*, *119*(4), 1183–1198.

Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996b). Throwing while looking through prisms: II. Specificity and storage of multiple gaze–throw calibrations. *Brain*, *119*(4), 1199–1211.

Mathys, C. D., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825.

McCoy, P. A., Huang, H.-S., & Philpot, B. D. (2009). Advances in understanding visual cortex plasticity. *Current Opinion in Neurobiology*, *19*(3), 298–304.

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

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of human learning and memory* (p. 65-77). New York: Guilford Press.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning.* Cambridge, MA: MIT press.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 689.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 855–874.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.

Nicoll, R. A. (2017). A brief history of long-term potentiation. *Neuron*, *93*(2), 281–290.

Ninkovic, J., & Bally-Cuif, L. (2006). The zebrafish as a model system for assessing the reinforcing properties of drugs of abuse. *Methods*, *39*(3), 262–274.

O'Reilly, R. C., Munakata, Y., Frank, M., Hazy, T., et al. (2012). *Computational cognitive neuroscience.* Mainz, Germany: PediaPress.

Ostfeld, A. M., & Aruguete, A. (1962). Central nervous system effects of hyoscine in man. *Journal of Pharmacology and Experimental Therapeutics*, *137*(1), 133–139.

Pakhotin, P., & Bracci, E. (2007). Cholinergic interneurons control the excitatory input to the striatum. *The Journal of Neuroscience*, *27*(2), 391–400.

Palay, S. L., & Chan-Palay, V. (2012). *Cerebellar cortex: Cytology and organization.* Springer Science & Business Media.

Paliwal, S., Mosley, P., Breakspear, M., Coyne, T., Silburn, P., Aponte, E., ... Klaas, S. (2018). Subjective estimates of uncertainty and volatility during gambling predict impulsivity after subthalamic deep brain stimulation for Parkinson's disease. *BioRxiv*, 477364.

Paliwal, S., Petzschner, F. H., Schmitz, A. K., Tittgemeyer, M., & Stephan, K. E. (2014). A model-based analysis of impulsivity using a slot-machine gambling paradigm. *Frontiers in Human Neuroscience*, *8*, 428.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* Translated and edited by Anrep, GV (Oxford University Press, London, 1927).

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*(1).

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*(1), 319–330.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83.

Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550.

Rall, W. (1967). Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. *Journal of Neurophysiology*, *30*(5), 1138-1168.

Ramnani, N. (2006). The primate cortico-cerebellar system: Anatomy and function. *Nature Reviews Neuroscience*, *7*(7), 511–522.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163–178.

Redding, G. M., Rossetti, Y., & Wallace, B. (2005). Applications of prism adaptation: A tutorial in theory and method. *Neuroscience & Biobehavioral Reviews*, *29*(3), 431–444.

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

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

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(3), 365–372.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rudy, J. W. (2020). *The neurobiology of learning and memory, Third edition.* Oxford University Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1.* Cambridge, MA: MIT Press.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative and Physiological Psychology*, *74*, 192-202.

Sawtell, N. B. (2010). Multimodal integration in granule cells as a basis for associative plasticity and sensory prediction in a cerebellum-like circuit. *Neuron*, *66*(4), 573–584.

Scheidt, R. A., Dingwell, J. B., & Mussa-Ivaldi, F. A. (2001). Learning to move amid uncertainty. *Journal of Neurophysiology*, *86*(2), 971–985.

Schmahmann, J. D., Guell, X., Stoodley, C. J., & Halko, M. A. (2019). The theory and neuroscience of cerebellar cognition. *Annual Review of Neuroscience*, *42*(1), 337–364.

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

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

Schweighofer, N., Arbib, M. A., & Kawato, M. (1998). Role of the cerebellum in reaching movements in humans. I. Distributed inverse dynamics control. *European Journal of Neuroscience*, *10*(1), 86–94.

Schweighofer, N., Spoelstra, J., Arbib, M. A., & Kawato, M. (1998). Role of the cerebellum in reaching movements in humans. II. A neural model of the intermediate cerebellum. *European Journal of Neuroscience*, *10*(1), 95–105.

Seamans, J. K., & Robbins, T. W. (2010). Dopamine modulation of the prefrontal cortex and cognitive function. In K. A. Neve (Ed.), *The dopamine receptors, 2nd edition* (pp. 373–398). New York: Springer.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental Brain Research*, *185*(3), 359–381.

Shidara, M., Kawano, K., Gomi, H., & Kawato, M. (1993). Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. *Nature*, *365*(6441), 50–52.

Sjöström, P. J., Rancz, E. A., Roth, A., & Häusser, M. (2008). Dendritic excitability and synaptic plasticity. *Physiological Reviews*, *88*(2), 769–840.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century.

Smith, J. D., & Ell, S. W. (2015). One giant leap for categorizers: One small step for categorization theory. *PloS One*, *10*(9).

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168.

Soliveri, P., Brown, R. G., Jahanshahi, M., Caraceni, T., & Marsden, C. D. (1997). Learning manual pursuit tracking skills in patients with Parkinson's disease. *Brain*, *120*(8), 1325–1337.

Soto, F. A., & Wasserman, E. A. (2010). Error-driven learning in visual categorization and object recognition: A common-elements model. *Psychological Review*, *117*(2), 349–381.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171-177.

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

Takahashi, Y., Schoenbaum, G., & Niv, Y. (2008). Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, *2*, 14.

Thomas-Ollivier, V., Reymann, J., Le Moal, S., Schück, S., Lieury, A., & Allain, H. (1999). Procedural memory in recent-onset Parkinson's disease. *Dementia and Geriatric Dognitive Disorders*, *10*(2), 172–180.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Thoroughman, K. A., & Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature*, *407*(6805), 742–747.

Thurstone, L. L. (1919). The learning curve equation. *Psychological Monographs*, *26*(3), i–51.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *The Journal of Neuroscience*, *23*(32), 10402–10410.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tulving, E., & Craik, F. I. (2000). *The Oxford handbook of memory.* New York: Oxford University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. *Frontiers in Psychology*, *5*(643).

van Woerden, G. M., Hoebeek, F. E., Gao, Z., Nagaraja, R. Y., Hoogenraad, C. C., Kushner, S. A., … Elgersma, Y. (2009). βCaMKII controls the direction of plasticity at parallel fiber–Purkinje cell synapses. *Nature Neuroscience*, *12*(7), 823–825.

Von Helmholtz, H. (1925). *Helmholtz's treatise on physiological optics* (Vol. 3). Optical Society of America.

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *NeuroImage*, *56*(3), 1791-1802.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177.

Weilnhammer, V. A., Stuke, H., Sterzer, P., & Schmack, K. (2018). The neural correlates of hierarchical predictions for perceptual decisions. *Journal of Neuroscience*, *38*(21), 5008–5021.

Weiss, E. J., & Flanders, M. (2011). Somatosensory comparison during haptic tracing. *Cerebral Cortex*, *21*(2), 425–434.

Welch, R. B. (1986). Adaptation of space perception. *Handbook of perception and human performance*, *1*(24), 2424–45.

Wickens, J. (1993). *A theory of the striatum.* Pergamon Press.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (Tech. Rep.). Stanford University, California, Stanford Electronics Labs.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*(3), 558-584.

Wills, A., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition*, *34*(1), 17-27.

Wilson, C. J. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 29-50). Cambridge, MA: MIT Press.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Computational Biology*, *9*(7).

Wolpert, D. M., Miall, R., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences*, *2*(9), 338–347.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*(2), 283-293.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., 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.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., & Poo, M.-M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, *395*(6697), 37–44.