Chapter 13

Neuroscience

Neuroscience is the multi-disciplinary study of nervous systems: how they keep animals' bodies functioning, how they control behavior, how they change over time, and how the underlying chemistry and physics makes all of this possible. Computer science and engineering are among many disciplines taking part in neuroscience through contributions to theories, models, and experimental technology. At the same time, neuroscience is contributing to computer science and engineering by inspiring the design of brain-like architectures for computing and engineering applications.

Neuroscience research has long tried to answer questions about how an animal's nervous system makes learning possible. Many different types of learning can be studied, ranging from animal learning in the laboratory to the learning we wish occurs in our schools. Not surprisingly, the former has received the most attention in neuroscience due to the relative ease of controlling conditions and making observations. Neuroscience research on learning has had a particularly strong influence on computer science and engineering, notably leading to artificial neural networks that implement learning algorithms using brain-like mechanisms. As we recounted in Section **??**, many aspects of the computational approach to reinforcement learning were inspired by what neuroscience tells us about brain mechanisms underlying learning.

It has turned out that computational reinforcement learning is having a remarkable influence in the other direction. Reinforcement learning theory and algorithms are providing neuroscientists with conceptual tools for thinking about reward-based learning in the brain. They are inspiring a literal flood of experiments investigating how reward-related processes actually work in the brains of vertebrate and invertebrate animals. In addition to its impact on the neuroscience of learning, reinforcement learning is among several disciplines contributing to the neuroscience of decision making in humans and non-human primates. Together with economics, evolutionary biology, and mathematical psychology, reinforcement learning theory is helping to formulate quantitative models of the neural mechanisms of choice.

At the root of reinforcement learning's influence on neuroscience is its embrace of learning scenarios that relate well to those faced by animals, combined with well-defined algorithmic structures that suggest a wealth of questions that lend themselves to experimental investigation. The most striking link between the computational study of reinforcement learning and neuroscience is the parallel between TD algorithms for predicting future reward and some properties of the mammalian dopamine system. The neurotransmitter dopamine had long been thought to transmit reward signals throughout the brain, but experiments in the 1990s revealed that story to be too simple. TD learning—which had been developed many years earlier for its computational properties—was quickly recognized as a ready-made model able to account for many of these new observations. Closely coupled to this is reinforcement learning's account of how a widely broadcast signal, like the brain's dopamine signal, can shape the behavior of a collection of many learning elements. Although the actions of dopamine are still not understood, computational reinforcement learning provides a compelling framework that is illuminating important aspects of dopamine's function as a widely distributed neuromodulator in the brain.

This chapter's objective is to provide the basic background needed to appreciate these and other correspondences responsible for the influence that reinforcement learning is having on neuroscience—and that it is likely to have over the future. But the chapter is too short to cover all the relevant correspondences. Neuroscience specifically focusing on brain reward processes is an extremely vast and fast-moving field. The correspondences we include here are those accounting for the strongest impact reinforcement learning has had, and is continuing to have, on neuroscience. But we can only provide a glimpse into this fascinating story.

It hardly needs emphasizing that this chapter is also too short to delve very deeply into the enormous complexity of the neural systems underlying reward-based learning and decision making. We provide a brief introduction to concepts and terminology from neuroscience that the reader needs for our discussion, but our treatment is very simplified from a neuroscience perspective. We do not attempt to describe—or even to name—the very many brain structures and pathways, or any of the molecular mechanisms, believed to be involved in these processes. We also do not do justice to the variety of hypotheses and models that represent alternative views.

The literature contains many excellent publications covering links between reinforcement learning and neuroscience at various levels and from various perspectives. Our treatment differs from most of these because we assume the reader is familiar with reinforcement learning as presented in the earlier chapters of this book, and we do not assume any knowledge of neuroscience. This chapter's final section provides information about the sources of our remarks along with commentary on how the various links between the computational theory and neuroscience developed.

13.1 Levels of Explanation

Understanding complex systems involves multiple levels of explanation, a perspective that is a valuable guide to understanding parallels between computational reinforcement learning and neuroscience. The three levels at which complex information-processing systems can be understood proposed by David Marr have proven to be useful in organizing thinking in both neuroscience and artificial intelligence (Marr, 1982). The *computational theory* level, the most abstract, is about *what problem* the system is solving and and theory related to that problem. The *representation and algorithm* level addresses *how* the system solves the problem in terms of representations and algorithms. The *hardware implementation* level, the most concrete, is about the physical substrate that actually performs the computation. There are no unique correspondences between these levels. Many different representations and algorithms are capable of solving any given problem, and many different physical substrates can implement any representation and algorithm.

At the computational theory level, reinforcement learning's focus on prediction and control in the context of stochastic sequential decisions problems is as applicable to animals as it is to artificial intelligence. Its emphasis on observing, predicting, and controlling in continual real-time interaction with an environment in order to achieve goals under conditions of uncertainty is a powerful guide to thinking in both cases. Reinforcement learning thus has an advantage over many other approaches to thinking about intelligence because it addresses a problem not unlike the problem that has driven the evolution of animal intelligence. At the computational level, therefore, the parallel between reinforcement learning and animal intelligence has a strong footing.

But the strongest parallels may be at the representation and algorithm level, the level of this chapter's focus. We have presented algorithms whose main features are numerical reward signals, value functions, TD errors, eligibility traces, environment models, and adjustable parameters, or weights. Here we examine parallels between these elements and what neuroscience tells us about how nervous systems address the kinds of problems in which we are interested.

Reinforcement learning theory has less to say at Marr's hardware implementation level. We describe algorithms using mathematics and computer code without paying attention to how they could be implemented in computational hardware. Nevertheless, some features of reinforcement learning algorithms were inspired by hypotheses about neural learning mechanisms, and some algorithms have proven to be highly suggestive about how neural hardware could implement them. We discuss hypothetical neural implementations of these features and algorithms, especially those that have influenced neuroscientists in their efforts to understand implementation-level information about reward-based learning and decision making.

Throughout all of what follows it is important to keep in mind that something at one level of analysis can be realized in many different ways at more detailed levels. Different algorithms can solve stochastic sequential decision problems, and each can be implemented with different hardware. A construct useful on one level may not correspond in a unique way to a construct at a more detailed level. For example, it can be useful for neuroscientists to think about a single real-valued reward signal even if there is no literal counterpart among the signals produced by any single neuron.

Despite the fact that any algorithm can be implemented in many different ways, hardware realizations are not arbitrary. Constraints from a number of sources narrow the set of plausible implementations. Energy requirements, signal timing, wiring efficiency and other aspects of packaging place strong constraints at the hardware implementation level. Computational efficiency and complexity, in both time and space, constrain the set of feasible algorithms for solving computational problems. The concern of reinforcement learning research with comparing algorithms in terms performance and complexity mirrors the role these same constraints played in the evolution of the algorithms implemented by animal nervous systems. We believe this is the reason that many features of reinforcement learning algorithms align well with what neuroscience is discovering about learning in animal brains.

13.2 Some Neuroscience Basics

Understanding some of the the very basics of neuroscience will help in following the content of this chapter. We briefly cover what is needed in this section, which can easily be skipped if you already know something about neuroscience. As is true in most introductions to the principal features of nervous systems, we describe just the most typical versions of various neural building blocks. Nervous systems employ an enormous range of variations of these basic elements.

Neurons, the main components of nervous systems, are cells specialized for processing and transmitting information using electrical and chemical signals. They come in many forms, but a neuron typically has a cell body, *dendrites*, and a single *axon*. Dendrites are fibers branching from the cell body to receive input from other neurons (or to receive external signals in the case of sensory neurons). A neuron's axon is a fiber that delivers the neuron's output to other neurons (or to muscles or glands). The output consists of sequences of electrical pulses called *action potentials* that travel along the axon. Action potentials are also called *spikes*, and a neuron is said to *fire* when it generates one. In models of neural networks it is common to use a real number to represent a neuron's *firing rate*.

A *synapse* is a structure generally at the termination of an axon branch that mediates the communication of one neuron to another. A synapse transmits information from the *presynaptic* neuron's axon to a dendrite or cell body of the *postsynaptic* neuron. Of major interest here are synapses that release a chemical *neurotransmitter* upon the arrival of an action potential from the presynaptic neuron. Neurotransmitter molecules bind to receptors on the surface of the postsynaptic neuron to excite or inhibit its spike-generating activity, or to modulate its behavior in other ways. A given neurotransmitter may bind to several different types of receptors, with each producing a different effect on the postsynaptic neuron. For example, there are at least five different receptor types by which the neurotransmitter dopamine can affect a postsynaptic neuron.

Neurons typically exhibit a *background* level of activity either when they are not being activated by synaptic input or their synaptic input is not related to task-specific brain activity. Background activity can be irregular as the result of noise within the neuron or its synapses, or it can be periodic due to dynamic processes intrinsic to the neuron. A neuron's *phasic* activity, in contrast, consists of bursts of spiking activity usually generated by synaptic input.

A neuron's axon can widely branch so that action potentials reach many targets. The branching structure of a neuron's axon is called the neuron's *axonal arbor*. Since the conduction of an action potential is an active process not unlike the burning of a fuse, when an action potential arrives at an axonal branching point it "lights up" action potentials on all the outgoing branches. This implies that a neuron with a large axonal arbor can exert approximately equal influence onto many target sites.

Neurotransmitters are called *neuromodulators* if they are distributed widely throughout the brain instead of targeting specific synapses. Brains contain several different neuromodulation systems consisting of clusters of neurons with widely branching axonal arbors, with each system using a different neurotransmitter. Neuromodulation can alter the function of neural circuits and mediate motivation, arousal, attention, memory, mood, emotion, sleep, and body temperature. Important here is that a neuromodulatory system can distribute something like a scalar signal, such as a reinforcement signal, to alter the operation of synapses in widely distributed sites critical for learning.

The strength or effectiveness by which a synapse's neurotransmitter release influences the postsynaptic neuron is the synapse's *efficacy.* One way a nervous system can change through experience is through changes in synaptic efficacies as a function of combinations of presynaptic, postsynaptic, and neuromodulatory activity. The ability of synaptic efficacies to change is called *synaptic plasticity.* It is a primary mechanism responsible for learning. The parameters, or weights, adjusted by learning algorithms correspond to synaptic efficacies. As we detail below, modulation of synaptic plasticity via dopamine is a plausible mechanism for how the brain might implement learning algorithms like many of those described in this book.

13.3 Reward Signals, Values, Prediction Errors, and Reinforcement Signals

Links between neuroscience and computational reinforcement learning begin by drawing parallels between neural signals and signals playing prominent roles in reinforcement learning algorithms. These include reward signals, values, prediction errors, and reinforcement signals. Searching for neural analogs of any of these involves many challenges. Signals related to reward processing can be found in nearly every part of the brain, but since representations of different reward-related signals tend to be highly correlated with one another it is difficult to interpret results unambiguously. Experiments need to be very carefully designed to create conditions under which one of these signals might be distinguished from the others—or from an abundance of other types of signals—with any degree of certainty. Despite these difficulties, many experiments have been conducted with the aim of reconciling aspects of reinforcement learning theory with neural mechanisms, and some compelling links have been established. In preparation for discussing these links, in this section we remind the reader of what various reward-related signals mean according to reinforcement learning theory, and we include some preliminary comments about how they might relate to signals in the brain.

The reward signal, R, as we define it, specifies what is intrinsically good or bad for an agent. It provides an immediate assessment of the desirability of a state, or of an action taken in a state. The reward signal defines the problem an agent is learning to solve. Relating this to biology, R signals what for an an animal would be *primary reward*, meaning the quality an animal attaches to biologically significant sensations or events, including those necessary for survival and reproduction, such as eating, sexual contact, and successful escape or aggression. While reinforcement learning theory focuses only on R's role in learning, neural signals related to primary rewards have many effects in addition their role in learning: they can produce subjective sensations of pleasure or pain (hedonic reactions), and they can activate and invigorate behavior (their role in engaging motivational systems). As we shall see, a unitary master primary reward signal like R may not exist in an animal's brain. In relating R to neural signals it is best to think of it as an abstraction summarizing the overall effect of a multitude of neural signals generated by multiple sources of primary reward.

Values, V, specify what is good or bad for the agent over the long run. They are estimates of the total reward an agent can expect to accumulate over the future. Values are attached to states or to state-action pairs. Agents make good decisions by selecting actions with the largest values for a current state, or by selecting actions leading to states with the highest values. Because they predict reward, values are the basis of *secondary reinforcement* (Section **??**). Researchers taking an economics perspective distinguish between goal, or outcome, values and decision values, where only the latter take the cost of a decision into account. In psychology and neuroscience stimuli are said to have reward values, which are measures of how avidly the animal will work for them. In our more abstract formulation, these types of values are implicitly present in the process that generates the reward signal R as a function of states and actions. We reserve the term value for an estimate of how much reward is expected over the future, where it can be either a state value or an action value.

Prediction errors measure discrepancies between expected and actual signals or observations. Reward prediction errors (RPEs) specifically measure discrepancies between the expected and the received reward signal, being positive when the reward signal is greater than expected, and negative otherwise. Temporal difference (TD) errors, δ, are special kinds of prediction errors that signal discrepancies between old and new expectations of a long-term measure of future observations of some numerical-valued environment feature. This feature does not have to be a reward signal, but most relevant in our treatment are TD errors where the predicted feature is a reward signal, making these errors examples of RPEs. When neuroscientists refer to an RPE they generally (though not always) mean a TD RPE, which we simply call a TD error throughout this chapter.

There are two basic types of TD errors depending on whether or not they involve the agent's actions. TD errors that involve actions appear in algorithms for learning action-value functions, such as Q-learning and Sarsa, and TD errors that do not involve on actions appear in algorithms for learning state-value functions, such as the TD(λ) family of algorithms. When we refer to a TD error in this chapter we are generally referring one that does not involve actions because the most well-known link to neuroscience is stated in terms of TD errors that do not involve actions. However, given the difficulty mentioned above of experimentally distinguishing among the abundance of reward-related signals that can be observed in the brain, this emphasis should not be interpreted as ruling out similar links involving action-dependent TD errors.

An additional type of signal is a *reinforcement signal*, A reinforcement signal modulates learning by directing what a learning algorithm should do. In reinforcement learning it is a signed scalar that is multiplied by a vector (and some constants) to determine parameter updates within some learning system. It is important to distinguish between reward signals and reinforcement signals. Reward signals might function as reinforcement signals, they can be components of reinforcement signals, as they are in RPEs, and some reinforcement signals may not involve reward signals at all, such as prediction errors for learning environment models. In neuroscience, reinforcement signals are often called "teaching signals," terminology we do not use because in machine learning it it is usually connected with supervised learning, or "learning with a teacher," in which the teacher is the source of training examples that specify desired responses—the labels of labeled examples. A reinforcement signal for us may signal errors, as in prediction and supervised learning tasks, or it may be an evaluation signal.

Important questions to ask about links between neuroscience data and these theoretical concepts is if an observed signal is more like a reward signal, a value signal, a prediction error, a reinforcement signal, or something altogether different. And if it is an error signal, is it an RPE, a TD error, or a simpler error like the Rescorla-Wagner error (Equation **??**)? And if it is a TD error, does it depend on actions like a Q-learning or a Sarsa TD error? As indicated above, probing the brain to try to answer questions like these is extremely difficult. Most of these theoretical distinctions lie beyond what current experimental techniques can determine with any degree of certainty.

However, experiments performed in the early 1990s in which activities of single dopamine neurons were recorded in awake behaving monkeys provided evidence that the signals produced by these neurons are RPEs. Specifically, the phasic activity of these neurons is well characterized as signaling TD errors, leading to the *reward prediction error hypothesis of dopamine neuron activity* which we describe in the next section. This hypothesis does not specify which exact variety of TD error best fits this activity, but it is based on convincing experimental evidence, described below, that dopamine neuron activity is consistent with the general principles of a TD error. The Bibliographical and Historical Remarks section at the end of this chapter chronicles the development of this influential hypothesis, including early contributions from experiments with honeybee learning where the octopamine system has many parallels with the mammalian dopamine system.

13.4 The Reward Prediction Error Hypothesis

The hypothesis that a part of the function of the phasic activity of dopamine-producing neurons in mammals is to signal an error between an old and a new estimate of expected future reward is called the *reward prediction error hypothesis of dopamine neuron activity* (Montague, Dayan, and Sejnowski,1996; Schultz, Dayan, and Montague, 1997). It is the hypothesis that the TD error concept from reinforcement learning accounts for many features of the phasic activity of dopamine-producing neurons in the mammalian brain. In our notation a basic TD error is $\delta_t = R_t + \gamma V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1})$, where R_t is the reward signal at time t, and $V_t(\mathbf{x}_t)$ and $V_t(\mathbf{x}_{t-1})$ are respectively the estimates at time t of the values of the states present at times t and $t-1$ represented by the feature vectors $\mathbf{x_t}$ and $\mathbf{x}_{t-1}$.

Modeling dopamine neuron activity in terms of this definition requires several assumptions as explained by Montague et al. (1996) and Schultz et al. (1997). First, since a TD error can become negative, but neurons cannot fire at negative rates, the quantity corresponding to dopamine neuron activity is assumed to be $\delta_t + b_t$, where b_t is the background firing rate of the neuron. A negative TD error therefore corresponds to a drop in dopamine neuron firing rate below its background rate. Second, an assumption is needed about the states visited in an analog of an animal experiment and how they are represented by sensory cues or internal signals. This is the same issue faced with the TD model of classical conditioning described in Section **??**. Montague et al. (1996) and Schultz et al. (1997) assumed that stimulus cues are represented over time as a complete serial compound (CSC) representation as shown in the first column of Figure **??**. As in modeling classical conditioning, very little is known about how the brain internally represents time sequences of stimuli. Representations other than the CSC representation are clearly possible, but the major implications of the reward prediction error hypothesis are not overly sensitive to details of the representation.

During TD learning with these assumptions, together with the assumption of a linear value function representation and weights updated as in the TD model of classical conditioning (Equations **??** and **??**), TD errors exhibit the most salient features observed in the phasic activity of dopamine-producing neurons while an animal is engaged in a variety of classical and instrumental learning tasks. These features—which we describe in detail in Section 13.5 along with the experimental evidence for them—are 1) the phasic response of a dopamine neuron only occurs when a rewarding event is unpredicted, 2) early in learning neutral cues that precede reward do not cause phasic dopamine responses, but with continued learning these cues gain predictive value and come to elicit phasic dopamine responses, 3) if an even earlier cue reliably precedes a cue that has already acquired predictive value, the phasic dopamine response shifts to the earlier cue, ceasing for the later cue, and 3) if after learning, the predicted rewarding event is omitted, a dopamine neuron's response decreases below its baseline level at the expected time of reward delivery.

Although the reward prediction error hypothesis does not account for all aspects of dopamine neuron activity, and there are some features of dopamine neuron phasic activity for which does not account (discussed in Section **??**), it has has received wide acceptance among neuroscientists studying the neuroscience of reward-based behavior.

Before turning to what what reinforcement learning theory suggests about the function of dopamine signaling in the brain, we look more carefully at dopamine and the main experimental evidence supporting the reward prediction error hypothesis.

13.5 Reward Prediction Error Hypothesis: Experimental Support

The neurotransmitter dopamine plays essential roles in many processes in the mammalian brain. Prominent among these are processes underlying motivation, learning, action-selection, most forms of addiction, and the disorders schizophrenia and Parkinson's disease. Dopamine is a neuromodulator because it performs many functions and is widely distributed across the brain. Although much remains unknown about dopamine's functions and details of its cellular effects, it is clear that it is fundamental to reward processing in the mammalian brain. It is not the only neuromodulator involved in reward processing, and it can function differently in non-mammals, but it is a solid fact that dopamine is essential for processing reward-related information in mammals, including humans.

The traditional view is that dopamine broadcasts a reward signal to multiple brain regions involved in motivation and learning. At the root of this view is a famous 1954 paper by James Olds and Peter Milner that described the effects of electrical stimulation on certain brain areas of a rat's brain. They found that electrical stimulation to particular regions acted as a very powerful reward in controlling the rats behavior: "... the control exercised over the animals behavior by means of this reward is extreme, possibly exceeding that exercised by any other reward previously used in animal experimentation" (Olds and Milner, 1954). Later research revealed that the sites at which stimulation was most effective in producing this rewarding effect excited dopamine pathways, either directly or indirectly, that ordinarily are excited by natural rewarding stimuli. Other studies showed that blocking the effects of dopamine on brain regions targeted by dopamine pathways impairs learning. Effects similar to these with rats have also been observed with human subjects.

Dopamine is produced as a neurotransmitter by neurons whose cell bodies lie mainly in two clusters of neurons in the midbrain of mammals: the substantia nigra pars compacta (SNpc) and the ventral tegmental area (VTA). The axons of these neurons have huge axonal arbors, each releasing dopamine at up to 10^6 sites, which is 100 to 1,000 times more than reached by typical axons. Figure 13.1 shows the axonal arbor of a single dopamine neuron whose cell body is in the SNpc of a rat's brain. A prevalent view is that dopamine neurons act in synchrony to send a common signal to wide areas of the brain. Accumulating evidence is leading to the conclusion that different subpopulations of dopamine neurons send different signals to different neural structures, but it is reasonable to think of any of these subpopulations as broadcasting a common signal to a broad neural territory.

Electrophysiological studies of the activity of dopamine neurons in anesthetized animals revealed that they respond to a variety of sensory stimuli, but it took studies of their activity in awake, behaving monkeys to show that their behavior is more complex than it would be if they were simply conveying a reward signal. The particular studies that attracted the attention of researchers familiar with reinforcement learning theory were conducted in the 1980s and early 1990s in the laboratory of neuroscientist Wolfram Schultz.

Early studies in this and other laboratories showed that dopamine neurons respond with bursts of activity to intense, novel, or unexpected visual and auditory stimuli that trigger eye and body movements. However, the neurons showed very little activity related to the movements themselves. This was surprising because degeneration of dopamine neurons is a cause of Parkinson's disease, whose symptoms include motor disorders, particularly deficits in self-initiated movement.

This surprising result motivated Romo and Schultz (1990) to more carefully investigate dopamine neuron activity to see how it is involved in the initiation of arm movements. They recorded the activity of dopamine neurons as well as muscle activity while monkeys moved their arms. They trained two monkeys to reach from a resting hand position into a bin containing a bit of apple, a piece of cookie, or a raisin when the monkey saw and heard the bin's door open. The monkey could then grab the food and bring it to its mouth. After a monkey became good at this, it was trained in two additional tasks.

The purpose of the first task was to see what dopamine neurons do when movements are self-initiated. The bin was left open but covered from above so that the monkey could not see inside but could reach in from below. No

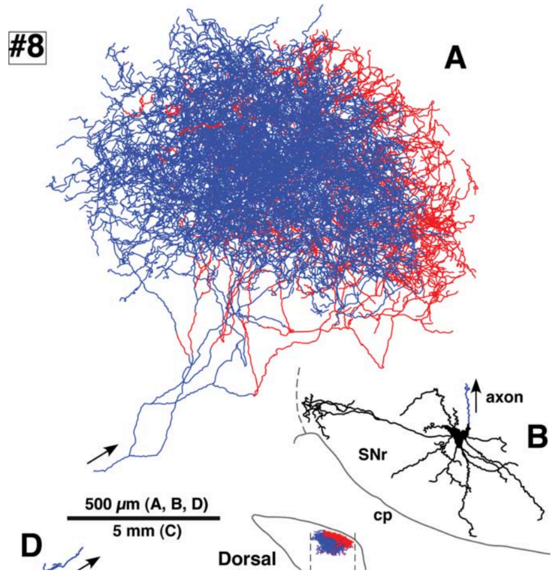

Figure 13.1: From Matsuda et al. (2009).

triggering stimuli were presented, and after the monkey reached for and ate the food morsel, the experimenter usually (though not always), silently and unseen by the monkey, replaced food in the bin by sticking it onto a rigid wire. Romo and Schultz observed that a large percentage of the dopamine neurons they monitored produced a phasic response whenever the monkey first touched a food morsel, but that these neurons did not respond when the monkey touched just the wire or explored the bin when no food was there. This was strong evidence that the neurons were responding to the food and not to other aspects of the task. In addition, the activity of the neurons was not related to the monkey's movements.

The second task's purpose was to see what happens when movements are triggered by stimuli. This task used a different bin with a moveable cover. The sight and sound of the bin opening triggered reaching movements to the bin. In this case, Romo and Schultz found that after a while the dopamine neurons did not respond to the touch of the food but instead responded to the sight and sound of the opening cover of the food bin, suggesting that the phasic responses of these neurons had shifted to stimuli predicting the availability of reward. In a followup study they found that most of the dopamine neurons whose activity they monitored did not respond to the sight and sound of the bin opening outside the context of the behavioral task, suggesting that the neurons were not responding to sensory properties of the stimuli but instead were signaling an expectation of reward (Schultz and Romo, 1990). This was the first step toward the reward prediction error hypothesis.

Schultz's group conducted many additional studies involving both SNpc and VTA neurons. In one influential study (Ljungberg, Apicella, and Schultz, 1992) monkeys were instrumentally conditioned to depress a lever after a light was illuminated to obtain a drop of apple juice. Dopamine neurons initially responded to the reward—the drop of juice—but they lost that response as conditioning continued and developed responses instead to the illumination of the light that predicted the reward. With continued training, lever pressing became faster while the responses of the dopamine neurons to the light decreased. Responses reappeared when training began for a different task.

In a more complicated task an instruction cue signaling which of two levers would be rewarding was followed a second later by a trigger cue signaling when the monkey should respond. In this task dopamine neuron activity shifted from initial responding to the reward to responding to the earlier predictive stimuli, first progressing to the trigger stimulus then to the still earlier instruction cue. As responding moved earlier in time it disappeared from the later stimuli (Figure 13.2). Here again the responses were much reduced when the task was well learned. Another finding was that during learning if a reward was not delivered at the time of its usual occurrence on correct trials, many of the dopamine neurons showed a sharp decrease in their activity below baseline shortly after the the reward's usual time of delivery, and this happened even without any external cue (Figure 13.3).

These observations led Schultz and his group to conclude that dopamine neurons respond to unpredicted rewards, to the earliest predictors of reward, and that dopamine neuron activity decreases below baseline if a reward, or a predictor of reward, does not occur at its expected time. These are the main observations that support the reward prediction error hypothesis.

In addition to these basic findings, other observations add to the plausibility of the hypothesis. When rewards of different magnitudes are delivered with different probabilities, phasic dopamine activity is consistent with how the TD error would behave in these situations, namely that dopamine activity is

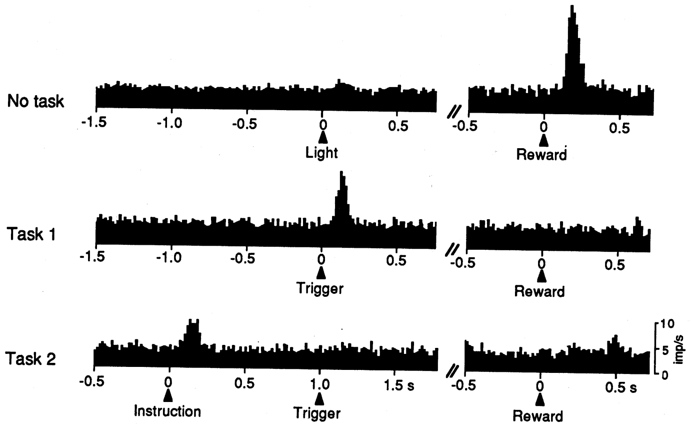

Figure 13.2: The response of dopamine neurons shifts from initial responses to primary reward to earlier predictive stimuli. From Schultz (1998).

roughly proportional to the error between the *expected reward magnitude* and the actual reward signal (Fiorillo, et al., 2003; Tobler et al., 2005). Experiments that manipulated the delay of reward showed that dopamine signaling decreases with delay as one would predict with temporal discounting (Roesch et al., 2007). Analysis of dopamine neuron activity in the blocking procedure of classical conditioning (Section **??**) demonstrated that this activity is consistent with the role of RPEs in theories of animal learning (Waelti, Dickinson, and Schultz, 2001). Another study suggests that animals' reward expectations depend on the reward history as computed by an iterative computation based on the history of prediction errors in the manner of TD learning (Bayer and Glimcher, 2005).

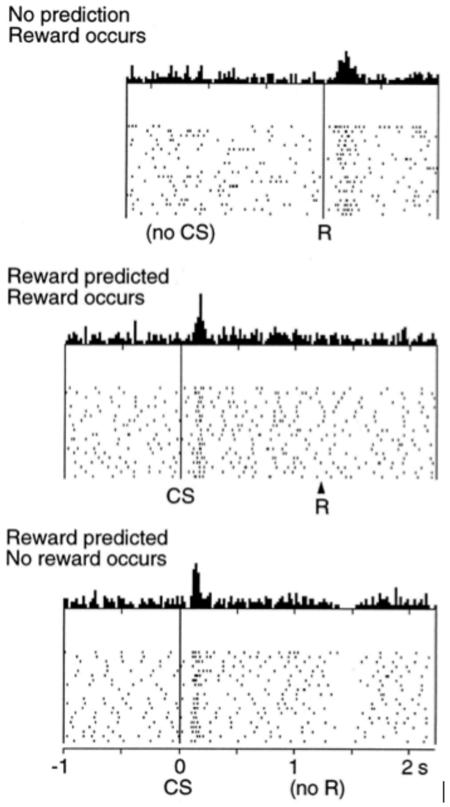

Figure 13.3: From Schultz (1998).

13.6 Dopamine and TD Learning

From all we have said about TD learning in previous chapters, it should be clear why the experimental results described above support the reward prediction error hypothesis. Nevertheless, here we take a close look because some of the important points are easy to overlook. Figure 13.4 illustrates the simplest scenario showing how the behavior of the TD error, δ, is consistent with the phasic responses of dopamine neurons in a task in which a sequence of actions has to be accomplished before a rewarding event occurs. This is an idealized version of the task described above in which a monkey rests its hand on a touch pad, then reaches for, and then grasps a morsel of apple, and finally brings it to its mouth. If we assume that the actions have already been learned and that they always produce the same results, this is an episodic prediction task.

Each episode (or trial, to use the animal-learning term) consists of a sequence of states leading up to a rewarding state. We call the states in this sequence *reward-predicting states* because in this task reward regularly follows these states and does not occur otherwise. Assume the algorithm TD(0) using a lookup table is learning a value function, V, stored in a lookup table initialized to be zero for all the states. Also assume that the discount factor, γ, is very nearly one so that we can ignore it.

The top graph in Figure 13.4 represents the sequence of states visited in each episode. Reward is zero throughout each episode except when the agent reaches the rewarding state, shown near the right end of the time line, when the reward signal becomes some positive number, say $r^\star$. Preceding the rewarding state is a sequence of reward-predicting states, with the *earliest reward-predicting state* shown near the left end of the time line. This is the state that produces the earliest cue or trigger stimulus in an episode. (Here we are assuming that states visited on preceding episodes are not counted as predicting reward on *this* episode because the inter-episode-interval is so long.) Other, later, reward-predicting states leading up to the reward are necessary so that TD learning can back up values along the trial's time line. These states need not produce external stimuli, but could be like the internal microstimuli forming a CS representation for the TD model of classical conditioning as illustrated in Figure **??**. The *latest reward-predicting state* in an episode is the state immediately preceding the episode's rewarding state. This is the state near the far right end of the time line in Figure 13.4. The rewarding state of an episode does not predict that episode's reward: the value of this state would come to predict reward for the *following* episodes, which for simplicity we are assuming to be zero because of a long time between episodes.

Figure 13.4 shows the first-episode time courses of V and δ as the graphs labeled "early in learning." Because reward is zero throughout the episode except when the rewarding state is reached, and all the V-values are zero, the TD error is also zero until it becomes $r^\star$ at the rewarding state. This follows from the definition $\delta_t = r_t + V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}) = r_t + 0 - 0 = r_t$, which is zero until it equals $r^\star$ when the reward occurs. Here $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are respectively vectors of features representing the states visited at times t an $t-1$ in a trial. The TD error at this stage of learning is analogous to a dopamine neuron responding to an unpredicted reward, e.g., a morsel of apple, at the start of learning.

Throughout this first episode and all successive episodes, TD(0) backups occur at each state transition as described in Chapter **??**. This successively increases the values of the reward-predicting states, with the increases spreading backwards from the rewarding state, until the values converge to the correct return predictions. In this case (since we are assuming no discounting) the correct predictions are equal to $r^\star$ for all the reward-predicting states. This can be seen in Figure 13.4 as the graph of V labeled "learning complete." The values of the states preceding the earliest reward-predicting state remain low (which Figure 13.4 shows as zero) because they are not reliable predictors of reward. (Below we explain why the TD algorithm, which backs values up to these states as well, leaves their values low.)

When learning is complete, that is, when V attains its correct values, the TD errors associated with transitions *from* any reward-predicting state are zero because the predictions are now accurate. This is because for a transition from a reward-predicting state to another reward-predicting state, we have $\delta_t = r_t + V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}) = 0 + r^\star - r^\star = 0$, and for the transition from the latest reward-predicting state, we have $\delta_t = r_t + V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}) = r^\star + 0 - r^\star - 0 = 0$. However, the TD error on a transition from any state *to* the earliest reward-predicting state is positive because of the mismatch between this state's low value and the larger value of the following reward-predicting state. Indeed, if the value of a state preceding the earliest reward-predicting state is zero, then after the transition to the earliest reward-predicting state $\delta_t = r_t + V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}) = 0 + r^\star - 0 = r^\star$. The "learning complete" graph of δ in Figure 13.4 shows this positive value at the earliest reward-predicting state, and zeros everywhere else.

This non-zero TD error upon transitioning to the earliest reward-predicting state is analogous to the persistence of dopamine responses to the earliest stimuli predicting reward. By the same token, when learning is complete a transition from the latest reward-predicting state to the reward state produces a zero TD error because the latest reward-predicting state's value, being correct, cancels the reward. This parallels the observation that fully predicted reward does not generate a burst of dopamine neuron activity.

If after learning, the reward is suddenly omitted, there is a negative TD error at the usual time of reward because the value of the latest reward-predicting state is then too high: $\delta_t = r_t + V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}) = 0 + 0 - r^\star = -r^\star$, as shown at the right end of the "r omitted" graph of δ in Figure 13.4. This is like dopamine neuron activity decreasing below baseline upon the omission of an expected reward.

This analysis raises the question of what exactly is an *earliest reward-predicting state*? In an animal's life, many different states may be followed by an earliest reward-predicting state. However, because these states are more often followed by *other* states that do not predict reward, their reward-predicting power, that is, their values, remain low. The TD algorithm, operating throughout the animal's life, backs up values to these states too, but the backups do not consistently accumulate because, by assumption, none of these states reliably precedes an earliest reward-predicting state. If any of them did, they would

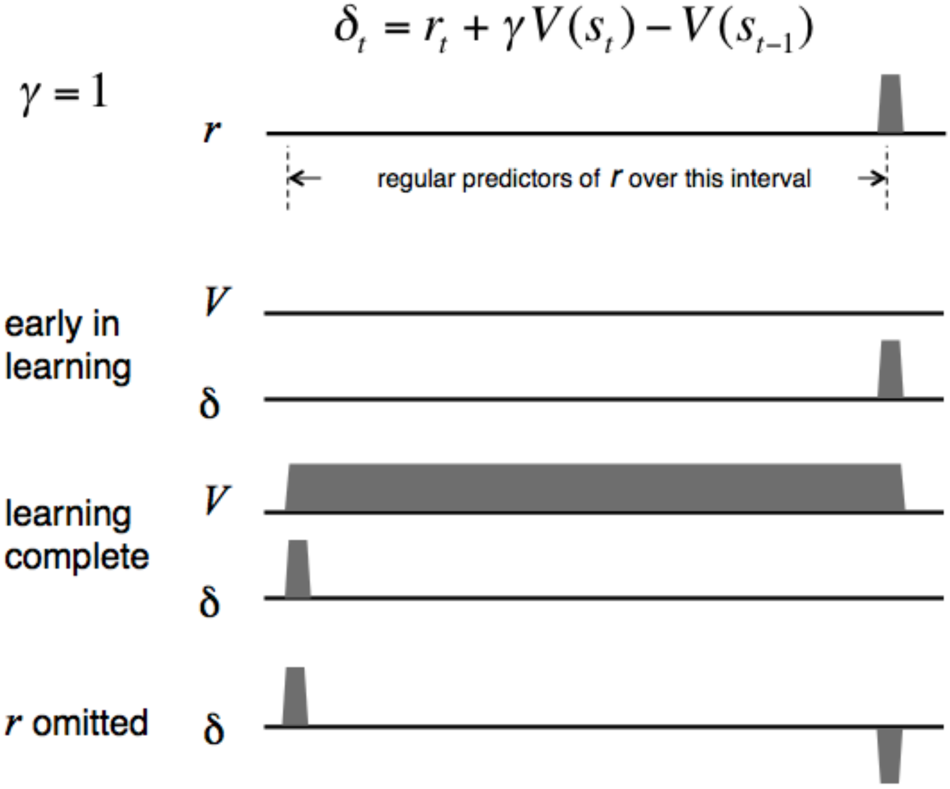
$\delta_t = r_t + \gamma V(s_t) - V(s_{t-1})$
$\gamma = 1$
r
regular predictors of r over this interval
early in learning
V
δ
learning complete
V
δ
r omitted
δ

Figure 13.4:

be reward-predicting states as well. Thus we can think of an earliest reward-predicting state as an *unpredicted predictor* of reward—and clearly there can be many such states.

This reasoning provides an explanation for the observation that with overtraining, dopamine responses decrease to even the earliest reward-predicting stimulus in a trial. With overtraining one would expect that even a formerly unpredicted predictor state would become predicted by stimuli associated with earlier states: the animal's life both inside and outside of an experimental task would become commonplace. Upon breaking this routine with the introduction of a new task, however, one would see TD errors reappear, as indeed one observes in dopamine neuron activity.

If the reward prediction error hypothesis is correct—even if it accounts for only a portion of a dopamine neuron's activity—then the phasic responses of dopamine neurons are neither reward signals nor value signals. A plausible alternative is that they are *reinforcement signals*, meaning that they modulate learning by providing information a learning algorithm needs to update a set of parameters. Many of the algorithms we have discussed in this book suggest ways to think about reinforcement signals in the brain, including algorithms for learning action-values, such as Q-learning and Sarsa, and algorithms that learn environment models. Some of the most influential ideas are related to the Actor-Critic class of reinforcement learning algorithms in which the Actor learns policies, the Critic learns state values, and TD errors are reinforcement signals for both. The next section describes a hypothesis based on an Actor-Critic algorithm about how the brain generates and uses dopamine's encoding of TD errors.

13.7 Actor-Critics and the Brain

Neuroscience is sill far from achieving complete understanding of how the brain generates and uses dopamine signals, but evidence from behavioral and physiological experiments, along with brain anatomy, suggests that something like an Actor-Critic algorithm may be responsible for many—though certainly not all—features of dopamine signaling and its function. As is true for any theoretically-motivated neural hypothesis, this hypothesis neglects innumerable neural details and is certainly incorrect in many ways. Nevertheless, the basic Actor-Critic algorithm has helped organize thinking about the role of dopamine in the brain's habit-learning system.

Actor-Critic algorithms (Section **??**) use explicit representations of both a policy and a state-value function. An architecture implementing this type of algorithm consists of two main components, the Actor and the Critic, which respectively update a policy and a state-value function. The Actor updates the current policy using the TD error computed by the Critic as the reinforcement signal. At any given time, the Actor outputs the action specified by the current policy (or an exploratory action), and the Critic outputs the current TD error. Figure 13.5 (a) is a diagram of the basic Actor-Critic architecture. In its simplest form, the Critic implements a TD algorithm to estimate state-values under the current policy, and it combines temporal changes in these estimated values with reward information to form a TD error δ. The Critic's algorithm is basically the TD model of classical conditioning described in Section **??** except that here the relevant output is δ instead of state-values V that contribute to conditioned responding in that model. Some versions of the architecture place the computation of δ outside of both the Critic and the Actor.

A distinctive feature of the Actor-Critic architecture is that the TD error, δ, produced by the Critic is the reinforcement signal for both the Actor and the Critic, but it modulates learning in different ways in each of these components. For the Actor, δ tells how to update the action probabilities in order to reach higher-valued states. The Actor increases the probability of any action leading to a state with a higher-than-expected value (a positive δ), and it decreases the probability of any action followed by a state with a less-than-expected value (a negative δ). On the other hand, δ tells the Critic how to change the value function parameters in order to improve predictive accuracy. This makes δ the appropriate reinforcement signal for the Critic. Because it uses an error-correcting algorithm, the Critic works to reduce δ to as close to zero as possible.

The computational rational for using δ as the Actor's reinforcement signal is that it is a better reinforcement signal than the "naked" reward signal R. The reward signal R might work as a reinforcement signal, but δ is better for a number of reasons. First, because it responds to the earliest reliable predictors of reward, it helps address the problem of delayed reward by being precisely timed to provide immediate evaluation of actions in terms of their expected consequences for bringing about future reward. Second, a given level of reward is not good or bad in itself, but only in comparison with expected levels of reward that might have been received had the system behaved differently. This is addressed by δ because it compares current reward with a reward expectation based on past experiences. A third feature of δ that makes it a better reinforcement signal than R is that it averages out reward variations due to uncontrollable and random aspects of the environment so that learning is less affected by these transitory variations.

Figure 13.5 (b) illustrates a hypothesis for how an Actor-Critic architecture might be implemented in the brain. This hypothesis is relevant only to the brain's habit system because the Actor-Critic architecture implements a model-free algorithm (see Section ??). The neuroscience of goal-directed behavior, likely relying on model-based methods, is discussed in Section 13.11.

The major brain structures involved comprise the basal ganglia, a collection neuron groups, or nuclei, lying at the base of the forebrain that participate

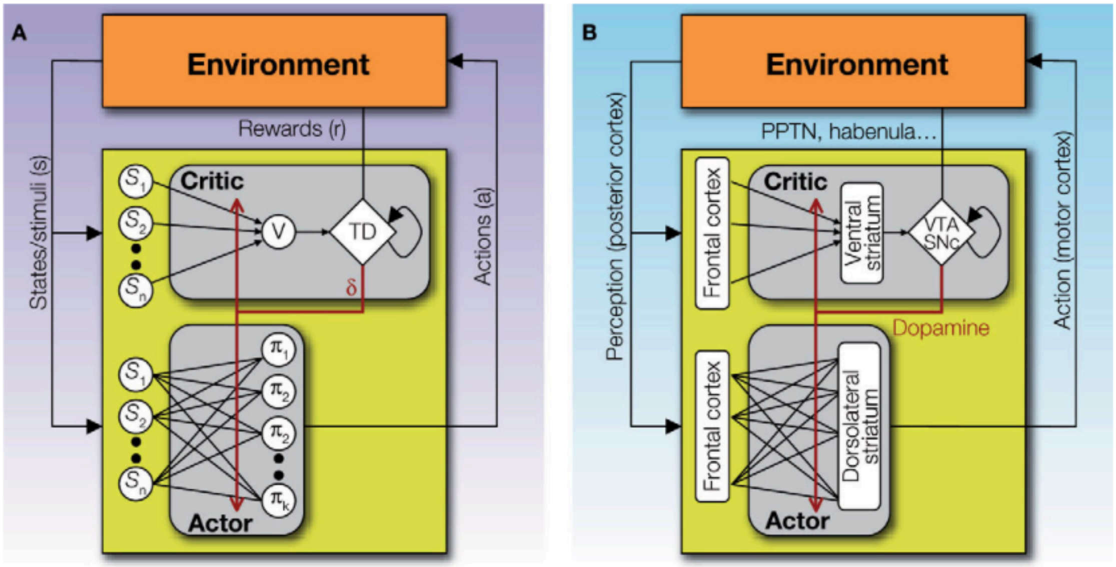

Figure 13.5: Actor-Critic architecture and a hypothetical neural implementation. (a) The Actor adjusts a policy based on the TD error δ it receives from the Critic, while the critic adjusts state-value parameters using the same TD error. The Critic produces a TD error from the reward signal, R, and the current change in its estimate state values. The Actor does not have direct access to the reward signal, and the Critic does not have direct access to the action. (b) In this hypothetical neural implementation the Actor and value learning part of the Critic are respectively placed in the ventral and dorsal subdivisions of the striatum. The TD error is transmitted from the VTA and SNpc to modulates changes in synaptic efficacies of input from cortical areas to the ventral and dorsal striatum. From Takahashi et al. (2008) permission pending.

in organizing behavior and cognitive function. The hypothesis illustrated in Figure 13.5 (b) places the Actor and the value-learning part of the Critic respectively in the dorsal and ventral subdivisions of the input structure of the basal ganglia known as the striatum. The ventral striatum sends value information to the VTA and SNpc, where dopamine neurons in these nuclei combine it with information about rewards to generate TD error signals. One idea for how this happens is that value information arrives to dopamine neurons via an excitatory direct pathway and a slower indirect inhibitory pathway to yield the temporal value difference that is combined with signals conveying information about reward. The axons of these dopamine neurons, in turn, project to the dorsal and ventral striatum, providing the reinforcement signal for adjusting the policy and the value function, respectively, through changes in the efficacies of the corticostriatal synapses, that is, the synapses by which cortical areas activate striatal neurons.

A detailed look at the rationale for this hypothetical neural implementation is beyond the scope of this book (see the references in the Historical and Bibliographic Remarks section at the end of this chapter), but here is the broad outline. Essentially all of the cerebral cortex, along with other structures, sends signals to the striatum. These signals convey a wealth of information about sensory input, internal states, as well as information about motor activity. Output from the striatum loops back via other basal ganglia nuclei and the thalamus to frontal areas of cortex as well as to motor areas making it possible for the striatum to influence movement, more abstract decision processes, and reward processing. The dorsal striatum is primarily implicated in influencing action selection, and the ventral striatum is thought to be critical for different aspects of reward processing, including the assignment of affective value to sensations. Thus, the anatomy and putative functions of these regions are in line with the idea that the dorsal and ventral striatum respectively implement Actor-like and Critic-like mechanisms.

A notable implication of the hypothesis of Figure 13.5 (b) is that the dopamine signal is a reinforcement signal, not a reward signal like the scalar signal R of reinforcement learning theory. In fact, the hypothesis implies that one should not necessarily be able to probe the brain and record such a “master” scalar reward signal in the activity of any single neuron. Many interconnected neural systems generate reward-related information, with different structures being recruited depending on different types of rewards. Dopamine neurons receive information from many different brain areas, so the input to the SNpc and VTA labeled “Reward” in Figure 13.5 (b) should be thought of as vector of reward-related information arriving to these neurons along multiple input channels. Some of these channels come from inside the animal’s body, conveying information about its state and needs, and some come from outside, conveying information about the animal’s environment related to food, social interaction, sex, danger, and other features essential for reproductive success. Reward-related information also includes less direct indicators that evolution has determined are also relevant to reproductive success, such as progress in learning about and exerting control over the environment. What the theoretical scalar reward signal R corresponds to, then, is the net contribution of all reward-related information to dopamine neuron activity. It is likely not a signal transmitted along any single axon in the brain but rather the result of a pattern of activity across many neurons in different areas of the brain.

The results of many experiments show that dopamine is critical for learning, and that it in fact conveys a reinforcement signal. One can go back to Olds' and Milner's (1954) famous observations of the rewarding effect of electrical stimulation of particular sites in a rat's brain. Later findings that these sites excited dopamine pathways led to the view that dopamine conveys a reward signal. A closer look at this paper shows that it describes the reinforcing effects of electrical stimulation in an instrumental conditioning task. Electrical stimulation not only energized the rats' behavior—through dopamine's effect on motivation—it also led to the rats quickly learning to stimulate themselves by pressing a lever, which they would do frequently for long periods of time. Dopamine signaling triggered by electrical stimulation acted as a reinforcement signal for learning.

Other experiments have shown that dopamine is critical for both classical and instrumental conditioning. Inactivating dopamine neurons and blocking their effect on target sites disrupts learning, as does genetic manipulation that effects dopamine neuron activity. Especially convincing support comes from the use of optogenetic methods, which allow neuroscientists to precisely control the activity of selected neuron types at a millisecond timescale in awake behaving animals. Optogenetic methods introduce light-sensitive proteins into selected neuron types so that these neurons can be activated or silenced by means of flashes of light. In the first study using optogenetic methods to study dopamine neurons, Tsai et al. (2009) showed that optogenetic stimulation producing phasic activation of dopamine neurons in mice was enough to condition them to prefer the side of a chamber where they received this stimulation as compared to the chamber's other side where they received no stimulation or lower-frequency stimulation. In another set of experiments using optogenetic activation of dopamine neurons, Steinberg et al. (2013) created artificial bursts of dopamine neuron activity in rats at the times when rewarding stimuli were expected but omitted—times when dopamine neuron activity normally pauses. With these pauses replaced by artificial bursts, responding was sustained when it would ordinarily decrease due to lack of reinforcement (in extinction trials), and learning was enabled when it would ordinarily be blocked due to the reward being already predicted (the blocking paradigm; see Chapter **??**). These results show that phasic dopamine neuron activity plays a causal role in behavioral conditioning.

What is known about how dopamine is distributed throughout the brain and how it influences target sites makes it plausible that dopamine signals from the VTA and SNpc function as reinforcement signals in the general manner suggested by Figure 13.5 (b). Although it is an oversimplification that all dopamine neurons send the same signal, the widely-branching axons of dopamine neurons are well-suited for rapidly sending a precisely-timed common reinforcement signal to the multiple sites where it modulates synaptic plasticity. Most of the neurons in the striatum are medium spiny neurons, so called because their dendrites are covered with spines on whose tips the inputs from the cortex make synaptic contacts. Medium spiny neurons are the main input/output neurons in the striatum. Dopamine neuron axons form synapses on the spine stems of medium medium spiny neurons (Figure 13.6). This brings pre- and postsynaptic fibers together with dopamine input. Neuroscientists have hypothesized that this is an ideal arrangement for a learning rule for adjusting the efficacies of corticostriatal synapses on the basis of three factors: presynaptic activity of cortical input fibers, activity of the postsynaptic medium spiny neurons, and modulation by the dopamine signal. In the next section we look at what the Actor and Critic learning rules suggest about how learning might depend on these factors.

Phasic bursts of dopamine neuron activity are particularly suitable for modulating changes at these synapses because they elevate extracellular dopamine concentration in the striatum more than slower firing does, thus producing a signal of significant strength. Further, timing precision is aided by the fact that dopamine released from presynaptic sites on the axons of dopamine neurons is rapidly reabsorbed into these axons so that the extracellular concentration of dopamine shows only a short-duration increase in concentration, roughly lasting a half second, as a result of a phasic burst. The credit-assignment utility of these mechanisms is supported by the effect of dopamine reuptake inhibitors, such as cocaine, that interfere with this reabsorbtion process, thereby leading to increased extracellular dopamine concentrations thought to disrupt the brain's credit assignment mechanism.

If the brain does implement something like the Actor-Critic architecture, and assuming populations of dopamine neurons broadcast a common reinforcement signal to the corticostriatal synapses of both the dorsal and ventral striatum as illustrated in Figure 13.5 (b) (which is an oversimplification as we mentioned above), then this signal affects the synapses of these two structures in different ways. The learning rules of the Critic and the Actor use the same reinforcement signal, but the signal's effect on learning is different for these two components. The difference in the Critic and Actor learning rules is relatively simple, but it has a profound effect on learning and is essential to how the Actor-Critic

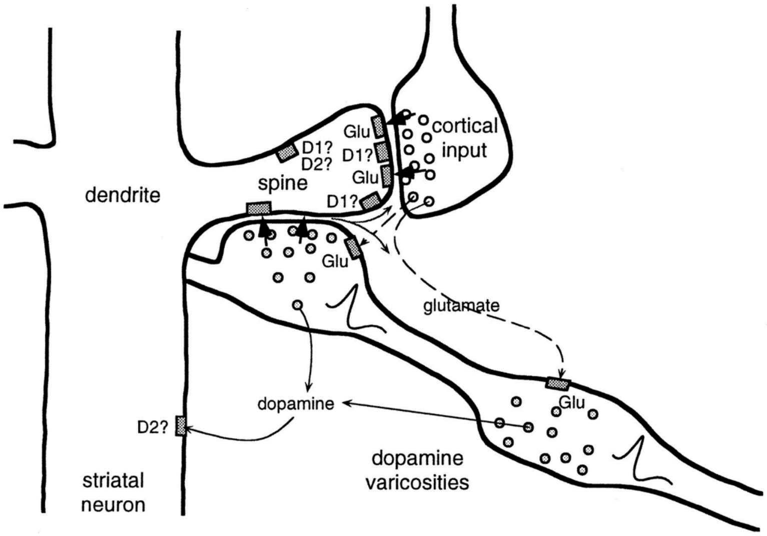

Figure 13.6: Spine of a striatal medium spiny neuron showing input from both cortical and dopamine neurons. Axons of cortical neurons influence striatal medium spiny neurons via synapses on the tips of spines covering the dendrites of these neurons. Each axon of a VTA or SNpc dopamine neuron makes synaptic contact with the stems of roughly 500,000 spines that it passes by, where dopamine is released from "dopamine varicosities." At each synapse by which the cortex influences the striatum, this arrangement brings together presynaptic input from cortex, postsynaptic activity of the medium spiny neuron, and a dopamine signal. This arrangement makes it possible that a either a two- or three-factor learning rule governs the plasticity of corticostriatal synapses. Hypothetically, a two-factor rule for Crtiic-like learning would use non-contingent eligibility traces not involving the postsynaptic activity, with dopamine providing the reinforcement signal, whereas a three-factor rule for Actor-like learning would use contingent eligibility involving pre- and postsynaptic activity, also with dopamine providing the reinforcement signal. What actually occurs at these spines is complex and not completely understood. The figure hints at the complexity that is possible by showing two types of receptors for dopamine, receptors for glutamate, the neurotransmitter of the cortical inputs, and multiple ways that the various signals can interact. From Schultz (1998) permission pending.

architecture functions. The major difference is in the kind of eligibility traces each type of learning rule uses, the topic to which we turn next.

13.8 Eligibility Traces

Eligibility traces are essential features of both the prediction and control algorithms presented in preceding chapters. They make predictive associations possible by linking events back to earlier-occurring states or state-action pairs. They enable control policies to be improved by linking states to actions on the basis of the later-occurring consequences of those actions. The eligibility traces in these algorithms are derived from Klopf's hypothesis of the "hedonistic neuron" (Klopf 1972, 1982). According to this hypothesis, eligibility traces are properties of synapses. When certain conditions (specified below) are satisfied, a synapse becomes eligible for modification and remains eligible for a limited period of time, but modification only occurs if certain other conditions (also specified below) are met during the period of eligibility. Klopf thought of eligibility as a synaptically-local molecular mechanism different from the electrical activity of both the presynaptic and postsynaptic neurons. We discuss Klopf's hypothesis in more below.

Section **??** related the idea of eligibility is related to similar ideas in animal learning theories, where stimulus traces have been proposed to bridge temporal intervals between stimuli. Here we look at eligibility traces from a neuroscience perspective, specifically focusing on the critical role eligibility traces play in the hypothetical neural implementation of an Actor-Critic architecture shown in Figure 13.5 (b). In the rat's brain the striatal subdivisions hypothesized to implement the Actor and the Critic each contain millions of medium spiny neurons. According to the hypothesis, Actor neurons learn policies and Critic neurons learn values, but they do this while using a common reinforcement signal to modulate changes in their synaptic efficacies. This happens because the eligibility traces for neurons in these two regions are triggered by different conditions. The mathematical definitions for eligibility traces and learning rules given below are based on computational principles rather than on neuroscience data, but these definitions allow us to be specific about issues that are likely also to be relevant for neural systems. Here we focus on a single medium spiny neuron in each striatal subdivision, reserving discussion of learning by the entire neuron populations for Section **??**.

An eligibility trace for a synapse providing input to an Actor neuron is triggered whenever activity of the presynaptic neuron takes part in causing the postsynaptic neuron to fire. This is related to Hebb's classic proposal that whenever a presynaptic signal participates in activating the postsynaptic neuron the synapse's efficacy increases (Hebb, 1949). Unlike Hebb's proposal, however, here whenever activity of the presynaptic neuron takes part in causing the postsynaptic neuron to fire, the synapse only becomes eligible for modification; how its efficacy changes depends on the reinforcement signal received during the window when its eligibility trace is non-zero. We call this a *contingent eligi-*

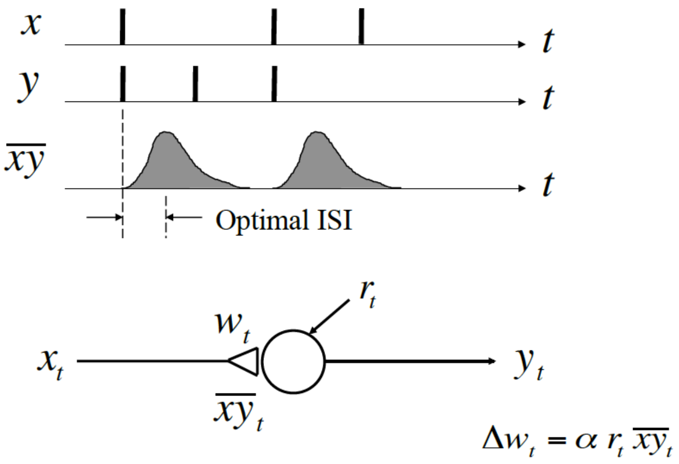

Figure 13.7:

bility trace: it is contingent on the firing of the postsynaptic neuron In contrast, the eligibility traces of a Critic neuron are triggered only by presynaptic activity. Activity of the postsynaptic neuron plays no role in the initiating synaptic eligibility. We call this a *non-contingent eligibility trace*. Contingent eligibility traces are related to instrumental conditioning, while non-contingent eligibility traces are related to classical conditioning (Chapter **??**).

In the hypothetical neural implementation of Figure 13.5 (b), a policy is stored in the efficacies of the synapses on the dendrites of medium spiny neurons in the dorsal striatum, the structure hypothesized to implement the Actor. The Actor's eligibility traces are part of the spine mechanisms at these synapses. If $\mathbf{x}_t$ denotes the input from the cortex to one of these neurons at time t, and y_t denotes the output of this neuron at time t, then the vector of the eligibility traces in this neuron's spines at time t, denoted $\mathbf{e}_t^A$ (where the superscript A identifies it as eligibility for the Actor), is updated according to the following equation:

$$\mathbf{e}_{t+1}^A = \gamma\lambda\mathbf{e}_t^A + y_t\mathbf{x}_t, \tag{13.1}$$

for discount factor γ and eligibility decay parameter λ. Here, as in many models of Hebbian-style synaptic plasticity, we assume that input from a presynaptic neuron influences the activity of the postsynaptic neuron so quickly that we can ignore any delay and represent the contingency condition as the product of simultaneous pre- and postsynaptic activities: $y_t\mathbf{x}_t$.

The Actor learns by changing the efficacies of the synapses by which it receives input from the cortex. Letting $\mathbf{w}_t$ denote the vector of efficacies at time t of these synapses, the Actor's learning rule is the following:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t\mathbf{e}_t^A, \tag{13.2}$$

for step-size parameter α, and where δ_t is the TD-error reinforcement signal at time t supplied by a burst of dopamine neuron activity as shown in Figure 13.6. This is a *three-factor learning rule* because it depends on presynaptic and postsynaptic activity—through their influence on contingent eligibility—and the reinforcement signal.

As in all reinforcement learning systems, the Actor-Critic architecture has to produce exploratory actions. In computer implementations of the architecture this is typically done simply by adding a random component to the Actor's behavior. The could occur in the brain simply by means of noise in the activity of Actor-like neurons. Medium spiny neurons the dorsal striatum in fact exhibit irregular spontaneous activity that could be a source of exploratory behavior, although more sophisticated forms of exploration are certainly possible as well.

The Critic's learning rule is basically the TD model of classical conditioning described in Section **??**. Sticking to the hypothetical neural implementation, let $\mathbf{e}_t^C$ denote the vector of eligibility traces at time t associated with the spines of a medium spiny neuron in the ventral striatum, the structure hypothesized to implement the value-learning part of the Critic. These traces are updated according to this equation:

$$\mathbf{e}_{t+1}^C = \gamma\lambda\mathbf{e}_t^C + \mathbf{x}_t, \tag{13.3}$$

where γ is the discount factor and λ is the eligibility trace-decay parameter. This differs from the Actor's eligibility update (Equation 13.1) only in having the term $\mathbf{x}_t$ on the far right instead of $y_t\mathbf{x}_t$. This means that the Critic's eligibilities are traces of presynaptic activity only—the neuron's output is not involved.

The Critic synaptic efficacies, $\mathbf{v}$, update according to Equation **??**, which we write again here:

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \alpha\delta_t \mathbf{e}_t^C, \tag{13.4}$$

for step-size parameter α, and where δ_t is the TD-error reinforcement signal at time t supplied by a dopamine signal as shown in Figure 13.6. This is a *two-factor learning rule* depending on presynaptic activity—through its influence on non-contingent eligibility—and the reinforcement signal.

Evidence is accumulating that synaptic eligibility traces in fact exist. Neuroscientists are intensely studying a form of synaptic plasticity called *spike-timing-dependent plasticity* (STDP) that involves eligibility-like traces. Experiments have shown that changes in many synapses depend on the relative timing of presynaptic and postsynaptic action potentials, i.e., spikes. The dependence can take different forms, but in the one most studied a synapse increases in strength if spikes incoming via that synapse arrive shortly before the postsynaptic neuron fires. If the timing relation is reversed, with a presynaptic spike arriving shortly after the postsynaptic neuron fires, then the strength of the synapse decreases. This form of STDP can be accounted for by assuming the existence of exponentially-decaying traces, one triggered by each presynaptic spike, and another one triggered by each postsynaptic spike. These are something like eligibility traces, although the their time courses are shorter than the time courses of the eligibility traces typically used in reinforcement learning algorithms.

The discovery of STDP has led neuroscientists to investigate the possibility of a three-factor form of STDP that depends on a neuromodulatory input in addition to pre- and postsynaptic activity. This form of synaptic plasticity, called *reward-modulated STDP*, is much like the Actor learning rule discussed above. Evidence is accumulating that this type of plasticity occurs at the spines of medium spiny neurons of the dorsal striatum, with dopamine providing the neuromodulatory factor. These are the sites where Actor learning takes place in the hypothetical neural implementation of the Actor-Critic architecture we have been discussing. Other experiments have demonstrated a form of reward-modulated STDP in which lasting changes in the efficacies of cortical synapses occur only if a neruromodulator pulse arrives within a time window that can last up to 10 seconds after a presynaptic spike is closely followed by a postsynaptic spike (He et al., 2015). (In these experiments neuromodulators other than dopamine produce the effect, so it may be inappropriate to call all forms of this type of plasticity *reward*-modulated STDP). Although the evidence is indirect, these experiments demonstrate the existence contingent eligibility traces with time courses like those of the eligibility traces Klopf postulated. The molecular mechanisms producing these eligibility traces, as well as those underlying

STDP, are not yet understood, but research focusing on time-dependent and neuromodulator-dependent forms of synaptic plasticity is continuing.

The hypothesis that Actor-like learning takes place at the level of single neurons derives from Klopf's "hedonistic neuron" proposal (Klopf 1972, 1982), which has had substantial influence on our computational approach to reinforcement learning. In the next section we look more closely at Klopf's proposal, which—among other things—explains both the philosophical and computational rationale for eligibility traces and Actor-type learning.

13.9 Hedonistic Neurons

Klopf's hedonistic neuron hypothesis is that individual neurons seek to obtain reward and avoid punishment, where rewards and punishments are conveyed through synaptic input from other neurons. Neurons do this, he conjectured, by implementing a form of the Law of Effect (see Section **??**) through a three-factor synaptic learning rule relying on contingent eligibility traces and a specific definition of rewards and punishments. Klopf argued that instead of the assumption common to traditional theories that homeostasis is the primary goal of behavior and learning, homeostasis is a subgoal and that organisms' primary goal is to increase the difference between the amounts of reward and punishment they receive. He conjectured that this organism-level goal arises from an intrinsic property of neurons by which they attempt to increase the difference between the amounts of reward and punishment they themselves receive. According to his hypothesis, then, individual neurons are self-interested hedonistic agents. In the Bibliographical and Historical Remarks section at the end of this chapter we mention similar ideas that have been proposed by others.

According to Klopf's hypothesis when a neuron fires an action potential, all of its synapses that were active in contributing to that action potential become eligible to undergo changes in their efficacies. If the action potential is followed within an appropriate time period by an increase of reward, the efficacies of all the eligible synapses increase. Symmetrically, if the action potential is followed within an appropriate time period by an increase of punishment, the efficacies of eligible synapses decrease. This is implemented by triggering an eligibility trace upon a coincidence of presynaptic and postsynaptic activity (or more generally upon pairing of presynaptic activity with the postsynaptic activity it influences). This is the inspiration for the Actor's learning rule described in the previous section and its use of contingent eligibility traces.

The shape and time course of an eligibility trace in Klopf's theory reflects the durations of the many feedback loops in which the neuron is imbedded, some of which lie entirely within the brain and body of the organism, while others extend out through the organism's external environment as mediated by its motor and sensory systems. His idea was that the shape of a synaptic eligibility trace is essentially a histogram of the durations of the feedback loops in which the neuron is embedded. The peak of an eligibility trace would then occur at the duration of the most prevalent feedback loops for that neuron, which Klopf identified with the optimal inter-stimulus-interval for conditioning, about 0.5 seconds for many paradigms. (Figure **??**).

The eligibility traces used by algorithms described in this book are simplified versions of Klopf's original idea of synaptic eligibility traces. Instead of taking the form of curves that reflect a distribution of feedback delay times, they are simply exponentially (or geometrically) decreasing functions controlled by the parameters λ and γ. This simplifies simulations as well as theory, but we regard this simple type of eligibility trace as a "place holder" for traces closer to Klopf's original conception, which may have computational advantages in complex reinforcement learning systems by refining the credit-assignment process.

Many of Klopf's ideas inspired, and are preserved in, our approach to reinforcement learning. This is especially true for the Actor-Critic architecture, where a hypothetical neural implementation of the Actor consists of neurons that are like his hedonistic neurons with δ acting as their common reward signal. The major exception is that an Actor neuron departs from Klopf's idea for how reward and punishment is conveyed to neurons. His idea was not that rewards and punishments are conveyed to neurons via neuromodulatory input specialized for this purpose. Instead, he wanted ordinary neural signals—action potentials of other neurons—to convey rewards and punishments because he believed it was essential to avoid a centralized source of these signals. He therefore defined neuron-local reward and punishment to be respectively the electrical excitation and inhibition a neuron experiences via input from other neurons. But a combination of the logic of reinforcement learning (How can it work if there is no definite objective function?), the implausibility that neurons work to maximize excitation and minimize inhibition (How can neurons function if they always try to be maximally excited?), as well as the now widely-accepted existence of fast-acting and precisely-timed modulatory signals in the brain, argue that a specialized reinforcement signal is both more plausible neuroscientifically and leads to a more manageable theoretical framework.

By itself, however, relying on a centralized source for rewards and punishments discards an important aspect of Klopf's intuition. He believed that it would merely be begging the question of how to create intelligent behavior by reducing the problem to the equally-difficult problem of designing this central source of rewards and punishments. In his scheme, neurons reward and punish each other as self-interested participants in an immense society or economic system making up the organism's nervous system. This is a thought-provoking idea, and some researchers have explored computational architectures based on similar ideas. In our approach, the Critic, and value-function methods in general, go part way toward letting ordinary signals convey reward and punishment. These methods synthesize out of all state information, a reinforcement signal—a TD error like δ—that is more informative than the "naked" reward signal R. However, while value-function methods address the temporal aspect of the credit-assignment problem (Section **??**) through prediction, they do not address *structural* aspects of the credit-assignment problem, which concern the problem of assigning credit or blame to particular components of a complex mechanism that produces behavior. A more distributed approach like the one Klopf envisioned may be an effective way to address the structural aspect of the credit-assignment problem.

The idea that a single neuron seeks reward and avoids punishment is not as outlandish as it may at first appear. A well-studied example of a single cell capable behaving something like this is the bacterium *Escherichia coli*. The movement of this single-cell organism is influenced by chemical stimuli in its environment, behavior known as chemotaxis. It swims in its liquid environment by rotating hairlike structures called flagella attached to its surface. (Yes, it rotates them!) Molecules in the bacterium's environment bind to receptors on its surface. Binding events modulate the frequency with which the bacterium reverses flagellar rotation. Each reversal causes the bacterium to tumble in place and then head off in a random new direction. A bit of chemical memory and computation causes the frequency of flagellar reversal to decrease when the bacterium swims toward higher concentrations of molecules it needs to survive (attractants) and increase when the bacterium swims toward higher concentrations of molecules that are harmful (repellants). The result is that the bacterium tends to persist in swimming up attractant gradients and tends to avoid swimming up repellant gradients.

This chemotactic behavior is called klinokinesis. It is a kind of trial-and-error behavior, although learning is unlikely to be involved: the bacterium needs a bit of short-term memory to detect molecular concentration gradients, but it probably does not maintain long-term memories. Artificial intelligence pioneer Oliver Selfridge called this strategy "run and twiddle," pointing out its utility as a basic adaptive strategy: "keep going in the same way if things are getting better, and otherwise move around" (Selfridge, 1978, 1984). Similarly, one might think of a neuron "swimming" (not literally of course) in a medium composed of the complex collection of feedback loops in which it is embedded, acting to obtain one type of input signal and to avoid others. According to this view, to fully understand a neuron's behavior it is necessary to take into account the closed-loop nature of its interaction with its environment. This implies that a neuron, like the bacterium, is actually a *control system*. Although control is usually thought of as controlling something external to the controller, what control engineers traditionally call the "plant," an equivalent—and perhaps more illuminating—view is that a controller is actually controlling features of its input (Powers, 1973).

13.10 Collective Behavior

The behavior of populations of reinforcement learning agents is deeply relevant to the study of social and economic systems—and if anything like Klopf's hedonistic neuron hypothesis is true—to neuroscience as well. The field of *multi-agent reinforcement learning* considers many aspects of the collective behavior of populations of reinforcement learning agents by extending the theoretical approach covered in this book. Although beyond the scope of this introductory book, familiarity with some basic concepts and results from the multi-agent case is important for thinking about the role that diffuse neuromodulatory systems may play in reinforcement learning in the brain.

The description above about how an Actor-Critic architecture might be implemented in the brain focussed on the learning rules of single medium spiny neurons in each of the Actor and Critic components, placed respectively in the dorsal and ventral subdivisions of the striatum. According to this hypothesis, changes in the efficacies of these neurons' synapses are modulated by a common reinforcement signal transmitted by dopamine released at corticostriatal synapses. But these striatal subdivisions contain millions of medium spiny neurons. The dorsal striatum in one hemisphere of rat's brain, for example, contains almost three million medium spiny neurons. It oversimplifies the situation to assume that exactly the same dopamine signal reaches all the synapses of all the neurons in both striatal subdivisions at exactly the same time, but it is instructive to consider this case as a starting point. Can reinforcement learning theory tell us anything about what happens when all members of a population of agents learn according to a common reinforcement signal?

The learning situation across a population of Critic neurons is relatively simple because they are all learning to predict the same quantity: the expected return, that is, the reward expected over the future. According to the Critic learning rule with its non-contingent eligibility traces, the efficacies of the synapses targeted by δ change in directions aimed at moving δ toward zero. How successful each member of the population can be in learning the expected return depends on the information it receives, which will differ across the population. But they are unified in aiming to implement the same input/output function, namely, the value function for whatever policy is driving behavior. (In the brain much more than this is likely going on, but we are considering an abstract situation).

The situation for Actor-like neurons is very different. Instead of each member of the population aiming to learn the same function, each is learning whatever function it needs to implement in order to make δ a positive as possible. By way of the hypothetical Actor-Critic neural implementation, suppose the *pattern* of activity over a large population of Actor neurons drives the behavior of the animal that influences the reward signal and the state transitions of the animal's environment. Instead of trying to learn to respond with identical levels of activity, as the Critic neurons do, these neurons must learn to contribute in the best way the can to the *collective action* of the population to produce favorable consequences in the animal's environment as signaled by more reward than expected, that is, by positive values of δ.

Each Actor neuron implements its own Law of Effect (Section **??**), with positive δ signaling satisfaction, and negative δ signaling discomfort—so δ is acting as a common reward signal. What makes this interesting and challenging is that individually any one of these neurons may have only a tiny bit of control over δ because it is contributing just one component of the overall pattern that influences δ. How can each Actor neuron learn under these conditions to "do the right thing" so that the collective action of the population produces positive reward?

A surprising result is that if the agents in a population implement certain reinforcement learning algorithms, the population as a whole can learn to produce collective actions that increase the common reward signal even when the agents cannot communicate with one another. Each agent faces its own reinforcement learning task in which its influence on the reward signal is deeply buried in the noise created by the influences of other agents. Moreover, since the other agent's are learning as well, each agent's task is nonstationary. Each agent faces its own nonstationary reinforcement learning task, being 'unaware' that its reward depends on the activity of other agents in addition to its own.

Two features of the Actor learning algorithm are essential for collective learning. First is its use of contingent eligibility traces. By keeping information about what actions were taken in what states, contingent eligibility traces allow credit for reward, or blame for punishment, to be apportioned among the agents according to the parts their individual actions played in the collective actions that influenced the reward signal. Learning with non-contingent eligibility traces does not work at all in these kinds of problems. Second, there has to be variability in the collective actions of the population in order to explore the space of collective actions. A population of Actor agents probabilistically explores the space of collective actions because each agent independently explores its own action space through the variability in its output. This is a very simple way for a population to explore; more sophisticated methods are possible if agents can communicate with one another to introduce dependencies among their actions.

In multi-agent reinforcement learning (and in game theory) this scenario is known as a *cooperative game* or a *team problem*. Each agent receives the same reward signal, but the signal is potentially influenced by the actions of other agents in the population, and possibly by all of them. It is a cooperative game, or a team problem, because the agents are united in seeking to increase the same reward signal: there are no conflicts of interest among the agents. An even more challenging multi-agent problem is when different agents receive different reward signals, where each reward signal can be influenced by the actions of other agents. This is called a *competitive game*. Agents might be able to cooperate to produce higher reward for each, or there might be conflicts of interest among them, meaning that actions that are good for some agents are bad for others. Even deciding what the right collective action should be is a non-trivial aspect of game theory. The competitive game scenario is likely also relevant to neuroscience. Details of the study of cooperative and non-cooperative game problems for populations of reinforcement learning elements is beyond the scope of this book. The Bibliographical and Historical Remarks section at the end of this chapter cites a selection of the relevant publications.

13.11 Model-Based Methods in the Brain

Reinforcement learning's distinction between model-free and model-based algorithms is proving to be useful for thinking about animal learning and decision processes. Section **??** discusses how this distinction aligns with that between habitual and goal-directed animal behavior. An agent using a model-free method adjusts its policy through direct experience with the reinforcement consequences of its actions. It has to execute an action in a state and observe the consequences in order to update the action's probability and/or it's action-value. A model-based agent, in contrast, selects actions based on explicit knowledge of the expected consequences of its actions and the rewards they are expected to deliver. Model-free methods allow efficient action selection but require relearning if reward contingencies change, whereas model-based methods can quickly adjust to changes without the need for acting under the new contingencies. Animals appear to use both methods, and various hypotheses have been put forward about the conditions under which one or the other predominates.

The hypothesis discussed above about how the brain might implement an Actor-Critic algorithm is relevant only to an animal's habitual mode of behavior because the basic Actor-Critic method is model-free. What neural mechanisms are responsible for producing goal-directed behavior, and how do they interact with those underlying habitual behavior? These questions are motivating an increasing number of neuroscience experiments.

One way to investigate the question of what brain structures are involved in these behavioral modes is to inactivate an area of a rat's brain and then observe what the rat does in a goal-devaluation experiment (Section **??**). Results from experiments like this indicate that the Actor-Critic hypothesis described above, that places the Actor in the dorsal striatum, is too simple. Inactivating one part of the dorsal striatum, the dorsolateral striatum (DLS), impairs habit learning, causing the animal to rely more on goal-directed processes. On the other hand, inactivating the dorsomedial striatum (DML) impairs goal-directed processes, requiring the animal to rely more on habit learning. Results like these support the view that the DLS in rodents is involved in model-free learning, whereas their DMS is involved in model-based learning. Results of studies with human subjects using functional imaging and with non-human primates support the view that the analogous structures in the primate brain are differentially involved in these two behavioral modes.

Other studies identify activity associated with model-based processes in the prefrontal cortex of the human brain, which is the front-most part of the frontal cortex implicated in executive function, including planning and decision making. Specifically implicated is the orbitofrontal cortex (OFC), the part of the prefrontal cortex immediately above the eyes. Functional neuroimaging reveals strong activity in the OFC related the subjective reward value of biologically significant stimuli, as well as activity related to the reward expected as a consequence of actions. In goal-devaluation experiments, the OFC is more strongly activated when the devalued choice is made as opposed to the nondevalued choice.

Another structure involved in model-based behavior is the hippocampus, a structure critical for memory and spatial navigation. A rat needs a functioning hippocampus to navigate a maze in the goal-directed manner that led Tolman to the idea of that animals use models, or cognitive maps, in selecting actions (Section **??**). The hippocampus is also essential for our human ability to imagine new experiences. The results most directly relevant to planning—the process needed to enlist an environment model in making decisions—come from experiments that decode the activity of neurons in the hippocampus to determine what part of space hippocampal activity is representing on a moment-to-moment basis. When a rat pauses at a choice point in a maze, the representation of space in the hippocampus sweeps forward (and not backwards) along the possible paths the animal can take from that point. This suggests that the hippocampus is part of a system that uses an environment model to simulate possible future state sequences in order to assess the consequences of possible actions in order to make decisions. This is a form of planning.

The results described above are just the tip-of-the-iceberg as far as understanding the neural mechanisms of goal-directed, or model-based, learning and decision making. There are many additional questions, some raised by the results obtained so far. For example, how can areas as structurally similar as the DLS and DMS be essential components of modes of learning and behavior that are as different as the model-free and model-based algorithms suggested by computational reinforcement learning? Are separate structures responsible for (what we call) the transition and reward components of an environment model? Is all planning conducted at decision time via simulation of possible future courses of action as the forward sweeping activity in the hippocampus suggests? Or are models sometimes engaged in the background to refine or recompute value information as illustrated by the Dyna architecture? How does the brain arbitrate between the use of the habit and goal-directed systems? Is there, in fact, a clear separation between these systems?

The evidence is not pointing to a positive answer to this last question. Summarizing the situation, Doll, Simon, and Daw (2012) write that "model-based influences appear ubiquitous more or less wherever the brain processes reward information," and this is true even in the regions thought to be critical for model-free learning. This includes the dopamine signals themselves, which can exhibit the influence of model-based information in addition to the reward prediction errors thought to be basis of model-free processes. Continuing neuroscience research informed by reinforcement learning's model-free and model-based distinction will sharpen the understanding of these processes. It is also likely that a better understanding of these neural mechanisms will suggest algorithms that combine model-free and model-based methods is novel ways.

13.12 Addiction

Understanding the neural basis of drug abuse is a high-priority goal of neuroscience with the potential to produce new treatments for this serious health problem. One view is that drug craving is the result of the same motivation and learning processes that lead us to seek the natural rewarding experiences that serve our biological needs. Addictive substances, by being intensely rewarding, effectively co-opt our natural mechanisms of learning and decision making. This is plausible given that many—though not all—drugs of abuse increase levels of dopamine either directly or indirectly in regions around terminals of dopamine neuron axons in the striatum, a brain structure firmly implicated in normal reward-based learning (Section 13.7). But the self-destructive behavior associated with drug addiction is not characteristic of normal learning. What is different about dopamine-mediated learning when the reward is an addictive drug? Is addiction the result of normal learning in response to substances that were largely unavailable throughout our evolutionary history, so that evolution could not select against their damaging effects? Or do addictive substances somehow interfere with normal dopamine-mediated learning?

The reward prediction error hypothesis of dopamine neuron activity and its connection to TD learning are the basis of an influential model of addiction due to Redish (2004). The model is based on the observation that administration of cocaine and some other addictive drugs produces a transient increase in dopamine. In the model this dopamine surge is assumed to increase the TD error δ in a way that cannot be cancelled out by changes in the value function. In other words, whereas δ is reduced to the degree that a normal reward is predicted by antecedent events (Section

This model accounts for some features of addicted behavior, but It is far from being a complete model of addiction (as Redish clearly notes). Dopamine appears not to play a critical role in all forms of addiction, and not everyone is equally susceptible to developing addictive behavior. Moreover, it does not include the changes in many circuits and brain regions that accompany chronic drug taking. Nevertheless, the model illustrates how reinforcement learning theory can be enlisted in the effort to understand a major health problem. In a similar manner, reinforcement learning theory has been influential in the recent development of Computational Psychiatry, which informs efforts to understand mental illness through mathematical and computational methods.

13.13 Summary

The neural pathways involved in the brain's reward system are enormously complex and incompletely understood, but neuroscience research directed toward understanding these pathways and their role in animal behavior is progressing rapidly. Some of this research has been influenced by the theory of reinforcement learning as presented in this book. The objectives of this chapter have been to assist readers in appreciating this influence and to acquaint them with theories of brain function that have played a part in shaping some features of reinforcement learning algorithms.

The most striking instance of how reinforcement learning has influenced neuroscience is the correspondence between the TD error and the phasic activity of dopamine neurons. We provided some detail about the series of experiments from the laboratory of Wolfram Schultz that convinced many neuroscientists that during learning dopamine neurons come to respond to unpredicted rewards and to the earliest predictors of reward, and we explained the basics of why the TD error closely matches these results. TD learning was developed independently of these experimental results, which did not exist at the time of its development, being motivated by the computational problem of efficient prediction and by animal behavior in classical conditioning experiments as described in Chapter **??**. When the data from Schultz's laboratory appeared in the early 1990s, TD learning was a ready-made model for the behavior of the dopamine neurons they observed. This correspondence opened a fruitful channel connecting the theoretical underpinnings of reinforcement learning, with its ties to optimal control and decision theory, to the study of reinforcement learning in the brain—a channel that is continuing to enrich both experimental and theoretical research in neuroscience.

We described some of the abundant evidence that dopamine plays a critical role in many forms of learning. A similar role is played by the TD error in reinforcement learning algorithms that use reward predictions to address the temporal credit assignment problem. The actor-critic type of reinforcement learning algorithm most clearly exemplifies a reinforcing role of reward prediction errors. We discussed the artificial neural-network version of an actor-critic algorithm in which a neuron-like Critic implements TD learning, sending the TD error as a reward signal to a neuron-like Actor that implements a Law-of-Effect type of learning rule.

Features of this actor-critic network turn out to align well with facts about reward processing in the brain. Although the critic and the actor elements implement different learning rules, the TD error is the reinforcing signal for both, acting as an error to be reduced by the critic and as a reward signal to be increased by the actor. The type of eligibility trace an element uses determines which role the TD error plays: non-contingent in the case of the critic element and contingent in the case of the actor element. It is consistent with the wide dispersion of dopamine to many brain areas that the dopamine signal can function in both roles, with properties of each target region determining which role it plays. The actor element's learning rule is a three-factor learning rule involving pre- and post-synaptic signals (to use the neural terms) plus a reward signal provided by the TD error.

Prime candidates for sites in the brain at which instrumental learning takes place are the synapses by which fibers from the cerebral cortex activate neurons in the dorsal striatum. Fibers carrying dopamine signals contact many of these synapses to bring together the three factors necessary for actor-like learning, and studies have demonstrated that the strengths of these synapses decrease with paired pre- and postsynaptic activity unless this is accompanied by stimulation of dopamine neurons, in which case the synaptic strengths increase. This suggests that something like the actor's learning rule may be at work in modifying the strengths of these synapses. Adding to the picture of dopamine acting analogously to the TD error's role in actor-like learning are mechanisms that make dopamine signaling to these synapses temporally precise, something that would facilitate a credit-assignment role of these signals.

Klopf's hypothesis of the "hedonistic neuron" had significant influence on our work on reinforcement learning, including an influence on the neural-network implementation of the actor-critic system. Arguing that maximizing is a better foundation for intelligent behavior than homeostasis, Klopf hypothesized that single neurons are self-interested agents that attempt of increase a neuron-local analog of pleasure and decrease a neuron-local analog of pain by means of a synaptic learning rule that works like the Law-of-Effect. We drew attention to the chemotactic behavior of the bacterium *Escherichia coli* as an example of a single cell that exhibits a related behavioral strategy as it is attracted or repelled by various molecules it encounters as it propels itself through its liquid environment. Imagining a neuron as a metaphorical "swimmer" in a medium composed of all the feedback loops in which it participates is a vivid way to appreciate the possibility that neuronal behavior, like the bacterium's, may best be understood as closed-loop, goal-directed interaction with a complex environment.

Klopf's hypothesis did not appeal to a specialized reward signal, as our approach to reinforcement learning does, but it included the idea of eligibility traces which play important roles in reinforcement leaning algorithms. We reviewed some of the history of stimulus traces in psychological learning theories in Chapter **??**. In this chapter we focused on eligibility traces in the neural terms of Klopf's original idea: synaptically-local traces of past pre- and postsynaptic activity. (We cite similar ideas by others in the Bibliographical and Historical Remarks section below.) Unlike the eligibility traces used in the algorithms presented in this book, which have exponentially decaying profiles, Klopf's idea was that eligibility profiles reflect the durations of the many feedback pathways in which the neuron is imbedded. We simplified this for the sake of theoretical and computationally expediency, but more complicated eligibility traces along the lines Klopf proposed may improve the performance of the algorithms.

Although the existence of synaptic eligibility traces like Klopf proposed has not been demonstrated, neuroscientists are actively studying synaptic plasticity that is sensitive to the timing of the various signals involved. We briefly described *spike-timing-dependent plasticity* (STDP), in which the relative timing of pre- and postsynaptic activity determine the direction of synaptic change. Models of this process include eligibility-like traces, although they have much shorter time courses than the traces Klopf hypothesized. Nevertheless, the growing understanding of STDP suggests that synaptically-localized eligibility-like traces are present in the nervous system. Evidence is also accumulating for the existence of *reward-modulated spike-timing-dependent plasticity*, a form of STDP that depends on a neuromodulator such as dopamine in addition to pre-

and post-synaptic activity. The result is very much like the actor's learning rule in the actor-critic network.

A conspicuous feature of the dopamine system is that fibers releasing dopamine project widely to multiple parts of the brain. If dopamine acts as a reward signal like the TD-error does for the actor in the actor-critic network, then a relevant question is how would an actor-critic network work if it consisted of many actor elements, each using the same TD-error to modulate learning? We addressed this question with a brief introduction to the *collective behavior* of reinforcement learning systems. This is properly part of the subject of *multi-agent reinforcement learning*, which is beyond the scope of this book, but we discussed *team* and *game* problems. In a team problem each agent receives the same reward signal which evaluates the collective behavior of the team. In a *game* problem, each agent receives an individualized reward signal, but every agent can influence the reward signals of every agent. Since it is a common assumption that each dopamine fiber carries basically the same signal, the team problem most relavent. With each team member employing a sufficiently capable learning algorithm, the team can act collectively to improve performance of the entire team as evaluated by the globally-broadcast reward signal, even if the team members do not directly communicate with one another. This ability or reinforcement learning agents to function as team members is another indicator of the importance of reinforcement learning to understanding brain function.

13.14 Conclusion

This chapter only touches the surface of how the neuroscience of reinforcement learning and the development of reinforcement learning algorithms in computer science and engineering have influenced one another. Most features of reinforcement learning algorithms owe their design to purely computational considerations, but some have been influenced by hypotheses about neural learning mechanisms. Remarkably, as experimental data has accumulated about the brain's reward processes, many of the purely computationally-motivated features of the algorithms are turning out to be consistent with neuroscience data.

Most striking is the correspondence between the TD error and the phasic responses of dopamine neurons in the brain. This correspondence is not the result of an attempt to model neuroscience data: the relevant behavior of dopamine neurons was not discovered until many years after the development of TD learning. Dopamine's role in reward-based learning is not its only function, and other chemical messengers are critical for learning, but we believe the correspondence between TD learning and dopamine signaling demonstrates a deep principle of reinforcement learning. Other features of computational reinforcement learning, such eligibility traces and the ability of teams of reinforcement learning agents to learn to act collectively under the influence of a globally-broadcast reward signal, may also turn out to parallel experimental data as neuroscientists continue to unravel the neural basis of animal learning.

13.15 Bibliographical and Historical Remarks

The number of publications treating parallels between the neuroscience of learning and decision making and the approach to reinforcement learning presented in this book is truly enormous. We can cite only a small selection. Niv (2009), Dayan and Niv (2008), and Gimcher (2011) are good places to start. Glimcher (2003) introduces the field of Neuroeconomics, in which reinforcement learning contributes to the study of the neural basis of decision making from an economics perspective.

13.1 Marr's three levels were originally four in Marr and Poggio (1976), separating the algorithmic and representational level into two.

13.2 There are many good expositions of basic neuroscience. Kandel, Schwartz, Jessell, Siegelbaum, and Hudspeth (2013) is an authoritative and very comprehensive source. Sterling and Laughlin (2015) consider neural design in terms of the engineering constraints nervous systems must satisfy.

13.3 Berridge and Kringelbach (2008) review the neural basis of reward and pleasure, pointing out that reward processing has many dimensions and involves many neural systems. Berridge and Robinson (1998) present experimental results supporting a distinction between the hedonic impact of a stimulus, which they call "liking" and the motivational effect, which they call "wanting." Hare, O'Doherty, Camerer, Schultz, and Rangel (2008) examine the neural basis of value-related signals from an economics perspective, distinguishing between goal values, decision values, and prediction errors. Decision value is goal value minus action cost. See also Rangel, Camerer, and Montague (2008), Rangel and Hare (2010). and Peters and Büchel (2010).

13.4 The connection between the TD errors and the phasic responses of dopamine neurons was most prominently introduced by Schultz, Montague, and Dayan (1997). The earliest recognition of this connection of which we are aware was made by Montague, Dayan, Nowlan, Pouget, and Sejnowski (1992) who proposed a TD-error-modulated Hebbian learning rule motivated by results on dopamine signaling from Schultz's group. They show how a diffuse modulatory system might guide map development in the vertebrate brain. The connection was also pointed out in a Neuroscience abstract by Quartz, Dayan, Montague, and Sejnowski (1992). Montague and Sejnowski (1994) emphasized the importance of prediction in the brain and outlined how predictive Hebbian learning modulated by TD errors could be implemented via a diffuse neuromodulatory system, such as the dopamine system. Montague, Dayan, Person, and Sejnowski (1995) presented a model of honeybee foraging using the TD error. The model was based on research by Hammer, Menzel, and colleagues (Hammer and Menzel, 1995; Hammer, 1997) showing that the neuromodulator octopamine acts as a reinforcement signal in the honeybee. Montague et al. pointed out that dopamine likely plays a similar role in the vertebrate brain. Barto (1995) related the Actor-Critic architecture to basal-ganglionic circuits and discussed the relationship between TD learning and the main results from Schultz's group. Houk, Adams, and Barto (1995) elaborated the Actor-Critic/basal ganglia hypothesis. The reward prediction error hypothesis of dopamine neuron activity was first explicitly put forward by Montague, Dayan, and Sejnowski (1996).

Dayan and Abbot's computational neuroscience book (Dayan and Abbott, 2001) contains a useful chapter focusing on TD learning and dopamine neuron activity. Dayan and Niv (2008) discuss strengths and weaknesses of the reward prediction error hypothesis. Gimcher (2011) reviews the empirical findings that support the hypothesis and emphasizes its significance for contemporary neuroscience.

13.5 Schultz's 1998 survey article (Schultz, 1998) is a good entrée into the very extensive literature on the reward predicting signaling of dopamine neurons. Sterling and Laughlin (2015) discuss the very extensive axonal branching of dopamine neuron axons and its significance as a neural design principle. Saddoris, Cacciapaglia, Wightmman, Carelli (2015) present results showing that dopamine neurons do not send the same signal to all target regions; the signals can be specialized for different target regions. O'Doherty, Dayan, Friston, Critchley, and Dolan (2003) describe a functional brain imaging study supporting the existence of signals like TD errors in the human brain.

13.6 This section roughly follows Barto (1995) in explaining how TD errors mimic the main results from Schultz's group on the phasic responses of dopamine neurons.

13.7 This section is largely based on Takahashi, Schoenbaum, and Niv (2008) and Niv (2009). . Barto (1995) and Houk, Adams, and Barto (1995) speculated about possible connections between the Actor-Critic algorithm (Barto, Sutton, and Anderson, 1983) and the basal ganglia. From functional magnetic resonance imaging of human subjects while engaged in instrumental conditioning, O'Doherty, Dayan, Schultz, Deichmann, Friston, and Dolan (2004) suggested that the Actor and the Critic are most likely located respectively in the dorsal and ventral striatum. Waelti, Dickinson, and Schultz (2001) demonstrated that dopamine responses follow the basic principles of psychological learning theory, including exhibiting the blocking phenomenon. Greybiel (2000) is a brief primer on the basal ganglia. Comments on the benefit of using δ as a reinforcement signal instead of R are from Sutton's dissertation (Sutton, 1984).

13.8 Frey and Morris (1997) proposed the idea of a "synaptic tag" for the induction of long-lasting strengthening of synaptic efficacy. Though not unlike an eligibility trace, the tag was hypothesized to consist of a temporary strengthening of a synapse that could be transformed into a long-lasting strengthening by subsequent neuron activation.

@ARTICLEHe-etal-2015, author = K. He and M. Huertas and S. Z. Hong and X. Tie and J. W. Hell and H. Shouval and A. Kirkwood, title Distinct Eligibility Traces for LTP and LTD in Cortical Synapses, journal = Neuron, volume = 88, number = 3, pages = 528-538, year = 2015

For plasticity and mechanism for eligibility traces: Wickens and Kotter (1995)

On STDP in medium spiny neurons: Dopamine Receptor Activation Is Required for Corticostriatal Spike-Timing-Dependent Plasticity, Verena Pawlak and Jason N. D. Kerr The Journal of Neuroscience, 5 March 2008, 28(10): 2435-2446; doi: 10.1523/JNEUROSCI.4402-07.2008

Timing is not Everything: Neuromodulation Opens the STDP Gate Verena Pawlak,1,* Jeffery R. Wickens,2 Alfredo Kirkwood,3 and Jason N. D. Kerr1,* Front Synaptic Neurosci. 2010; 2: 146. Published online Oct 25, 2010. doi: 10.3389/fnsyn.2010.00146

Relevant to this may be the model by computational neuroscientists Rajesh Rao and Terrence Sejnowski showing that STDP could be the result of a TD-like process in which a postsynaptic potential combines with a 10 milliseconds later backpropagating potential from a postsynaptic spike to effect a synapse's weight as a prediction error of the form $V_t - V_{t-1}$. This model requires a non-contingent eligibility trace that lasts only about 10 milliseconds. Commentary on Rao and Sejnoswki's implemention of TD with STDP. He says that R and S argue that "if synapses were to implement a temporal-difference learning rule, then they would be expected to exhibit the sort of temporally asymmetric plasticity that has indeed been observed." Dayan says the their scheme would compute an S-B type signal and not a TD signal. So basically: R and T discuss possible biophysical mechanisms by which neurons could compute signals like TD errors, and Dayan (2002) discusses some of the computational issues that arise from their ideas. Dayan also emphasizes that temporally asymmetric Hebbian learning rules are best seen as predictive rather than correlational.

How TD could be implemented with STDP

Neuroscientists Verena Pawlak and Jason Kerr have shown that increases in the strengths of synapses that cortical inputs make with striatal medium spiny neurons requires dopamine in addition to appropriate timing of pre- and post-synaptic spiking. Determining how plasticity depends on details of the timing of dopaminergic input remains to be worked out. Critical questions like this, as well questions about the existence and nature of eligibility traces, remain to be answered by continuing experimental research.

STDP model: Jesper Sjstrm and Wulfram Gerstner (2010), Scholarpedia, 5(2):1362.

Robert Legenstein, Dejan Pecevski, and Wolfgang Maass of Austria's Graz University of Technology showed that a model of reward-modulated STDP together with variable spontaneous activity of neuron-like elements could account for the results of Fetz's experiment.

Selectionist Theories of the Brain: Edelman, Adams, Fernando, Changeaux, etc. (Fernando et al., 2012), (Adams, 1998)

13.9 hedonistic neurons

Klopf's hypothesis of the "hedonistic neuron" (Klopf 1972, 1982) influenced us to present in 1983 the actor-critic algorithm as an artificial neural network with a single neuron-like element implementing a Law-of-Effect-like learning rule employing eligibility traces at its "synapses" (Barto, Sutton, and Anderson, 1983). Unknown to us at that time (and also unknown to us when we wrote the first edition of this book) were similar theories by others. Physiologist T. J. Crow of Scotland's University of Aberdeen presented a hypothesis in 1968 that emphasized the need to address the time delay between neural activity and its consequences in a reward-modulated form of synaptic plasticity. His solution was that a wave of neuronal activity

> leads to a short-term change in the cells involved in the wave such that they are picked out from a background of cells not so activated. ... such cells are rendered sensitive by the short-term change to a reward signal ... in such a way that if such a signal occurs before the end of the decay time of the change the synaptic connexions between the cells are made more effective. (Crow, 1968)

Crow argued against previous proposals that reverberating neural circuits play this role by pointing out that the effect of a reward signal on such a circuit would "... establish the synaptic connexions leading to the reverberation (that is to say, those involved in activity at the time of the reward signal) and not those on the path which led to the adaptive motor output." Crow further postulated that reward signals are delivered via a "distinct neural fiber system," presumably the one into which Olds and Milner (1954) tapped, that would transform synaptic connections "from a short into a long-term form."

In another farsighted hypothesis about how the brain might implement instrumental learning about which we were unaware when writing the first edition of this book, Robert Miller of New Zealand's University of Otago in 1981 proposed a learning process for synapses following the Law of Effect that included the eligibility concept:

> ... it is envisaged that in a particular sensory situation neurone B, by chance, fires a 'meaningful burst' of activity, which is then translated into motor acts, which then change the situation. It must be supposed that the meaningful burst has an influence, *at the neuronal level*, on all of its own synapses which are active at the time ... thereby making a preliminary selection of the synapses to be strengthened, though not yet actually strengthening them. ...The strengthening signal ... makes the final selection ... and accomplishes the definitive change in the appropriate synapses. ((Miller, 1981), p. 81)

Miller's hypothesis also included a critic-like mechanism, that he called a "sensory analyzer unit," that worked according to classical conditioning principles to provide reinforcement signals to neurons so that they would learn to move from lower- to higher-valued states, thus paralleling the use of the TD error as a reward signal instead of just state values themselves.

A related though different idea, which MIT's Sebastian Seung (2003) called the "hedonistic synapse," is that synapses individually adjust the probability that they release neurotransmitter in the manner of the Law of Effect: if reward follows release, the release probability increases, and decreases if reward follows failure to release. This is essentially the same as the learning scheme Marvin Minsky used in his 1954 Princeton Ph.D. dissertation, where he called the synapse-like learning element a SNARC (Stochastic Neural-Analog Reinforcement Calculator). Synaptic eligibility is involved in these ideas too, although it is contingent on the activity of an individual synapse instead of the postsynaptic neuron.

The metaphor of a neuron using a learning rule related to bacterial chemotaxis was discussed by Barto (1989) in relation to reinforcement learning algorithms. Berg (1975) Koshland's extensive study of bacterial chemotaxis was in part motivated by similarities between some features of bacteria and those of neurons (Koshland,1980). The classic work on chemotaxis and other animal movement strategies is Fraenkel and Gunn (1961). Shimansky (2009) proposed a synaptic learning rule somewhat similar to Seung's mentioned above in which each synapse individually acts like a chemotactic bacterium. In this case a collection of synapses "swims" toward attractants in the high-dimensional space of synaptic weight values. Montague, Dayan, Person, and Sejnowski (1995) proposed a chemotaxic-like model of the bee's foraging behavior involving the neuromodulator octopamine. The view that a controller is actually engaged in controlling its inputs rather than an external system was influenced by the Perceptual Control Theory of behavior developed by Powers (1973).

13.10 Collective behavior

ALOPEX

Cite Hayak machine for society idea: whatshisname?

P. L. Bartlett and J. Baxter, A biologically plausible and locally optimal learning algorithm for spiking neurons.

(Yagishita et al., 2014) Critical time window for dopamine action on MSN spines: 0.3 to 2 seconds. Maximal at 0.6 sec.

Theoretical papers on reward-modulated STDP: Baras and Meir, 2007; Florian 2007, Izhikevich 2007, Legenstein et al., 2008, Vasilaki et al. 2009; Fremaux et al. 2010; Potjans et al. 2010 (see bibliography of "Timing is not Everything: Neuromodulation Opens the STDP Gate")

@ARTICLEFremaux-etal-2010, author = N. Frémaux and H. Sprekeler and W. Gerstner, title = Functional requirements for reward-modulated spike-timing-dependent plasticity, journal = The Journal of Neuroscience, volume = 30, number = 40, year = 2010, pages = 13326-13337

Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity R. V. Florian (XOR net) Neural Computation, 2007 - MIT Press June 2007, Vol. 19, No. 6, Pages 1468-1502

R. Legenstein, D. Pecevski, W. Maass A Learning Theory for Reward-Modulated Spike-Timing-Dependent Plasticity with Application to Biofeedback PLoS Computational Biology, october 2008, Vol 4, No. 10, pp. ??

Adams has it right that a global reward signal is needed for an ensemble to learn, but he regards it as an error signal, which is not the correct way to think about it. But then says that it can be a scalar: so not really an error at all. Essentially talks about reward modulated Hebbian synapses (though without the timing considerations?) Calls it ”synaptic Darwinism”

A special kind of team problem occurs in biofeedback training. Suppose the reward signal broadcast to a population of reinforcement learning agents depends on the activity of only one of the agents. With repeated trials, the responsible agent will learn to do the right thing because its activity has a clear causal influence on the reward signal, making its reinforcement learning problem easy. The other agents will continue to explore without finding any correlation between their actions and the reward. Of course, a key difficulty in doing this in practice is to make the reward signal depend on just one agent.

University of Washington’s Eberhard Fetz did just this in a striking experiment in which he conditioned monkeys to increase the firing rates of specific cortical neurons (Fetz, 1969). Recording the activity of a single neuron in the motor cortex of an unanesthetized monkey, Fetz and his assistants rewarded high rates of this neuron’s activity by delivering banana-flavored pellets to the monkey. This was done with several monkeys, who also heard sounds or saw the deflection of a meter that varied with the activity of the cortical neuron. After sufficient training, “monkeys consistently and rapidly increased the activity of newly isolated cells” (Fetz, 1969). They could increase the neuron’s activity up to between 50 and 500 percent above the unconditioned level of activity. Monkeys already experienced with the task could rapidly increase the firing rate of the isolated neurons without audio or visual feedback. As a control, the experimenters presented pellets and feedback according to a record of their delivery during a previous reinforcement period, but with no relation to the activity of the monitored neuron. In these cases, the firing rate remained at or below the unconditioned level. They could also condition neurons to fire more slowly than their unconditioned rates by rewarding decreases in activity rate. Although the rationale of this experiment was to develop a method to study the influence of single motor cortical neurons on movement, the experiment supports the ubiquity of reinforcement learning in the brain.

13.11 Model based

Daw and Shohamy (2008) proposed that while dopamine signaling connects well to habitual, or model-free, behavior, other processes are involved in goal-directed, or model-based, behavior.

@ARTICLEJohnson-Redish-2007, author = A. Johnson and A. D. Redish, title = Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point, journal = The Journal of neuroscience, volume = 27, number = 45, year = 2007, pages = 12176-12189

humans with functional imaging lots of model-based activity in the striatum: support their view that no clear separation between habit and goal-dircted substrates hippocampus for imagining fictitious experiences review goal-devaluation experiments showing DLS and DMS distinction

Model-based RL: @ARTICLEDoll-etal-2012, author = B. B. Doll and D. A. Simon and N. D. Daw, title = The Ubiquity of Model-Based Reinforcement Learning, journal = Current Opinion in Neurobiology, volume = 22, pages = 1-7, year = 2012

Goal-Directed learning Brain planning mechanisms: Daw, Gershman, Seymour, Dayan, Dolan (2011) @ARTICLEValentin-etal-2007, author = V. V. Valentin and A. Dickinson and J. P. O'Doherty, title = Determining the Neural Substrates of Goal-Directed Learning in the Human Brain, journal = The Journal of Neuroscience, volume = 27, number = 15, pages = 4019-4026, year = 2007 @ARTICLERangel-Hare-2010, author = A. Rangel and T. Hare, title = Neural computations associated with goal-directed choice, journal = Current opinion in neurobiology, volume = 20, number = 2, pages = 262-270, year = 2010

13.12 The model of addiction that eliminates negative TD errors for addictive stimuli is due to Redish (2004). Keiflin and Janak (2015) review connections between TD errors and addiction. Nutt, Lingford-Hughes, Erritzoe, and Stokes (2015) critically evaluate the hypothesis that addiction is due to a disorder of the dopamine system. Adams, Huys, and Roiser (2015) review the new field of Computational Psychiatry

Bibliography

Adams, P. (1998). Hebb and Darwin. *Journal of Theoretical Biology*, 195:419–438.

Adams, R. A., Huys, Q. J. M., and Roiser, J. P. (2015). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, doi:10.1136/jnnp-2015-310737.

Barto, A. G. (1989). From chemotaxis to cooperativity: Abstractexercises in neuronal learning strategies. In Durbin, R., Maill, R., and Mitchison, G., editors, *The Computing Neuron*, pages 73–98. Addison-Wesley, Reading, MA.

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

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike elements that can solve difficult learningcontrol problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846. Reprinted in J. A. Anderson and E. Rosenfeld (eds.), *Neurocomputing: Foundations of Research*, pp. 535-549, MIT Press, Cambridge, MA, 1988.

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

Berg, H. C. (1975). Chemotaxis in bacteria. *Annual review of biophysics and bioengineering*, 4(1):119–136.

Berridge, K. C. and Kringelbach, M. L. (2008). Affective neuroscience of pleasure: reward in humans and animals. *Psychopharmacology*, 199(3):457–480.

Berridge, K. C. and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, 104(2):1068–1076.

Crow, T. J. (1968). Cortical synapses and reinforcement: a hypothesis. *Nature*, 219:736–737.

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

Daw, N. D. and Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5):593–620.

Dayan, P. (2002). Matters temporal. *Trends in Cognitive Sciences*, 6(3):105–106.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA.

Dayan, P. and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22:1–7.

Fernando, C., Szathmáry, E., and Husbands, P. (2012). Selectionist and evolutionary approaches to brain function: A critical appraisal. *Frontiers in Computational Neuroscience*, 6.

Fetz, E. E. (1969). Operant conditioning of cortical unit activity. *Science*, 28:955–958.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902.

Fraenkel, G. S. and Gunn, D. L. (1961). *The orientation of animals: kineses, taxes and compass reactions*. Dover, Oxford, England. This an expanded version of the first edition published by Oxford University Press in 1940.

Frey, U. and Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(6616):533–536.

Gimcher, 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.

Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. MIT Press, Cambridge, MA.

Glimcher, P. W. and Fehr, E., editors (2013). *Neuroeconomics: Decision making and the brain, Second Edition*. Academic Press.

Greybiel, A. M. (2000). The basal ganglia. *Current Biology*, 10(14):R509–R511.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, 28(22):5623–5630.

Hassabis, D. and Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306.

He, K., Huertas, M., Hong, S. Z., Tie, X., Hell, J. W., Shouval, H., and Kirkwood, A. (2015). Distinct eligibility traces for LTP and LTD in cortical synapses. *Neuron*, 88(3):528–538.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* John Wiley and Sons Inc., New York. Reissued by Lawrence Erlbaum Associates Inc., Mahwah NJ, 2002.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. MIT Press, Cambridge, MA.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J., editors (2013). *Principles of Neural Science, Fifth Edition*. McGraw-Hill Companies, Inc.

Keiflin, R. and Janak, P. H. (2015). Dopamine prediction errors in reward learning and addiction: Ffrom theory to neural circuitry. *Neuron*, 88(2):247–263.

Klopf, A. H. (1972). Brain function and adaptive systems—A heterostatic theory. Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in *Proceedings of the International Conference on Systems, Man, and Cybernetics*, 1974, IEEE Systems, Man, and Cybernetics Society, Dallas, TX.

Klopf, A. H. (1982). *The Hedonistic Neuron: A Theory of Memory, Learning, andIntelligence*. Hemisphere, Washington, D.C.

Koshland, D. E. (1980). *Bacterial Chemotaxis as a Model Bhavioral System*. Raven Press, New York.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1):145–163.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. Technical Report AI Memo 357, Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., Arai, R., and Kaneko, T. (2009). Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. *The Journal of Neuroscience*, 29(2):444–453.

Miller, R. (1981). *Meaning and Purpose in the Intact Brain: A Philosophical, Psychological, and Biological Account of Conscious Process*. Clarendon Press, Oxford.

Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., and Sejnowski, T. J. (1992). Using aperiodic reinforcement for directed self-organization during development. In Hanson, S. J., Cohen, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems: Proceedings of the 1992 Conference*, pages 969–976, San Mateo, CA. Morgan Kaufmann.

Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377(6551):725–728.

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

Montague, P. R. and Sejnowski, T. J. (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learningmechanisms. *Learning & Memory*, 1:1–33.

Niv, Y. (2009). Reinforcement leaning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154.

Nutt, D. J., Lingford-Hughes, A., Erritzoe, D., and Stokes, P. R. A. (2015). The dopamine theory of addiction: 40 years of highs and lows. *Nature Reviews Neuroscience*, 16:305–312.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.

Olds, J. and Milner, P. (1954). Positive reinforcement produced by electrical stimulation of the septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.

Peters, J. and B"

chel, C. (2010). Neural representations of subjective reward value. *Behavioural brain research*, 213(2):135–141.

Pezzulo, G., van der Meer, M. A. A., Lansink, C. S., and Pennartz, C. M. A. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Science*, 18(12):647–657.

Powers, W. T. (1973). *Behavior: The Control of Perception.* Aldine de Gruyter, Chicago. 2nd expanded edition 2005.

Quartz, S., Dayan, P., Montague, P. R., and Sejnowski, T. J. (1992). Expectation learning in the brain using diffuse ascending connections. In *Society for Neuroscience Abstracts*, volume 18, page 1210.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556.

Rangel, A. and Hare, T. (2010). Neural computations associated with goal-directed choice. *Current opinion in neurobiology*, 20(2):262–270.

Rao, R. P. N. and Sejnowski, T. J. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Computation*, 13:2221–2237.

Redish, D. A. (2004). Addiction as a computational process gone awry. *Science*, 306(5703):1944–1947.

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

Romo, R. and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, 63(3):592–624.

Saddoris, M. P., Cacciapaglia, F., Wightmman, R. M., and Carelli, R. M. (2015). Differential dopamine release dynamics in the nucleus accumbens core and shell reveal complementary signals for error prediction and incentive motivation. *The Journal of Neuroscience*, 35(33):11572–11582.

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

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1598.

Schultz, W. and Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, 63(3):607–624.

Selfridge, O. G. (1978). Tracking and trailing: Adaptation in movement strategies. Technical report, Bolt Beranek and Newman, Inc. Unpublished report.

Selfridge, O. G. (1984). Some themes and primitives in ill-defined systems. In Selfridge, O. G., Rissland, E. L., and Arbib, M. A., editors, *Adaptive Control of Ill-Defined Systems*, pages 21–26. Plenum Press, NY. Proceedings of the NATO Advanced Research Institute on Adaptive Control of Ill-defined Systems, NATO Conference Series II, Systems Science, Vol. 16.

Sescousse, G., Caldú, X., Segura, B., and Dreher, J.-C. (2013). Precessing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(4):681–696.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6):1063–1073.

Shimansky, Y. P. (2009). Biologically plausible learning in neural networks: a lesson from bacterial chemotaxis. *Biological Cybernetics*, 101(5-6):379–385.

Simon, D. A. and Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision tasks in humans. *The Journal of Neuroscience*, 31(14):5526–5539.

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

Sterling, P. and Laughlin, S. (2015). *Principles of Neural Design*. MIT Press, Cambridge, MA.

Sutton, R. S. (1984). *Temporal Credit Assignment in ReinforcementLearning*. PhD thesis, University of Massachusetts, Amherst, MA.

Takahashi, Y., Schoenbaum, G., and 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(1):86–99.

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645.

Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, 27(15):4019–4026.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412(6842):43–48.

Wickens, J. and Kötter, R. (1995). Cellular models of reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 187–214. MIT Press, Cambridge, MA.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1619.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476.