Psychological and Neuroscientific Connections with Reinforcement Learning (preprint)

Ashvin Shah
Department of Psychology
University of Sheffield

2012

Abstract

The field of Reinforcement Learning (RL) was inspired in large part by research in animal behavior and psychology. Early research showed that animals can, through trial and error, learn to execute behavior that would eventually lead to some (presumably satisfactory) outcome, and decades of subsequent research was (and is still) aimed at discovering the mechanisms of this learning process. This chapter describes behavioral and theoretical research in animal learning that is directly related to fundamental concepts used in RL. It then describes neuroscientific research that suggests that animals and many RL algorithms use very similar learning mechanisms. Along the way, I highlight ways that research in computer science contributes to and can be inspired by research in psychology and neuroscience.

Please note:

This is a preprint of a chapter for the book *Reinforcement Learning: State of the Art*, edited by Marco Wiering and Martijn van Otterlo. This document does not follow the format used in the book, but the text should be pretty much the same. If you cite this chapter, below is a bibtex you can use. Thanks.

```
@InCollection{Shah-PsychRL-2012,
author={A. Shah},
title={Psychological and neuroscientific connections with {R}einforcement {L}earning},
booktitle={Reinforcement Learning: State of the Art},
chapter={16},
year={2012},
editor={M. Wiering and M. {van Otterlo}},
publisher={Springer-Verlag},
address={Berlin Heidelberg},
pages = {507--537}
}
```

1 Introduction

In the late nineteenth century, the psychologist Edward L. Thorndike conducted experiments in which he placed a hungry animal (most famously a cat) in a "puzzle box," the door of which could be opened only after a certain sequence of actions, such as pulling a chain and then pressing a lever, have been executed (Thorndike, 1911). Placed outside the box and in view of the animal was some food. A naive animal would exhibit many behaviors that had no effect on the door, such as batting at the door or even grooming itself. At some point, the animal might, by chance, pull the chain and later, also by chance, press the lever, after which the door would open and the animal could escape and consume the food. When that animal was again hungry and placed into the box, it would execute less of the useless behaviors and more of the ones that led to the opening door. After repeated trials, the animal could escape the box in mere seconds.

The animal's behavior is familiar to readers of this book as it describes well the behavior of a reinforcement learning (RL) agent engaged in a simple task. The basic problems the animal faces—and solves—in the puzzle box are those that an RL agent must solve: given no instruction and only a very coarse evaluation signal, how does an agent learn what to do and when to do it in order to better its circumstances? While RL is not intended to be a model of animal learning, animal behavior and psychology form a major thread of research that led to its development (Chapter 1, Sutton and Barto 1998). RL was also strongly influenced by the work of Harry Klopf (Klopf, 1982), who put forth the idea that hedonistic ("pleasure seeking") behavior emerges from hedonistic learning processes, including processes that govern the behavior of single neurons.

In this chapter I describe some of the early experimental work in animal behavior that started the field and developed the basic paradigms that are used even today, and psychological theories that were developed to explain observed behavior. I then describe neuroscience research aimed at discovering the brain mechanisms responsible for such behavior. Rather than attempt to provide an exhaustive review of animal learning and behavior and their underlying neural mechanisms in a single chapter, I focus on studies that are directly-related to fundamental concepts used in RL and that illustrate some of the experimental methodology. I hope that this focus will make clear the similarities—in some cases striking—between mechanisms used by RL agents and mechanisms thought to dictate many types of animal behavior. The fact that animals can solve problems we strive to develop artificial systems to solve suggests that a greater understanding of psychology and neuroscience can inspire research in RL and machine learning in general.

2 Classical (or Pavlovian) Conditioning

Prediction plays an important role in learning and control. Perhaps the most direct way to study prediction in animals is with classical conditioning, pioneered by Ivan Pavlov in Russia in the early 1900s. While investigating digestive functions of dogs, Pavlov noticed that some dogs that he had worked with before would salivate before any food was brought out. In what began as an attempt to account for this surprising behavior, Pavlov developed his theory of conditioned reflexes (Pavlov, 1927): mental processes (e.g., perception of an auditory tone) can cause a physiological reaction (sali-

vation) that was previously thought to be caused only by physical processes (e.g., smell or presence of food in the mouth). Most famously, the sound of a ringing bell that reliably preceded the delivery of food eventually by itself caused the dog to salivate. This behavior can be thought of as an indication that the dog has learned to predict that food delivery will follow the ringing bell.

2.1 Behavior

While Pavlov's drooling dog is the enduring image of classical conditioning, a more studied system is the nictitating membrane (NM) of the rabbit eye (Gormezano et al., 1962), which is a thin "third eyelid" that closes to protect the eye. Typically, the rabbit is restrained and an air puff or a mild electric shock applied to the eye (the unconditioned stimulus, US) causes the NM to close (the unconditioned response, UR). If a neutral stimulus that does not by itself cause the NM to close (conditioned stimulus, CS), such as a tone or a light, is reliably presented to the rabbit before the US, eventually the CS itself causes the NM to close (the conditioned response, CR). Because the CR is acquired with repeated pairings of the CS and the US (the *acquisition* phase of an experiment), the US is often referred to as a *reinforcer*. The strength of the CR, often measured by how quickly the NM closes or the likelihood that it closes before the US, is a measure of the predictive strength of the CS for the animal. After the CR is acquired, if the CS is presented but the US is omitted (*extinction* phase), the strength of the CR gradually decreases.

Manipulations to the experimental protocol can give us a better idea of how such predictions are learned. Particularly instructive are manipulations in the timing of the CS relative to the US and how the use of multiple CSs affects the predictive qualities of each CS. (These manipulations are focused on in Sutton and Barto (1987) and Sutton and Barto (1981), which describe temporal difference models of classical conditioning.)

Two measures of timing between the CS and the subsequent US are (Figure 1, top): 1) interstimulus interval (ISI), which is the time between the onset of the CS and the onset of the US, and 2) trace interval (TI), which is the time between the offset of the CS and the onset of the US. A simple protocol uses a short and constant ISI and a zero-length TI (*delay conditioning*). For example, the tone is presented briefly (500 ms) and the air puff is presented at the end of the tone. When the TI is greater than zero (*trace conditioning*), acquisition and retention of the CR are hindered. If the ISI is zero, i.e., if the CS and US are presented at the same time, the CS is useless for prediction and the animal will not acquire a CR. There is an optimal ISI (about 250 ms for the NM response) after which the strength of the CR decreases gradually. The rate of decrease is greater in trace conditioning than it is in delay conditioning. In addition, the rate of acquisition decreases with an increase in ISI, suggesting that it is harder to predict temporally distant events.

The use of several CSs (*compound stimuli*) reveals that learning also depends on the animal's ability to predict the upcoming US. Figure 1, middle, illustrates example protocols in which one CS (e.g., a tone, referred to as CSA) is colored in light gray and the other (a light, CSB) is colored in dark gray. In *blocking* (Kamin, 1969), the US is paired with CSA alone and the animal acquires a CR. Afterwards, the simultaneous presentation of CSA and CSB is paired with the US for a block of trials. Subsequent presentation of CSB alone elicits no CR. Because CSA was already a predictor of the US, CSB holds no predictive power. In *conditioned inhibition*, two types of stimulus

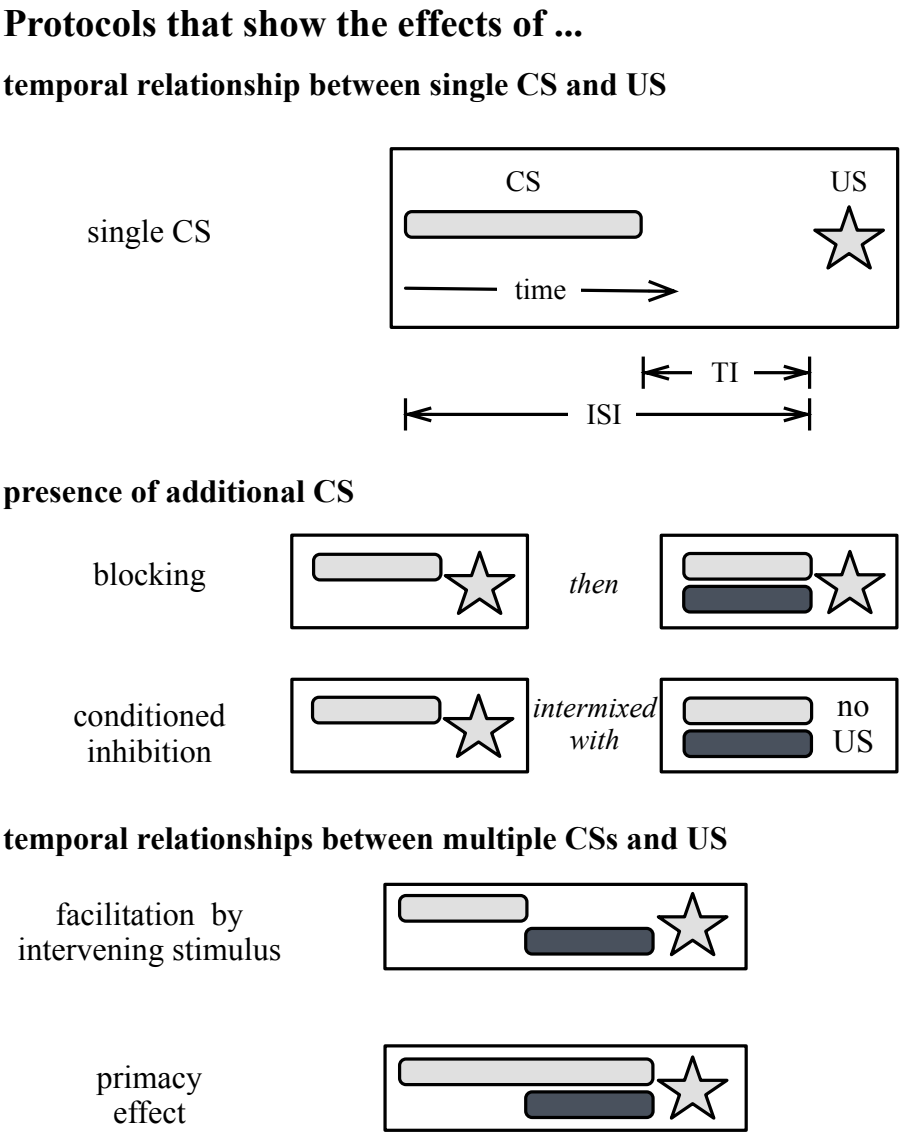

Figure 1: Schematic of stimulus presentation for different types of classical conditioning tasks. Time progresses along the horizontal axis. The star indicates the occurrence of the US. The light gray rectangle indicates the occurrence and time course of one CS, while the dark gray rectangle indicates that for another CS.

presentations are intermixed during training: CSA alone paired with the US, and the simultaneous presentation of CSA and CSB with the US omitted. Subsequent pairing of CSB alone with the US results in a lower rate of CR acquisition relative to animals that did not experience the compound stimulus. CSB was previously learned as a reliable predictor that the US will not occur.

In the above two examples, the two CSs had identical temporal properties. Other protocols show the effects of presenting compound stimuli that have different temporal properties (serial compound stimuli) (Figure 1, bottom). As mentioned earlier, acquisition of a CR is impaired in trace conditioning (TI > 0). However, if another CS is presented during the TI, the acquisition of the CR in response to the first CS is facilitated (*facilitation by intervening stimulus*). A related protocol results in *higher-order* conditioning, in which a CR is first acquired in response to CSB. Then, if CSA is presented prior to CSB, a CR is acquired in response to CSA. In a sense, CSB plays the role of reinforcer.

In the *primacy effect*, CSA and CSB overlap in time. The offset time of each is the same and immediately precedes the US, but the onset time of CSA is earlier than that of CSB (Figure 1, bottom). Because CSB has a shorter ISI than CSA, one may expect that a CR would be elicited more strongly in response to CSB alone than to CSA alone. However, the presence of CSA actually results in a decrease in the strength of the CR in response to CSB alone. More surprising is a prediction first discussed in Sutton and Barto (1981). They presented a model of classical conditioning that was first trained with CSB paired with the US (delay conditioning with a short ISI), and then CSA (with an earlier onset time) was presented as well. Even though the strength of the association between CSB and the response—which represents the predictive qualities the CS—had reached its asymptotic level, it decreased when CSA was presented. Such a finding is seemingly incongruous with the effects of the ISI and the phenomenon of blocking. Sutton and Barto (1987) replicated this result, and Kehoe et al. (1987) confirmed this prediction experimentally.

2.2 Theory

The blocking effect suggests that learning occurs when the unexpected happens (Kamin, 1969) as opposed to when two things are correlated. This idea led to the development of the famous *Rescorla-Wagner* model of classical conditioning (Rescorla and Wagner, 1972), in which the presence of a US during a trial is predicted by the sum of associative strengths between each CS present during the trial and the US. Changes in associative strengths depend on the accuracy of the prediction. For every CS i present during a trial:

$$\Delta w(i) = \alpha \left(r - \sum_i w(i)x(i) \right),$$

where $w(i)$ is the associative strength between CS i and the US, $x(i) = 1$ if CS i is present and 0 otherwise, r is the maximum amount of conditioning the US can produce (analogous to the "magnitude" of the US), and α is a step-size parameter. (Note that this notation differs from the original version.) If the US is perfectly-predicted (i.e., if $r - \sum_i w(i)x(i) = 0$), the associative strength of another CS subsequently added (with an initial associative strength of zero) will not increase. This influential model captures several features of classical conditioning (e.g., blocking). Also, as first noted in Sutton

and Barto (1981), it is similar to the independently-developed *Widrow-Hoff* learning rule (Widrow and Hoff, 1960), showing the importance of prediction-derived learning.

The Rescorla-Wagner model is a *trial level* account of classical conditioning in that learning occurs from trial to trial as opposed to at each time point within a trial. It cannot account for the effects that temporal properties of stimuli have on learning. In addition, Sutton and Barto (1987) point out that animal learning processes may not incorporate mechanisms that are dependent on the concept of the trial, which is essentially a convenient way to segregate events.

Temporal difference (TD) models (Sutton, 1988; Sutton and Barto, 1998), which have foundations in models of classical conditioning (Sutton and Barto, 1981; Barto and Sutton, 1982; Sutton and Barto, 1987), were developed in part to do prediction learning on a *real-time* level. In the following TD model of classical conditioning (Sutton and Barto, 1987) (using notation that is similar to the equation above), let r_t be the presence (and magnitude) of the US at time step t, $x_t(i)$ be 1 if CS i is present at time t and 0 otherwise, and $w_t(i)$ be the associative strength between CS i and the US at time t. At each time point,

$$w_{t+1}(i) = w_t(i) + \alpha \left(r_t + \gamma \left[\sum_i w_t(i) x_t(i) \right]^+ - \left[\sum_i w_t(i) x_{t-1}(i) \right]^+ \right) \overline{x}_t(i),$$

where γ is the familiar temporal discount factor, $[y]^+$ returns zero if $y < 0$, and $\overline{x}_t(i)$ is an eligibility trace, e.g., $\beta \overline{x}_{t-1}(i) + (1 - \beta) x_{t-1}(i)$, where $0 \leq \beta < 1$.

At each time point, weights are adjusted to minimize the difference between $r_t + \gamma [\sum_i w_t(i) x_t(i)]^+$ and $[\sum_i w_t(i) x_{t-1}(i)]^+$ (i.e., the *temporal difference error*). These are temporally successive predictions of the same quantity: upcoming USs (more precisely, $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$). The former prediction incorporates more recent information (r_t and $x_t(i)$) and serves as a target for the latter, which uses information from an earlier time point ($x_{t-1}(i)$). The eligibility trace (which restricts modification to associations of CSs that were recently present), discount factor, and explicit dependence on time allow the model to capture the effect on learning of temporal relationships among stimuli within a trial. Also, because the prediction at time t trains the prediction at time $t-1$, the model accounts for higher order conditioning.

2.3 Summary and additional considerations

Classical conditioning reveals behavior related to an animal's ability to predict that a US will follow a CS. The prediction is thought to arise from the strengthening of an association between the CS and the US. Associations are strengthened when there is a prediction error (i.e., the animal does not already predict that the US will follow the CS), there is contingency (the US follows the CS most of the time), and there is contiguity (the US occurs shortly after the CS) (Schultz, 2006).

Methods used in TD models were developed in large part to capture the types of behaviors observed in animals engaged in classical conditioning experiments. Whereas other accounts of prediction use only the actual outcome (e.g., the US) as a training signal, TD models use the difference between temporally successive predictions of the outcome (the TD error) as a training signal. Besides providing for a better account of animal behavior, such bootstrapping has proven to be a powerful computational technique (Sutton, 1988).

Additional experiments in classical conditioning include characterizing how the form of the CR changes depending on contingencies between the CS and US and how brain mechanisms mediate such behavior. Computational accounts devised to explain behavior based on neural mechanisms further increase our understanding of how animals learn to predict (Gluck, 2008; Mirolli et al., 2010; Moore and Choi, 1997; Brandon et al., 2002).

Finally, the UR and CR can be considered *Pavlovian actions* in that the animal does not develop or choose to execute them. Rather, the US is a salient stimulus that causes the animal to emit the UR. The CR is not learned, but arises from the learned association between the stimulus (CS) and the eventual outcome (US). Because the animal has very little control over its environment and behavior, classical conditioning allows us to focus on prediction rather than decision-making. Of course, one of the benefits of prediction is that it allows us to better control what we experience. Such control is the focus of the next section.

3 Operant (or Instrumental) Conditioning

Classical conditioning experiments present the animal with a salient stimulus (the US) contingent on another stimulus (the CS) regardless of the animal's behavior. Operant conditioning experiments present a salient stimulus (usually a "reward" such as food) contingent on specific actions executed by the animal (Thorndike, 1911; Skinner, 1938). (The animal is thought of as an instrument that operates on the environment.) Thorndike's basic experimental protocol described at the beginning of this chapter forms the foundation of most experiments described in this section.

3.1 Behavior

The simplest protocols use single-action tasks such as a rat pressing the one lever or a pigeon pecking at the one key available to it. Behavior is usually described as the "strength" of the action, measured by how quickly it is initiated, how quickly it is executed, or likelihood of execution. Basic protocols and behavior are analogous to those of classical conditioning. During acquisition (action is followed by reward), action strength increases, and the reward is referred to as the reinforcer. During extinction (reward is omitted after the acquisition phase), action strength decreases. The rates of action strength increase and decrease also serve as measures of learning.

Learning depends on several factors that can be manipulated by the experimenter. In most experiments, the animal is put into a deprived state before acquisition (e.g., it's hungry if the reward is food). Action strength increases at a faster rate with deprivation; hence, deprivation is said to increase the animal's "drive" or "motivation." Other factors commonly studied include the magnitude of reward (e.g., volume of food, where an increase results in an increase in learning) and delay between action execution and reward delivery (where an increase results in a decrease).

Factors that affect learning in single-action tasks also affect selection, or decision-making, in *free choice* tasks in which more than one action is available (e.g., there are two levers). The different actions lead to outcomes with different characteristics, and the strength of an action relative to others is usually measured by relative likelihood (i.e., choice distribution). Unsurprisingly, animals more frequently choose the action

that leads to a reward of greater magnitude and/or shorter delay.

We can quantify the effects of one factor in terms of another by examining choice distribution. For example, suppose action A leads to a reward of a constant magnitude and delay and action B leads to an immediate reward. By determining how much we must decrease the magnitude of the immediate reward so that the animal shows no preference between the two actions, we can describe a temporal discount function. Although most accounts in RL use an exponential discount function, behavioral studies support a hyperbolic form (Green and Myerson, 2004).

Probability of reward delivery is another factor that affects choice distribution. In some tasks where different actions lead to rewards that are delivered with different probabilities, humans and some animals display *probability matching*, where the choice distribution is similar to the relative probabilities that each action will be reinforced (Siegel and Goldstein, 1959; Shanks et al., 2002). Such a strategy is clearly suboptimal if the overall goal is to maximize total reward received. Some studies suggest that probability matching may be due in part to factors such as the small number of actions that are considered in most experiments or ambiguous task instructions (i.e., participants may be attempting to achieve some goal other than reward maximization) (Shanks et al., 2002; Gardner, 1958; Goodnow, 1955).

Naive animals usually do not execute the specific action(s) the experimenter wishes to examine; the animal must be "taught" with methods drawn from the basic results outlined above and from classical conditioning. For example, in order to draw a pigeon's attention to a key to be pecked, the experimenter may first simply illuminate the key and then deliver the food, independent of the pigeon's behavior. The pigeon naturally pecks at the food and, with repeated pairings, pecks at the key itself (*autoshaping*, Brown and Jenkins 1968). Although such Pavlovian actions can be exploited to guide the animal towards some behaviors, they can hinder the animal in learning other behaviors (Dayan et al., 2006).

If the movements that compose an action can vary to some degree, a procedure called *shaping* can be used to teach an animal to execute a specific movement by gradually changing the range of movements that elicit a reward (Eckerman et al., 1980). For example, to teach a pigeon to peck at a particular location in space, the experimenter defines a large imaginary sphere around that location. When the pigeon happens to peck within that sphere, food is delivered. When the pigeon consistently pecks within that sphere, the experimenter decreases the radius, and the pigeon receives food only when it happens to peck within the smaller sphere. This process continues until an acceptable level of precision is reached.

Shaping and autoshaping modify the movements that compose a single action. Some behaviors are better described as a chain of several actions executed sequentially. To train animals, experimenters exploit higher-order conditioning (or *conditioned reinforcement*): as with classical conditioning, previously neutral stimuli can take on the reinforcing properties of a reinforcer with which they were paired. In *backward chaining*, the animal is first trained to execute the last action in a chain. Then it is trained to execute the second to last action, after which it is in the state from which it can execute the previously acquired action (Richardson and Warzak, 1981). This process, in which states from which the animal can execute a previously learned action act as a reinforcers, continues until the entire sequence is learned.

3.2 Theory

To account for the behavior he had observed in his experiments, Thorndike devised his famous Law of Effect:

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond. (Chapter 5, page 244 of Thorndike 1911.)

Learning occurs only with experience: actually executing the action and evaluating the outcome. According to Thorndike, action strength is due to the strength of an association between the response (or action) and the situation (state) from which it was executed. The basic concepts described in the Law of Effect are also used in many RL algorithms. In particular, action strength is analogous to *action value*, $Q(s, a)$, which changes according to the consequences of the action and determines behavior in many RL schemes.

This type of action generation, where the action is elicited by the current state, is sometimes thought of as being due to a *stimulus-response* (SR) association. Though the term is a bit controversial, I use it here because it is used in many neuroscience accounts of behavior (e.g., Yin et al. 2008) and it emphasizes the idea that behavior is due to associations between available actions and the current state. In contrast, actions may be chosen based on explicit predictions of their outcomes (discussed in the next subsection).

Thorndike conducted his experiments in part to address questions that arose from Charles Darwin's Theory of Evolution: do humans and animals have similar mental faculties? His Law of Effect provides a mechanistic account that, coupled with variability (exploration), can explain even complicated behaviors. The SR association, despite its simplicity, plays a role in several early psychological theories of behavior (e.g., Thorndike 1911; Hull 1943; Watson 1914). New SR associations can be formed from behavior generated by previously-learned ones, and a complicated "response" can be composed of previously-learned simple responses. (Such concepts are used in hierarchical methods as well, Grupen and Huber 2005; Barto and Mahadevan 2003; Hengst 2012.) This view is in agreement with Evolution: simple processes of which animals are capable explain some human behavior as well. (In Chapter 6 of his book, Thorndike suggests that thought and reason are human qualities that arise from our superior ability to learn associations.)

As with animals, it may be very difficult for a naive artificial agent to learn a specific behavior. Training procedures developed by experimentalists, such as backward chaining and shaping, can be used to aid artificial agents as well (Konidaris and Barto, 2009; Selfridge et al., 1985; Ng et al., 1999). A type of shaping also occurs when reinforcement methods are used to address the structural credit assignment problem (Barto, 1985): when an "action" is composed of multiple elements that can each be modified, exactly *what* did an agent just do that led to the reward?

Psychological theories that primarily use concepts similar to the SR association represent a view in which behavior is accounted for only by variables that can be directly-observed (e.g., the situation and the response). Taken to the extreme, it is controversial and cannot explain all behavior. The next subsection discusses experiments that show that some behavior is better explained by accounts that do allow for variables that cannot be directly-observed.

3.3 Model-based versus model-free control

The idea that an action is elicited from an SR association is a *model-free* account of behavior—an explicit prediction of the outcome of the action plays no role in its execution. However, the results of other experiments led to theories in which an action is elicited from an explicit prediction of its outcome—*model-based* accounts. For example, rats that had prior exposure to a maze in the absence of any reward quickly learned the same route to a reward later presented as rats that had been trained with the reward all along. This behavior, and that from other maze experiments, led the American psychologist Edward Tolman to suggest that rats form models of the environment which specify relations between states and between actions and their outcomes (Tolman, 1948). Whereas an SR association is strengthened in model-free accounts, an association between an action and its predicted outcome is strengthened in many types of model-based accounts.

To more directly determine if an explicit representation of outcome had any effect on behavior, *goal devaluation* (or *revaluation*) procedures have been devised (Dickinson, 1985; Dickinson and Balleine, 1994; Balleine et al., 2009). In a typical experiment, an animal is trained to perform some operant task and then, in a separate situation, its motivation for the reward is changed (e.g., by making it ill just after it eats the food). The animal is then placed in the original task environment during an extinction test. (To prevent new learning, it is not given the outcome.) If behavior is different than that of animals that have not undergone the devaluation procedure (e.g., if action strength during extinction declines at a faster rate), we can infer that it is controlled at least in part by some representation of its outcome.

This is indeed the case for animals that have not had extensive training, suggesting that they used model-based mechanisms (in many cases, the animal does not execute the action even on the first trial after devaluation). However, the behavior of animals that have had extensive training is not affected by devaluation. Such behavior is considered habitual (Dickinson, 1985), and it is better accounted for by model-free mechanisms. Dickinson (1985) suggests that behavior is the result of one of two processes: as the animal learns the task, there is a high correlation between rate of behavior and rate of reinforcer delivery, and this correlation strengthens action-outcome (AO) associations. Meanwhile, SR associations are being learned as well. However, with extensive training, behavior rate plateaus and, because its variance is very small, its correlation with reinforcer delivery is much weaker. Thus the AO association is decreased and SR associations dominate control.

The availability of multiple control mechanisms has functional advantages. Because the outcome is explicitly represented, a model-based controller does not need much experience in order to make appropriate decisions once it has encountered the reward. Consequently, it can adapt quickly—even immediately—if the reward structure changes (i.e., it's flexible). However, making decisions based on predictions generated by a model

has high computational and representational demands. A model-free controller requires fewer resources but much more experience in order to develop SR associations so that appropriate decisions are made. Such requirements make them less flexible.

Further insights can be drawn from computational models of decision-making that use multiple controllers. In a particularly relevant account (Daw et al., 2005), model-based and model-free RL algorithms were used to select actions in a simulated goal devaluation task. Because the task changes when the goal is devalued, a measure of uncertainty was incorporated into the algorithms. At each state, the action suggested by the controller with the lower uncertainty was executed. Because the model-free controller required much experience, the model-based controller dominated control early on. However, because the model-based controller relied on multiple predictions to select actions, the minimum uncertainty it could achieve was higher than that of the model-free controller, and the model-free controller dominated later. The multiple controller scheme accounted for much of the behavior observed experimentally and provided computationally-grounded explanations for how such behavior is developed. Other types of multiple controller models suggest a more passive arbitration scheme, where the simpler controller simply selects an action faster if it is sufficiently trained (Ashby et al., 2007; Shah and Barto, 2009; Shah, 2008). The simpler mechanisms may be useful in the construction of hierarchical behavior such as skills or "chunks" of actions (Shah, 2008).

Goal devaluation experiments provide elegant behavioral evidence that animals use different control mechanisms at different stages in learning. Equally elegant experiments show that different brain systems are involved in mediating the different control mechanisms (Balleine et al., 2009). Computational models inspired by these results use learning and control mechanisms that are thought to occur in the corresponding brain areas. These connections are discussed further in Section 5.3.

3.4 Summary and additional considerations

Operant conditioning shows us that the consequences of behavior can be used to train animals to develop specific actions and adjust the strength—including likelihood of execution—of those specific actions. Such actions may be referred to as *operant actions* to distinguish them from Pavlovian actions. Psychological theories that explain such development have much in common with RL: that a scalar reinforcement signal strengthens an association between stimulus and response (SR, consistent with model-free mechanisms) or action and outcome (AO, model-based).

Several other types of processes gleaned from experiments in animal behavior may be of interest to the reader. As discussed earlier, an animal's motivational state affects its behavior. Early attempts to explain this effect have led to the drive theory of Hull, where *any* learned SR association is strengthened when the animal is more motivated (Hull, 1943), and the incentive value theory of Tolman, where the strength of an action depends on the motivational state of the animal when it learned that action, even if it is no longer motivated (Tolman, 1949). These ideas are further explored experimentally (Dickinson and Balleine, 1994; Pompilio and Kacelnik, 2005) and within an RL framework (Niv et al., 2006b).

The experiments described in this section deliver a reward when an action is executed, but other experiments examine behavior when the action results in an aversive outcome (such as a shock) or the absence of a salient outcome. Other types of manip-

ulations include delivering the outcome only after some number of actions have been executed or amount of time has elapsed since the last delivery (Ferster and Skinner, 1957). These manipulations result in interesting behavior (e.g., the *matching law*, Herrnstein 1961) that may further reveal the underlying processes of learning and control in animals (Staddon and Cerutti, 2003).

In general, we can think of an animal as doing what we develop RL agents to do: modify its behavior so as to maximize satisfaction using only a coarse evaluation of the consequences of behavior as feedback. Although much can be learned by examining animal behavior, we must turn to neuroscience in order to better understand the neural processes that govern such behavior. The next two sections describe evidence suggesting that animals use learning processes remarkably similar to those used in many RL algorithms.

4 Dopamine

Although different types of rewards could have similar effects on behavior, it was unknown if similar or disparate neural mechanisms mediate those effects. In the early 1950s, it was discovered that if an action led to electrical stimulation applied (via an electrode) to the brain, action strength would increase (Olds and Milner, 1954). This technique, known as *intracranial self-stimulation* (ICSS), showed greater effects when the electrodes were strongly stimulating the projections of neurons that release dopamine (DA) from their axon terminals. DA neurons are mostly located in the substantia nigra pars compacta (SNpc, also called the A9 group) (a part of the basal ganglia, BG) and the neighboring ventral tegmental area (VTA, group A10) (Björklund and Dunnett, 2007). As described in the next section, many of the areas DA neurons project to are involved with decision-making and movement. ICSS shows that behavior may be modified according to a global signal communicated by DA neurons. Original interpretations suggested, quite reasonably, that DA directly signals the occurrence of a reward. Subsequent research, some of which is described in this section, shows that DA plays a more sophisticated role (Wise, 2004; Schultz, 2007; Bromberg-Martin et al., 2010; Montague et al., 2004).

4.1 Dopamine as a reward prediction error

To better characterize the role of DA in behavior, researchers recorded the activity of single neurons from VTA and SNpc in awake behaving animals. DA neurons exhibit low baseline firing-rates (< 10 Hz) with short (*phasic*) bursts of firing (henceforth referred to as "*the DA burst*," and the magnitude of the burst refers to the frequency of firing within a burst). Early studies showed that a DA burst occurred in response to sensory events that were task-related, intense, or surprising (Miller et al., 1981; Schultz, 1986; Horvitz, 2000).

In one of these most important studies linking RL to brain processes, Ljungberg et al. (1992) recorded from DA neurons in monkeys while the monkeys learned to reach for a lever when a light was presented, after which they received some juice. Initially, a DA burst occurred in response only to juice delivery. As the task was learned, a DA burst occurred at both the light and juice delivery, and later only to the light (and *not* juice delivery). Finally, after about $30,000$ trials (over many days), the DA burst even

in response to the light declined by a large amount (perhaps due to a lack of attention or motivation, Ljungberg et al. 1992). Similarly, Schultz et al. (1993) showed that as monkeys learned a more complicated operant conditioning task, the DA burst moved from the time of juice delivery to the time of the stimuli that indicated that the monkey should execute the action. If the monkey executed the wrong action, DA neuron activity at the time of the expected juice delivery decreased from baseline. Also, juice delivered before its expected time resulted in a DA burst; the omission of expected juice (even if the monkey behaved correctly) resulted in a decrease in activity from baseline; and the DA burst at the time of juice delivery gradually decreased as the monkey learned the task (Schultz et al., 1997; Hollerman and Schultz, 1998).

The progression of DA neuron activity over the course of learning did not correlate with variables that were directly-manipulated in the experiment, but it caught the attention of those familiar with RL (Barto, 1995). As noted by Houk et al. (1995), there "is a remarkable similarity between the discharge properties of DA neurons and the effective reinforcement signal generated by a TD algorithm..." (page 256). Montague et al. 1996 hypothesized that "the fluctuating delivery of dopamine from the VTA to cortical and subcortical target structures in part delivers information about prediction errors between the expected amount of reward and the actual reward" (page 1944). Schultz et al. (1997) discuss in more detail the relationship between their experiments, TD, and psychological learning theories (Schultz, 2010, 2006, 1998). The importance of these similarities cannot be overstated: a fundamental learning signal developed years earlier within the RL framework appears to be represented—almost exactly—in the activity of DA neurons recorded during learning tasks.

Several studies suggest that the DA burst is, like the TD error, influenced by the prediction properties of a stimulus, e.g., as in blocking (Waelti et al., 2001) and conditioned inhibition (Tobler et al., 2003). When trained monkeys were faced with stimuli that predict the probability of reward delivery, the magnitude of the DA burst at the time of stimuli presentation increased with likelihood; that at the time of delivered reward decreased; and DA neuron activity at the time of an omitted reward decreased to a larger extent (Fiorillo et al., 2003) (but see Niv et al. 2005). When reward magnitude was drawn from a probability distribution, the DA burst at the time of reward delivery reflected the difference between delivered and expected magnitude (Tobler et al., 2005). Other influences on the DA burst include motivation (Satoh et al., 2003), delay of reward delivery, (Kobayashi and Schultz, 2008), and history (if reward delivery was a function of past events) (Nakahara et al., 2004; Bayer and Glimcher, 2005).

In a task that examined the role of DA neuron activity in decision-making (Morris et al., 2006), monkeys were presented with visual stimuli indicating probability of reward delivery. If only one stimulus was presented, DA neuron activity at the time of the stimulus presentation was proportional to the expected value. If two stimuli were presented, and the monkey could choose between them, DA neuron activity was proportional to the expected value of the eventual choice rather than representing some combination of the expected values of the available choices. The authors suggest that these results indicate that DA neuron activity reflects the perceived value of a chosen (by some other process) action, in agreement with a SARSA learning scheme (Niv et al., 2006a) (though the results of other experiments support a Q-learning scheme, Roesch et al. 2007).

4.2 Dopamine as a general reinforcement signal

The studies reviewed above provide compelling evidence that the DA burst reflects a reward prediction error very similar to the TD error of RL. Such an interpretation is attractive because we can then draw upon the rich computational framework of RL to analyze such activity. However, other studies and interpretations suggest that DA acts as a general reinforcer of behavior, but perhaps not just to maximize reward.

The *incentive salience* theory (Berridge and Robinson, 1998; Berridge, 2007; Berridge et al., 2009) separates "wanting" (a behavioral bias) from "liking" (hedonic feelings). Experiments using pharmacological manipulations suggest that opioid—not DA—systems mediate facial expressions associated with pleasure (Berridge and Robinson, 1998). DA could increase action strength without increasing such measures (Tindell et al., 2005; Wyvell and Berridge, 2000), and DA-deficient animals could learn to prefer pleasurable stimuli over neutral ones (Cannon and Palmiter, 2003). The separation offers an explanation for some experimental results and irrational behaviors (Wyvell and Berridge, 2000).

Redgrave et al. (2008) argue that the latency of the DA burst, < 100 ms after stimulus presentation in many cases, is too short and uniform (across stimuli and species) to be based on identification—and hence predicted value—of the stimulus. Rather, the DA burst may be due to projections from entirely subcortical pathways that respond quickly—faster than DA neurons—to coarse perceptions that indicate that something has happened, but not what it was (Dommett et al., 2005) . More recent experimental results provide evidence that additional phasic DA neuron activity that occurs with a longer latency (< 200 ms) (Joshua et al., 2009; Bromberg-Martin et al., 2010; Nomoto et al., 2010) may be due to early cortical processing and does provide some reward-related information. Very early (< 100 ms) DA activity may signal a sensory prediction error (e.g., Horvitz 2000) that biases the animal to repeat movements so as to determine what it had done, if anything, to cause the sensory event (Redgrave et al., 2008; Redgrave and Gurney, 2006). Later DA activity may be recruited to bias the animal to repeat movements in order to maximize reward.

4.3 Summary and additional considerations

The experiments described this section show that DA acts not as a "reward detector," but rather as a learning signal that reinforces behavior. Pharmacological treatments that manipulate the effectiveness of DA further support this idea. For example, in humans that were given DA agonists (which increase the effectiveness of DA on target neurons) while performing a task, there was an increase in both learning and a representation of the TD error in the striatum (a brain area targeted by DA neurons) (Pessiglione et al., 2006). DA antagonists (which decrease the effectiveness of DA) had the opposite effect.

There are a number of interesting issues that I have not discussed but deserve mention. Exactly how the DA burst is shaped is a matter of some debate. Theories based on projections to and from DA neurons suggest that they are actively suppressed when a predicted reward is delivered (Hazy et al., 2010; Houk et al., 1995). Also, because baseline DA neuron activity is low, aversive outcomes or omitted rewards cannot be represented in the same way as delivered rewards. Theories based on stimulus representation (Ludvig et al., 2008; Daw et al., 2006a), other neurotransmitter systems (Daw

et al., 2002; Phelps and LeDoux, 2005; Wrase et al., 2007; Doya, 2008), and/or learning rules (Frank, 2005; Frank et al., 2004) address this issue. While this section focused on phasic DA neuron activity, research is also examining the effect that long-term (*tonic*) DA neuron activity has on learning and behavior (Schultz, 2007; Daw and Touretzky, 2002). Finally, recent experimental evidence suggests that the behavior of DA neurons across anatomical locations may not be as uniform as suggested in this section (Haber, 2003; Wickens et al., 2007).

While several interpretations of how and to what end DA affects behavior have been put forth, of most interest to readers of this chapter is the idea that the DA burst represents a signal very similar to the TD error. To determine if that signal is used in the brain in ways similar to how RL algorithms is it, we must examine target structures of DA neurons.

5 The Basal Ganglia

Most DA neuron projections terminate in frontal cortex and the basal ganglia (BG), areas of the brain that are involved in the control of movement, decision-making, and other cognitive processes. Because DA projections to the BG are particularly dense relative to projections in frontal cortex, this section focuses on the BG. What follows is a brief (and necessarily incomplete) overview of the BG and its role in learning and control.

5.1 Overview of the basal ganglia

The BG are a set of interconnected subcortical structures located near the thalamus. Scientists first connected their function with voluntary movement in the early twentieth century when post-mortem analysis showed that a part of the BG was damaged in Parkinson's disease patients. Subsequent research has revealed a basic understanding of their function in terms of movement (Mink, 1996), and research over the past few decades show that they mediate learning and cognitive functions as well (Packard and Knowlton, 2002; Graybiel, 2005).

A part of the BG called the striatum receives projections from DA neurons and excitatory projections from most areas of cortex and thalamus. A striatal neuron receives a large number of weak inputs from many cortical neurons, suggesting that striatal neurons implement a form of pattern recognition (Houk and Wise, 1995; Wilson, 2004). Striatal neurons send inhibitory projections to the internal segment of the globus pallidus (GPi), the neurons of which are tonically-active and send inhibitory projections to brain stem and thalamus. Thus, excitation of striatal neurons results in a disinhibition of neurons targeted by GPi neurons. In the case in which activation of neurons targeted by the GPi elicits movements, their disinhibition increases the likelihood that those movements will be executed.

On an abstract level, we can think of pattern recognition at striatal neurons as analogous to the detection of state as used in RL, and the resulting disinhibition of the targets of GPi neurons as analogous to the selection of actions. Corticostriatal synapses are subject to DA-dependent plasticity (Wickens, 2009; Calabresi et al., 2007), e.g., a learning rule roughly approximated by the product of the activity of the striatal neuron and the activities of the cortical and DA neurons that project to it. Thus, the DA

burst (e.g., representing the TD error) can modify the activation of striatal neurons that respond to a particular state according to the consequences of the resulting action. In other words, the BG possess characteristics that enable them to modify the selection of actions through mechanisms similar to those used in RL (Barto, 1995; Doya, 1999; Daw and Doya, 2006; Doll and Frank, 2009; Graybiel, 2005; Joel et al., 2002; Wickens et al., 2007).

Additional pathways within the BG (which are not described here) impart a functionality to the BG useful for behavioral control. For example, actions can be actively facilitated or inhibited, possibly through different learning mechanisms (Frank, 2005; Frank et al., 2004; Hikosaka, 2007). Also, intra-BG architecture appears to be well-suited to implement selection between competing actions in an optimal way (Bogacz and Gurney, 2007; Gurney et al., 2001).

Different areas of cortex—which are involved in different types of functions—project to different areas of the striatum. These pathways stay segregated to a large degree through the BG, to the thalamus, and back up to the cortex (Alexander et al., 1986; Middleton and Strick, 2002). The parallel loop structure allows the BG to affect behavior by shaping cortical activity as well as through descending projections to brain stem. The functional implications of the parallel loop structure are discussed later in this section. The next subsection describes studies that aim to further elucidate the functions of the BG, focusing mostly on areas involved with the selection of movements and decisions.

5.2 Neural activity in the striatum

In most of the experiments described in this subsection, the activities of single neurons in the striatum were recorded while the animal was engaged in some conditioning task. As the animal learned the task, neural activity began to display task-related activity, including activity modulated by reward (Schultz et al., 2003; Hikosaka, 2007; Barnes et al., 2005).

In a particularly relevant study, Samejima et al. (2005) recorded from dorsal striatum (where dorsal means toward the top of the head) of monkeys engaged in a two-action free choice task. Each action led to a large reward (a large volume of juice) with some probability that was held constant over a block of trials, and a smaller reward (small volume) the rest of the time. For example, in one block of trials, the probability that action A led to the large reward was 0.5 and that of action B was 0.9, while in another block the probabilities were, respectively, 0.5 and 0.1. Such a design dissociates the absolute and relative action values (the expected reward for executing the action): the value of action A is the same in both blocks, but it is lower than that of action B in the first block and higher in the second. Choices during a block were distributed between the two actions, with a preference for the more valuable one.

The recorded activity of about one third of the neurons during a block covaried with the value of one of the actions (a lesser proportion covaried with other aspects such as the difference in action values or the eventual choice). In addition, modelling techniques were used to estimate action values online based on experience, i.e., past actions and rewards, and to predict choice behavior based on those estimated action values. Many neurons were found whose activities covaried with estimated action value, and the predicted choice distribution agreed with the observed distribution.

The temporal profile of neural activity within a trial yields further insights. Lau and

Glimcher (2008) showed that dorsal striatal neurons that encoded the values of available actions were more active before action execution, and those that encoded the value of the chosen action were more active after action execution. The results of these and other studies (Balleine et al., 2007; Kim et al., 2009) suggest that action values of some form are represented in dorsal striatum and that such representations participate in action selection and evaluation (though some lesion studies suggest that dorsal striatum may not be needed for learning such values, Atallah et al. 2007).

Analyses described in Samejima et al. (2005) illustrate an approach that has been growing more prominent over the past decade: to correlate neural activity not only with variables that can be directly-observed (e.g., reward delivery or choice), but also with variables thought to participate in learning and control according to theories and computational models (e.g., expected value) (Corrado and Doya, 2007; Daw and Doya, 2006; Niv, 2009). This approach is especially useful in analyzing data derived from functional magnetic resonance imaging (fMRI) methods (Gläscher and O'Doherty, 2010; Montague et al., 2006; Haruno and Kawato, 2006), where the activity of many brain areas can be recorded simultaneously, including in humans engaged in complex cognitive tasks. Note that the precise relationship between the measured signal (volume of oxygenated blood) and neural activity is not known and the signal has a low temporal and spatial resolution relative to single neuron recordings. That being said, analyses of the abundance of data can give us a better idea of the overall interactions between brain areas.

Using fMRI, O'Doherty et al. (2004) showed that, in a task in which the human participant must choose an action from a set, dorsolateral striatum and ventral striatum (where ventral means toward the bottom) exhibited TD error-like signals, while in a task in which the participant had no choice, only ventral striatum exhibited such a signal. These results suggest that dorsal and ventral striatum implement functions analogous to the Actor and Critic (see also Barto 1995; Joel et al. 2002; Montague et al. 2006), respectively, in the Actor-Critic architecture (Barto et al., 1983).

Recordings from single neurons in the ventral striatum of animals support this suggestion to some degree. Information which can be used to evaluate behavior, such as context (or state), some types of actions, and outcome, appear to be represented in ventral striatum (Ito and Doya, 2009; Roesch et al., 2009; Kim et al., 2009). Roughly speaking, while dorsal striatum is more concerned with actions in general, ventral striatum may participate in assigning value to stimuli, but it may also participate in controlling some types of actions and play a more complicated role in behavior (Yin et al., 2008; Humphries and Prescott, 2010; Nicola, 2007).

5.3 Cortico-basal ganglia-thalamic loops

As described earlier, pathways from cortex to the BG to thalamus and back up to cortex form parallel segregated loops (Alexander et al., 1986; Middleton and Strick, 2002). The basic functionality of the BG within a loop—RL-mediated selection or biasing—is though to apply to information that is represented by neural activity within that loop (Wickens et al., 2007; Samejima and Doya, 2007; Graybiel et al., 1994; Houk et al., 2007; Yin et al., 2008; Haruno and Kawato, 2006; Redgrave et al., 2010; Packard and Knowlton, 2002; Cohen and Frank, 2009). Also, because the different cortical areas are involved in different types of functions, the loop structure allows the BG to affect behavior through different mechanisms (e.g., those that govern behavior in classical and operant conditioning, see Sections 2.4 and 3.3). In an interpretation similar to that

put forth by Yin et al. (2008), these are: 1) operant actions learned through model-free mechanisms (i.e., generated by a stimulus-response, SR, association); 2) operant actions learned through model-based mechanisms (action-outcome, AO, association); or 3) the learning of stimulus-outcome, SO, associations, which can result in the execution of Pavlovian actions.

A model-free mechanism is thought to be implemented in a loop involving the dorsolateral striatum (DLS, also called the putamen in primates). The DLS receives cortical projections mainly from primary sensory, motor, and premotor cortices (which will be referred to collectively as sensorimotor cortices, or SMC), providing the DLS with basic sensory and movement information.

A model-based mechanism is thought to be implemented in a loop involving the dorsomedial striatum (DMS, also called the caudate), which receives cortical projections mainly from prefrontal cortex (PFC). The PFC is on the front part of the cortex and has reciprocal connections with many other cortical areas that mediate abstract representations of sensations and movement. PFC neurons exhibit sustained activity (working memory, Goldman-Rakic 1995) that allows them to temporarily store information. RL mechanisms mediated by the DMS may determine which past stimuli should be represented by sustained activity (O'Reilly and Frank, 2006). Also, the PFC is thought to participate in the construction of a model of the environment (Gläscher et al., 2010), allowing it to store predicted future events as well (Mushiake et al., 2006; Matsumoto et al., 2003). Thus, the PFC can affect behavior through a planning process and even override behavior suggested by other brain areas (Miller and Cohen, 2001; Tanji and Hoshi, 2008).

The assignment of value to previously neutral stimuli may be implemented in a loop involving the ventral striatum (VS, also called the nucleus accumbens), which receives cortical projections mainly from the oribtofrontal cortex (OFC, the underside part of the PFC that is just behind the forehead.) The VS and the OFC have connections with limbic areas, such as the amygdala and hypothalamus. These structures, and VS and OFC, have been implicated in the processing of emotion, motivations, and reward (Wallis, 2007; Cardinal et al., 2002; Mirolli et al., 2010).

Most theories of brain function that focus on the loop structure share elements of the interpretation of Yin et al. (2008) (though of course there are some differences) (Samejima and Doya, 2007; Haruno and Kawato, 2006; Daw et al., 2005; Balleine et al., 2009; Ashby et al., 2010; Balleine and O'Dohrety, 2010; Pennartz et al., 2009; Houk et al., 2007; Wickens et al., 2007; Redgrave et al., 2010; Mirolli et al., 2010; Packard and Knowlton, 2002; Cohen and Frank, 2009). Common to all is the idea that different loops implement different mechanisms that are useful for the types of learning and control described in this chapter. Similar to how behavioral studies suggest that different mechanisms dominate control at different points in learning (e.g., control is transferred from model-based to model-free mechanisms, as discussed in Section 3.3), neuroscience studies suggest that brain structures associated with different loops dominate control at different points in learning (previous references and Doyon et al. 2009; Poldrack et al. 2005).

For example, in humans learning to perform a sequence of movements, the PFC (model-based mechanisms) dominates brain activity (as measured by fMRI) early in learning while the striatum and SMC (model-free) dominates activity later (Doyon et al., 2009). These results can be interpreted to suggest that decision-making mechanisms during model-based control lie predominantly within the PFC, while those during

model-free control lie predominantly at the cortico-striatal synapses to the DLS. That is, control is transferred from cortical to BG selection mechanisms (Daw et al., 2005; Niv et al., 2006b; Shah, 2008; Shah and Barto, 2009), in rough agreement with experimental studies that suggest that the BG play a large role in encoding motor skills and habitual behavior (Graybiel, 2008; Pennartz et al., 2009; Aldridge and Berridge, 1998). However, other theories and experimental results suggest that control is transferred in the opposite direction: the BG mediate trial and error learning early on (Pasupathy and Miller, 2005; Packard and Knowlton, 2002), but cortical areas mediate habitual or skilled behavior (Ashby et al., 2007, 2010; Frank and Claus, 2006; Matsuzaka et al., 2007). As discussed in Ashby et al. (2010), part the discrepancy may be due to the use of different experimental methods, including tasks performed by the animal and measures used to define habitual or skilled behavior.

Finally, some research is also focusing on how control via the different loops is coordinated. Behavior generated by mechanisms in one loop can be used to train mechanisms of another, but there is also some communication between the loops (Haber, 2003; Haber et al., 2006; Pennartz et al., 2009; Yin et al., 2008; Balleine and O'Dohrety, 2010; Mirolli et al., 2010; Joel and Weiner, 1994; Graybiel et al., 1994). Such communication may occur within the BG (Pennartz et al., 2009; Graybiel, 2008; Joel and Weiner, 1994; Graybiel et al., 1994) or via connections between striatum and DA neurons (some of which are also part of the BG). The latter connections are structured such that an area of the striatum projects to DA neurons that send projections back to it and also to a neighboring area of striatum (Haber, 2003). The pattern of connectivity forms a spiral where communication is predominantly in one direction, suggestive of a hierarchical organization in which learned associations within the higher-level loop are used to train the lower-level loop (Haruno and Kawato, 2006; Yin et al., 2008; Samejima and Doya, 2007). Again following an interpretation similar to that of Yin et al. (2008), the OFC-VS loop (SO association) informs the PFC-DMS loop (AO), which informs the SMC-DLS loop (SR).

5.4 Summary and additional considerations

BG function is strongly affected by DA, which we can approximate as representing the TD error. The architectural and physiological properties of the BG endow it with the ability to do RL. Experimental studies show that neurons in the dorsal striatum, which communicates with motor-related and decision-making areas of the brain, represent action values. These results suggest that the BG mediate learning and control in ways similar to those used in many RL algorithms.

Studies also suggest that different cortico-basal ganglia-thalamic loops use RL mechanisms on different types of information. The parallel loop structure allows for behavior to be affected through different mechanisms, and communication between loops allows for learning in one mechanism to drive learning in another. The conceptual architecture derived from biological systems suggests a potentially advantageous way to construct hierarchical control architectures for artificial agents, both in what types of information different levels of the hierarchy represent and how the different levels communicate to each other.

Of course, pathways, brain areas, and functional mechanisms other than those described in this section influence learning and control as well. For example, DA affects neural activity directly, striatal activity is shaped by the activities of interneurons (neu-

rons that project to other neurons within the same structure), which change dramatically as an animal learns a task (Graybiel et al., 1994; Graybiel, 2008; Pennartz et al., 2009), and the BG also affects behavior through recurrent connections with subcortical structures (McHaffie et al., 2005). Reward-related activity in frontal cortical areas (Schultz, 2006) are shaped not only through interactions with the BG, but also by direct DA projections and interactions with other brain areas.

Computational mechanisms and considerations that were not discussed here but are commonly used in RL and machine learning have analogs in the brain as well. Psychological and neuroscientific research address topics such as behavior under uncertainty, state abstraction, game theory, exploration versus exploitation, and hierarchical behavior (Dayan and Daw, 2008; Doya, 2008; Gold and Shadlen, 2007; Seger and Miller, 2010; Glimcher and Rustichini, 2004; Daw et al., 2006b; Yu and Dayan, 2005; Wolpert, 2007; Botvinick et al., 2009; Grafton and Hamilton, 2007). Determining how brain areas contribute to observed behavior is a very difficult endeavour. The approach discussed earlier—analyzing brain activity in terms of variables used in principled computational accounts (Corrado and Doya, 2007; Daw and Doya, 2006; Niv, 2009)—is used for a variety of brain areas and accounts beyond that discussed in this section. For example, the field of neuroeconomics—which includes many ideas discussed in this chapter—investigates decision-making processes and associated brain areas in humans engaged in economic games (Glimcher and Rustichini, 2004; Glimcher, 2003).

6 Chapter Summary

RL is a computational formulation of learning, through interaction with the environment, to execute behavior that delivers satisfying consequences. Animals possess this ability, and RL was inspired in large part by early studies in animal behavior and the psychological theories they spawned (Sections 2 and 3). Methodological advances in biology enabled scientists to better characterize learning and control mechanisms of animals by recording the neural activity of animals engaged in learning tasks. RL provides a framework within which to analyze such activity, and it was found that some mechanisms the brain uses are remarkably similar to mechanisms used in RL algorithms (Sections 4 and 5).

In particular, the characterization of DA neuron activity in terms of the TD error (Section 4.1) has forged a stronger and more explicit communication between RL (and machine learning) and psychology and neuroscience. A number of outstanding reviews have been published recently that focus on these connections: Niv 2009; Dayan and Daw 2008; Daw and Doya 2006; Cohen 2008; Doya 2007; Maia 2009; Dayan and Niv 2008. In addition, studies in computational neuroscience show how physiological and architectural characteristics of brain areas can implement specific functions: Wörgötter and Porr 2005; Cohen and Frank 2009; Gurney et al. 2004; Gurney 2009; Doya 2008; Bar-Gad et al. 2003; Joel et al. 2002; Hazy et al. 2010; Mirolli et al. 2010. The integration of computational methods with psychology and neuroscience not only gives us a better understanding of decision-making processes, but it also presents us with new approaches to study disorders in decision-making such as mental disorders and addiction (Maia and Frank, 2011; Kishida et al., 2010; Redish et al., 2008; Redgrave et al., 2010; Graybiel, 2008; Belin et al., 2009).

Finally, more recent research in neuroscience is beginning to further unravel the

mechanisms by which animals develop complex behavior. In particular, rather than use one monolithic control mechanism, animals appear to use multiple control mechanisms (Sections 3.3 and 5.3). Neuroscientific research suggests how these different mechanisms cooperate (Section 5.3). Such schemes contribute to the sophisticated behavior of which animals are capable and may serve as inspiration in our quest to construct sophisticated autonomous artificial agents.

Acknowledgements

I am grateful for comments from and discussions with Andrew Barto, Tom Stafford, Kevin Gurney, and Peter Redgrave, comments from anonymous reviewers, and financial support from the European Community's 7th Framework Programme grant 231722 (IM-CLeVeR).

References

Aldridge, J. W. and Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *The Journal of Neuroscience*, 18:2777–2787.

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

Ashby, F. G., Ennis, J., and Spiering, B. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, 114:632–656.

Ashby, F. G., Turner, B. O., and Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, 14:208–215.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., and O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in ventral and dorsal striatum. *Nature Neuroscience*, 10:126–131.

Balleine, B. W., Delgado, M. R., and Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, 27:8161–8165.

Balleine, B. W., Liljeholm, M., and Ostlund, S. B. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behavioural Brain Research*, 199:43–52.

Balleine, B. W. and O'Dohrety, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35:48–69.

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction, and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71:439–473.

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

Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, 4:229–256.

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*, chapter 11, pages 215–232. MIT Press, Cambridge, Massachuetts, USA.

Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13:341–379.

Barto, A. G. and Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research*, 4:221–235.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernectics*, 13:835–846.

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

Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W., and Everitt, B. J. (2009). Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. *Behavioural Brain Research*, 199:89–102.

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology*, 191:391–431.

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*, 28:309–369.

Berridge, K. C., Robinson, T. E., and Aldridge, J. W. (2009). Dissecting components of reward: 'Liking,' 'wanting,' and learning. *Current Opinion in Pharamacology*, 9:65–73.

Björklund, A. and Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, 30:194–202.

Bogacz, R. and Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, 19:442–477.

Botvinick, M. M., Niv, Y., and Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. *Cognition*, 113:262–280.

Brandon, S. E., Vogel, E. G., and Wagner, A. R. (2002). Computational theories of classical conditioning. In Moore, J. W., editor, *A Neuroscientist's Guide to Classical Conditioning*, chapter 7, pages 232–310. Springer-Verlag, New York, USA.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, 68:815–834.

Brown, P. L. and Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior*, 11:1–8.

Calabresi, P., Picconi, B., Tozzi, A., and DiFilippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neuroscience*, 30:211–219.

Cannon, C. M. and Palmiter, R. D. (2003). Reward without dopamine. *Journal of Neuroscience*, 23:10827–10831.

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

Cohen, M. X. (2008). Neurocomputational mechanisms of reinforcement-guided learning in humans: a review. *Cognitive, Affective, and Behavioral Neuroscience*, 8:113–125.

Cohen, M. X. and Frank, M. J. (2009). Neurocomputational models of the basal ganglia in learning, memory, and choice. *Behavioural Brain Research*, 199:141–156.

Corrado, G. and Doya, K. (2007). Understanding neural coding through the model-based analysis of decision-making. *The Journal of Neuroscience*, 27:8178–8180.

Daw, N. D., Courville, A. C., and Tourtezky, D. S. (2006a). Representation and timing in theories of the dopamine system. *Neural Computation*, 18:1637–1677.

Daw, N. D. and Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16:199–204.

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

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

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006b). Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879.

Daw, N. D. and Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, 14:2567–2583.

Dayan, P. and Daw, N. D. (2008). Connections between computational and neurobiological perspectives on decision making. *Cognitive, Affective, and Behavioral Neuroscience*, 8:429–453.

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

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, 19:1153–1160.

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

Dickinson, A. and Balleine, B. W. (1994). Motivational control of goal-directed action. *Animal Learning and Behavior*, 22:1–18.

Doll, B. B. and Frank, M. J. (2009). The basal ganglia in reward and decision making: computational models and empirical studies. In Dreher, J. and Tremblay, L., editors, *Handbook of Reward and Decision Making*, chapter 19, pages 399–425. Academic Press, Oxford, UK.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Mayhew, N. W. J. E., Overton, P. G., and Redgrave, P. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, 307:1476–1479.

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

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, 1:30–40.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, 11:410–416.

Doyon, J., Bellec, P., Amsel, R., Penhune, V., Monchi, O., Carrier, J., Lehéricy, S., and Benali, H. (2009). Contributions of the basal ganglia and functionally related brain structures to motor learning. *Behavioural Brain Research*, 199:61–75.

Eckerman, D. A., Hienz, R. D., Stern, S., and Kowlowitz, V. (1980). Shaping the location of a pigeon's peck: Effect of rate and size of shaping steps. *Journal of the Experimental Analysis of Behavior*, 33:299–310.

Ferster, C. B. and Skinner, B. F. (1957). *Schedules of Reinforcement*. Appleton-Century-Crofts, New York, USA.

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

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17:51–72.

Frank, M. J. and Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113:300–326.

Frank, M. J., Seeberger, L. C., and O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, 306:1940–1943.

Gardner, R. (1958). Multiple-choice decision behavior. *American Journal of Psychology*, 71:710–717.

Gläscher, J. P., Daw, N. D., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66:585–595.

Gläscher, J. P. and O'Doherty, J. P. (2010). Model-based approaches to neuroimaging combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1:501–510.

Glimcher, P. W. (2003). *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. MIT Press, Cambridge, Massachusetts, USA.

Glimcher, P. W. and Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science*, 306:447–452.

Gluck, M. A. (2008). Behavioral and neural correlates of error correction in classical conditioning and human category learning. In Gluck, M. A., Anderson, J. R., and Kosslyn, S. M., editors, *Memory and Mind: A Festschrift for Gordon H. Bower*, chapter 18, pages 281–305. Lawrence Earlbaum Associates, New York, NY, USA.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, 14:447–485.

Goodnow, J. T. (1955). Determinants of choice-distribution in two-choice situations. *The American Journal of Psychology*, 68:106–116.

Gormezano, I., Schneiderman, N., Deaux, E. G., and Fuentes, I. (1962). Nictitating membrane: Classical conditioning and extinction in the albino rabbit. *Science*, 138:33–34.

Grafton, S. T. and Hamilton, A. F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, 26:590–616.

Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology*, 15:638–644.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, 31:359–387.

Graybiel, A. M., Aosaki, T., Flahrety, A. W., and Kimura, M. (1994). The basal ganglia and adaptive motor control. *Science*, 265:1826–1831.

Green, L. and Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130:769–792.

Grupen, R. and Huber, M. (2005). A framework for the development of robot behavior. In *2005 AAAI Spring Symposium Series: Developmental Robotics*, Palo Alta, California, USA. American Association for the Advancement of Artificial Intelligence.

Gurney, K. (2009). Reverse engineering the vertebrate brain: Methodological principles for a biologically grounded programme of cognitive modelling. *Cognitive Computation*, 1:29–41.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, 84:401–410.

Gurney, K., Prescott, T. J., Wickens, J. R., and Redgrave, P. (2004). Computational models of the basal ganglia: From robots to membranes. *Trends in Neuroscience*, 27:453–459.

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

Haber, S. N., Kim, K. S., Mailly, P., and Calzavara, R. (2006). Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical inputs, providing a substrate for incentive-based learning. *The Journal of Neuroscience*, 26:8368–8376.

Haruno, M. and Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, 19:1242–1254.

Hazy, T. E., Frank, M. J., and O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine repsonses in learning. *Neuroscience and Biobehavioral Reviews*, 34:701–720.

Hengst, B. (2012). Approaches to hierarchical Reinforcement Learning. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State of the Art*, chapter 9, pages 293–323. Springer-Verlag, Berlin Heidelberg.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4:267–272.

Hikosaka, O. (2007). Basal ganglia mechanisms of reward-oriented eye movement. *Annals of the New York Academy of Science*, 1104:229–249.

Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309.

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

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use 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*, chapter 13, pages 249–270. MIT Press, Cambridge, Massachuetts, USA.

Houk, J. C., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P. J., Roy, S. A., and Simo, L. S. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362:1573–1583.

Houk, J. C. and Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex*, 5:95–110.

Hull, C. L. (1943). *Principles of Behavior*. Appleton-Century-Crofts, New York, USA.

Humphries, M. D. and Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in Neurobiology*, 90:385–417.

Ito, M. and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *The Journal of Neuroscience*, 29:9861–9874.

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

Joel, D. and Weiner, I. (1994). The organization of the basal ganglia-thalamocortical circuits: Open interconnected rather than closed segregated. *Neuroscience*, 63:363–379.

Joshua, M., Adler, A., and Bergman, H. (2009). The dynamics of dopamine in control of motor behavior. *Current Opinion in Neurobiology*, 19:615–620.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and Aversive Behavior*, pages 279–296. Appleton-Century-Crofts, New York, USA.

Kehoe, E. J., Schreurs, B. G., and Graham, P. (1987). Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response. *Animal Learning and Behavior*, 15:455–464.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience*, 29:14701–14712.

Kishida, K. T., King-Casas, B., and Montague, P. R. (2010). Neuroeconomic approaches to mental disorders. *Neuron*, 67:543–554.

Klopf, A. H. (1982). *The Hedonistic Neuron: A Theory of Memory, Learning and Intelligence*. Hemisphere Publishing Corporation, Washington DC, USA.

Kobayashi, S. and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *The Journal of Neuroscience*, 28:7837–7846.

Konidaris, G. D. and Barto, A. G. (2009). Skill discovery in continuous reinforcement learning domains using skill chaining. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pages 1015–1023, Cambridge, Massachusetts, USA. MIT Press.

Lau, B. and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, 58:451–463.

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

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

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, and Behavioral Neuroscience*, 9:343–364.

Maia, T. V. and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurobiological disorders. *Nature Neuroscience*, 14:154–162.

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301:229–232.

Matsuzaka, Y., Picard, N., and Strick, P. (2007). Skill representation in the primary motor cortex after long-term practice. *Journal of Neurophysiology*, 97:1819–1832.

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., and Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends in Neurosciences*, 28:401–407.

Middleton, F. A. and Strick, P. L. (2002). Basal-ganglia"projections" to the prefrontal cortex of the primate. *Cerebral Cortex*, 12:926–35.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202.

Miller, J. D., Sanghera, M. K., and German, D. C. (1981). Mesencephalic dopaminergic unit activity in the behaviorally conditioned rat. *Life Sciences*, 29:1255–1263.

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50:381–425.

Mirolli, M., Mannella, F., and Baldassarre, G. (2010). The roles of the amygdala in the affective regulation of body, brain, and behaviour. *Connection Science*, 22:215–245.

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., Hyman, S. E., and Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, 431:760–767.

Montague, P. R., King-Casas, B., and Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, 29:417–448.

Moore, J. W. and Choi, J. S. (1997). Conditioned response timing and integration in the cerebellum. *Learning and Memory*, 4:116–129.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, 9:1057–1063.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., and Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, 50:631–641.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron*, 41:269–280.

Ng, A., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: theory and applications to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287.

Nicola, S. M. (2007). The nucleus accumbens as part of a basal ganglia action selection circuit. *Psychopharmacology*, 191:521–550.

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

Niv, Y., Daw, N. D., and Dayan, P. (2006a). Choice values. *Nature Neuroscience*, 9:987–988.

Niv, Y., Duff, M. O., and Dayan, P. (2005). Dopamine, uncertainty, and TD learning. *Behavioral and Brain Functions*, 1:6.

Niv, Y., Joel, D., and Dayan, P. (2006b). A normative perspective on motivation. *Trends in Cognitive Sciences*, 10:375–381.

Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *The Journal of Neuroscience*, 30:10692–10702.

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*, 304:452–454.

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

O'Reilly, R. C. and Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18:283–328.

Packard, M. G. and Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, 25:563–593.

Pasupathy, A. and Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433:873–876.

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, Toronto, Ontario, Canada.

Pennartz, C. M., Berke, J. D., Graybiel, A. M., Ito, R., Lansink, C. S., van der Meer, M., Redish, A. D., Smith, K. S., and Voorn, P. (2009). Corticostriatal interactions during learning, memory processing, and decision making. *The Journal of Neuroscience*, 29:12831–12838.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442:1042–1045.

Phelps, E. A. and LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48:175–87.

Poldrack, R. A., Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y., and Knowlton, B. J. (2005). The neural correlates of motor skill automaticity. *The Journal of Neuroscience*, 25:5356–5364.

Pompilio, L. and Kacelnik, A. (2005). State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. *Animal Behaviour*, 70:571–578.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, 7:967–975.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, 58:322–339.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M. R., and Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience*, 11:760–772.

Redish, A. D., Jensen, S., and Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences*, 31:415–487.

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

Richardson, W. K. and Warzak, W. J. (1981). Stimulus stringing by pigeons. *Journal of the Experimental Analysis of Behavior*, 36:267–276.

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*, 10:1615–1624.

Roesch, M. R., Singh, T., Brown, P. L., Mullins, S. E., and Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *The Journal of Neuroscience*, 29:13365–13376.

Samejima, K. and Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, 1104:213–228.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, 310:1337–1340.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *The Journal of Neuroscience*, 23:9913–9923.

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology*, 56:1439–1461.

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

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, 57:8–115.

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, 30:259–288.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions*, 6:24.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience*, 13:900–913.

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

Schultz, W., Tremblay, L., and Hollerman, J. R. (2003). Changes in behavior-related neuronal activity in the striatum during learning. *Trends in Neuroscience*, 26:321–328.

Seger, C. A. and Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, 33:203–219.

Selfridge, O. J., Sutton, R. S., and Barto, A. G. (1985). Training and tracking in robotics. In Joshi, A., editor, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 670–672, San Mateo, CA, USA. Morgan Kaufmann.

Shah, A. (2008). *Biologically-based functional mechanisms of motor skill acquisition.* PhD thesis, University of Massachusetts Amherst.

Shah, A. and Barto, A. G. (2009). Effect on movement selection of an evolving sensory representation: A multiple controller model of skill acquisition. *Brain Research*, 1299:55–73.

Shanks, D. R., Tunney, R. J., and McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, 15:233–250.

Siegel, S. and Goldstein, D. A. (1959). Decision making behaviour in a two-choice uncertain outcome situation. *Journal of Experimental Psychology*, 57:37–42.

Skinner, B. F. (1938). *The Behavior of Organisms*. Appleton-Century-Crofts, New York, USA.

Staddon, J. E. R. and Cerutti, D. T. (2003). Operant behavior. *Annual Review of Psychology*, 54:115–144.

Sutton, R. S. (1988). Learning to predict by methods of temporal differences. *Machine Learning*, 3:9–44.

Sutton, R. S. and Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170.

Sutton, R. S. and Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pages 355–378.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts, USA.

Tanji, J. and Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological Reviews*, 88:37–57.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Macmillan, New York, USA.

Tindell, A. J., Berridge, K. C., Zhang, J., Pecina, S., and Aldridge, J. W. (2005). Ventral pallidal neurons code incentive motivation: Amplification by mesolimbic sensitization and amphetamine. *European Journal of Neuroscience*, 22:2617–2634.

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

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

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

Tolman, E. C. (1949). There is more than one kind of learning. *Psychological Review*, 56:44–55.

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

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience*, 30:31–56.

Watson, J. B. (1914). *Behavior: An Introduction to Comparative Psychology*. Holt, New York, USA.

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behavioural Brain Research*, 199:119–128.

Wickens, J. R., Budd, C. S., Hyland, B. I., and Arbuthnott, G. W. (2007). Striatal contributions to reward and decision making. Making sense of regional variations in a reiterated processing matrix. *Annals of the New York Academy of Sciences*, 1104:192–212.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 WESCON Convention Record Part IV, New York: Institute of Radio Engineers*, pages 96–104.

Wilson, C. J. (2004). Basal ganglia. In Shepherd, G. M., editor, *The Synaptic Organization of the Brain*, chapter 9, pages 361–414. Oxford University Press, Oxford, United Kingdom, 5 edition.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5:483–494.

Wolpert, D. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, 27:511–524.

Wörgötter, F. and Porr, B. (2005). Temporal sequence learning, prediction, and control: A review of different models and their relation to biological mechanisms. *Neural Computation*, 17:245–319.

Wrase, J., Kahnt, T., Schlagenhauf, F., Beck, A., Cohen, M. X., Knutson, B., and Heinz, A. (2007). Different neural systems adjust motor behavior in response to reward and punishment. *NeuroImage*, 36:1253–1262.

Wyvell, C. L. and Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: Enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *Journal of Neuroscience*, 20:8122–8130.

Yin, H. H., Ostlund, S. B., and Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, 28:1437–1448.

Yu, A. and Dayan, P. (2005). Uncertainty, neuromodulation and attention. *Neuron*, 46:681–692.