A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives

Elliot A. Ludvig, Marc G. Bellemare, & Keir G. Pearson

University of Alberta

Abstract

In the last 15 years, there has been a flourishing of research into the neural basis of reinforcement learning, drawing together insights and findings from psychology, computer science, and neuroscience. This remarkable confluence of three fields has yielded a growing framework that begins to explain how animals and humans learn to make decisions in real time. Mastering the literature in this sub-field can be quite daunting as this task can require mastery of at least three different disciplines, each with its own jargon, perspectives, and shared background knowledge. In this chapter, we attempt to make this fascinating line of research more accessible to researchers in any of the constitutive sub-disciplines. To this end, we develop a primer for reinforcement learning in the brain that lays out in plain language many of the key ideas and concepts that underpin research in this area. This primer is embedded in a literature review that aims not to be comprehensive, but rather representative of the types of questions and answers that have arisen in the quest to understand reinforcement learning and its neural substrates. Drawing on the basic findings in this research enterprise, we conclude with some speculations about how these developments in computational neuroscience may influence future developments in Artificial Intelligence.

Keywords: Reinforcement Learning, Classical Conditioning, Operant Conditioning, Dopamine, Temporal-difference (TD) algorithm, Striatum, Reward, Neuroeconomics

A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives.

The last decade has seen a proliferation of research exploring the neural and psychological mechanisms of reinforcement learning (for some good reviews and perspectives, see Dayan & Daw, 2008; Doya, 2007; Maia, 2009; Niv, 2009; Rangel, Camerer, & Montague, 2008; Schultz, 2002, 2007). This flourishing area of computational neuroscience draws on the expertise and knowledge in many sub-disciplines, including psychology, neuroscience, computer science, philosophy, and economics, amongst others. This remarkable confluence of fields was catalyzed by the discovery of a close correspondence between the behaviour of dopamine neurons in classical conditioning tasks and the prediction error in the temporal-difference (TD) algorithm from reinforcement learning (Montague, Dayan & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Sutton, 1988; Sutton & Barto, 1990; see Figure 5). The import of this finding has filtered outward from a strikingly successful model of the neural basis of a simple conditioning behavior in animals to theoretical models of human economic decision making and, in part, to an entire field of neuroeconomics (e.g., Glimcher et al., 2009; Platt & Huettel, 2008; Rangel et al., 2008; Schultz, 2009).

Our goal in this chapter is two-fold. First, we aim to provide a primer of basic introductory materials in three of the constitutive disciplines of this enterprise—psychology, computer science, and neuroscience—to facilitate access by Artificial Intelligence (AI) researchers and other computational neuroscientists into this exciting field. As our second goal, we will not directly re-tread the ground covered in detail by the many comprehensive recent reviews, but rather we use some selective examples of reinforcement-learning research and show how this multi-disciplinary enterprise has helped inform and been informed by these basic lines

of inquiry.

In considering the relationship between observed behaviour, computational models, and neural mechanisms, Marr's (1982) three levels of analysis prove very instructive. Marr proposed that any information-processing system can be analyzed at three different levels: the computational or functional, the algorithmic or representational, and the implementational. At the computational level, one specifies the goals and objectives of the system. What does the system do? For example, the computational goal for classical conditioning might be the prediction of important biological events. Second, at the algorithmic level, one specifies the step-by-step procedure by which this function is accomplished. What algorithm or procedure does the system use to accomplish the computational goals? Again, for classical conditioning, this might be the Rescorla-Wagner rule (Rescorla & Wagner, 1972) or the TD algorithm (Sutton & Barto, 1990) or any other set of rules that describe how the computation happens. Finally, at the implementational level, the important details of how these different algorithms and representations can be instantiated in neural tissue or other mediums are laid out. How are these algorithms physically realized? One example would be the equating of the reward-prediction error from reinforcement learning with the burst firing of dopamine neurons (Schultz et al., 1997). A full explanation of any information-processing system would require adequate accounts at each of the three levels of analysis.

Our chapter follows Marr's proposal by dividing this introduction to neural reinforcement learning into three sections that roughly correspond to his three levels of analysis. Section 1 describes the computational problems facing creatures in simple learning and decision-making situations and summarizes some of the attempts within psychology to characterize the algorithms that may be used by animals to solve these problems. Section 2 introduces the core

computational ideas of reinforcement learning (RL) and shows how ideas from RL are particularly well suited as potential algorithms for models of natural learning. Section 3 ties the first two sections together in the brain, by introducing the neurobiological evidence that different RL algorithms have strong neurophysiological correlates. Our general strategy in each of the three sections is to introduce some of the fundamental problems and terminology in the sub-field before detailing the role and contribution of RL.

1. The Psychology of Learning and Decision Making

The first step towards understanding the modern study of reinforcement learning in the brain is a basic grasp of the major behavioral phenomena within the realms of animal learning and behavioral economics. We start by discussing two simple forms of learning—classical and operant conditioning—before progressing to more complex value-based decision making in humans and animals. In each case, we attempt to illustrate the major empirical phenomena and the functional goals for these behaviors, in addition to touching on some of the explanations that have been proffered within psychology to deal with these data. Our goal in this section is to lay out the behavioral puzzles for which the RL algorithms discussed in Section 2 provide a potential computational mechanism.

1.1. Predicting the Future: Classical Conditioning

In a recent episode of the American TV show *The Office*, the protagonist Tim decides to play a prank on his co-worker Dwight. Repeatedly, each time Tim shuts down his computer (making a tell-tale beep), he offers Dwight a small candy. After several iterations, Dwight begins to stick his hand out for the candy immediately upon hearing the beep, even claiming a bad taste in his mouth on a "trial" when Tim fails to present the candy. This little fictional snippet evokes (and was inspired by) the classic work of the Russian physiologist Ivan Pavlov, who spent many

years studying the salivary reactions of dogs to various sounds that were reliably followed by food delivery (Pavlov, 1927). This simple form of associative learning, known as classical or Pavlovian conditioning, is widely exhibited in the natural world, spanning many species from insects to fish to dogs to humans (for some good reviews and perspectives, see Domjan, 2005; Pearce & Bouton, 2001; Rescorla, 1988).

More precisely, classical conditioning is said to occur whenever a previously neutral stimulus (the CS or conditioned stimulus), such as the beep, is paired with a rewarding stimulus (the US or unconditioned stimulus). This reward can either be positive, such as the candy in the *Office* episode, or aversive, such as an electric shock or puff of air to the eye. Initially, only the reward elicits a response, such as reaching out a hand or salivation, but after sufficient training, the CS will also elicit a conditioned response (CR).

Most early views of classical conditioning proposed that animals learn an association between the CS and reward solely due to temporal contiguity (e.g., Guthrie, 1930; Pavlov, 1927; but really back to Aristotle). This simple idea entailed that whenever the CS and reward occurred around the same time, an association formed between them—the simple co-occurrence of the beep and candy was sufficient for Dwight to learn a link between them. Three major empirical findings from the animal learning literature in the late 1960's helped upend this contiguity-centered point-of-view: blocking, contingency effects, and conditioned taste aversion.

The first finding, blocking, showed that stimuli that perfectly predict reward do not always elicit conditioning responding. Only if the reward is unpredicted or surprising does learning occur. In the blocking procedure, a CS is paired with reward until the association is well learned. At this point, a second CS is introduced, and both CSs are now paired with the reward. Typically, the newly introduced CS, when presented alone, does not elicit conditioned

responding, even after substantial training with both CSs and the reward. It is as though the pre-trained cue "blocks" any learning from happening to the newly introduced cue, despite the temporal contiguity of this cue and reward (Kamin, 1969; Waelti et al., 2001). As a real-life example, imagine your friend has a known peanut allergy, and she experiences an allergic reaction after eating shrimp satay with peanut sauce at a Thai restaurant. Your friend may also be allergic to the shellfish, but from the restaurant incident, you would not make this connection because the potential association between shellfish and the allergic reaction was "blocked" by the known allergy to peanuts.

Second, Rescorla (1968) found in a series of experiments that a contingency or predictive relationship is crucial for establishing an association between two stimuli. As opposed to contiguity, which only requires that two events occur at the same time, contingency requires that the predicted stimulus (US) be more probable during the CS than at other times. For example, Rescorla found that inserting extra rewards into the experiment when the CS was not present (i.e., during the inter-trial intervals) eliminated responding to that cue (but the temporal relations might matter; see Williams et al., 2008). Only when the CS predicted a reliably higher rate of reward than the background rate did conditioned responding emerge.

These two experiments—blocking and contingency effects—demonstrated how temporal contiguity by itself was insufficient for classical conditioning; some form of contingency or predictive relationship was needed for conditioning to occur. Finally, a third set of experiments firmly established that temporal contiguity was not even necessary for conditioning. In these *conditioned taste aversion* experiments, rats were presented with flavoured water and then made ill several hours later through a dose of radiation (e.g., Garcia & Koelling, 1966). Rats subsequently avoided drinking this flavoured water, even though there was no temporal

contiguity between the cue and reward, as several hours intervened between the initial drinking session and the illness. The key ingredient for successful conditioning in this case was that a valid predictive relationship (contingency) existed between the water and the illness, even when there was no temporal contiguity whatsoever.

This perspective that surprise and contingency, rather than contiguity, are the most important factors for conditioning found its most succinct expression in what has become known as the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972). In the RW model, learning occurs whenever the reward exceeds expectations. Figure 1A depicts how learning occurs in the RW model for a negative shock reward. There is an initial reward prediction (read: associative strength) on a given trial, followed by the actual outcome. At the end of the trial, this outcome is compared with the prediction, and the difference between the two, the *reward-prediction error*, is used to improve the prediction for next time. This simple learning rule describes precisely how the associative strength or reward prediction V for each CS present in a trial changes as a result of experience:

$$V = V + \alpha\,[r - V_{\text{Sum}}] \tag{1}$$

where r is the reward on that trial, V_{Sum} is the net associative strength based on all available CSs, and α is a parameter that controls the speed of learning. According to the RW model, there is an increment (or decrement) in the strength of an association based on the discrepancy between the reward received and the expected reward on a given trial. The expected reward (i.e., the net associative strength) is derived from all CSs present on a given trial as a simple sum of their associative strengths with the reward in question. The reward-prediction error drives all learning, leading to the experimental prediction that associative learning should only occur when expectations about rewards are violated. As a result, blocking is quite naturally explained. When

the second cue is introduced, the first cue already perfectly predicts the reward, thus no reward-prediction error occurs on that trial, and no new learning occurs to the second cue.

The RW model has been remarkably successful in explaining and predicting many phenomena in animal and human conditioning (Rescorla & Wagner, 1972; Miller, 1996), but there are numerous empirical difficulties and theoretical alternatives (e.g., Gallistel & Gibbon, 2000; Pearce & Hall, 1980; Sutton & Barto, 1981, 1990; Wagner, 1981). Miller et al. (1996) compiled an extensive list of these empirical challenges to the RW model, and many of these challenges also extend to newer extensions of this error-correcting learning rule, such as those from reinforcement learning (see below). Beyond these empirical concerns, one major conceptual problem that confronts the Rescorla-Wagner model is that the model is not real time and relies quite heavily on the concept of a trial, which is not immediately apparent in the experience of an animal (see Gallistel & Gibbon, 2000; Sutton & Barto, 1981; 1990). In fact, the relative times of the stimulus-reward and inter-trial intervals might be the most important determinants of the speed of learning, rather than trials themselves (e.g., Gottlieb, 2008). In addition, animals can learn to predict stimuli other than rewards (e.g., Brogden, 1939; Ludvig & Koop, 2008; Rescorla, 1980), perhaps even taking into account the causal structure of the environment (e.g., Blasidell et al., 2006; Dwyer, Starns, & Honey, 2009). These simple learning phenomena lie beyond the explanatory scope of the RW model.

The basic idea of error-driven learning has also been adopted into many of the learning rules that characterize modern RL (see Sutton & Barto, 1990, 1998). For example, in temporal-difference (TD) learning (see also Section 2.2), the discrepancy between prediction and outcome is also used to drive the value function or the prediction of future rewards on a moment-to-moment basis. This value function is roughly equivalent to the net associative strength (V_{Sum}) in

the Rescorla-Wagner model. Figure 1B depicts how learning proceeds on a single time step in the TD algorithm. The learning proceeds along quite similar lines to the RW model, but with one major difference: Instead of comparing the reward received with the reward predicted on a given trial, on every time step *t*, the reward received is compared with *the change* in reward prediction to generate a reward-prediction error (δ):

$$\delta(t) = r(t) - [V(t\text{-}1) - V(t)]. \quad (2)$$

With some easy algebra, this equation can be written to make it clear that the TD or reward-prediction error reflects how much better the world is at this time step (current reward plus newly predicted upcoming reward) versus what it was expected to be (the old prediction of reward):

$$\delta(t) = [r(t) + V(t)] - V(t\text{-}1). \quad (3)$$

This error can then be used to update the old reward prediction to bring it more in line with what was experienced by adding a portion of the error to that old reward prediction:

$$V(t\text{-}1) = V(t\text{-}1) + \alpha\, \delta(t), \quad (4)$$

where α is a parameter that controls the speed of learning. What the error does is change the value function or the way that predictions about rewards are made based on the stimuli. This updated value function can now be used to make a new prediction about the upcoming reward for the next time step.

As a result of this real-time updating, TD learning can make moment-to-moment predictions about reward and does not operate solely at the trial level, leading to improved correspondence with numerous conditioning behaviors in models based on this learning rule (e.g., Sutton & Barto, 1981, 1990; Ludvig et al., 2009). In addition to a better fit with some conditioning data, what is probably most compelling about this alternative learning rule as a model of conditioning is the strong correspondence between the error term and the behaviour of

dopamine neurons in the midbrain (e.g., Montague et al., 1996; Schultz et al., 1997; see Fig. 5). Section 3.3 will discuss these correspondences in detail.

1.2. Controlling the Future: Operant Conditioning

Classical conditioning is restricted in scope because most conditioned responses already exist as reflexive reactions to the rewarding or conditioned stimuli (but see Domjan, 2005). What classical conditioning does is tune when and how strongly animals perform these reactive responses. In contrast, novel responses that are shaped and reinforced by rewards from the environment are not possible within the classical conditioning framework. Indeed, the RW model of classical conditioning ignores responding altogether, providing a model for how associative strength changes over time, but leaving out the important issue of how this associative strength might get translated into behavior (see Stout & Miller, 2007; Ludvig et al., 2009). Operant or instrumental conditioning deals directly with how animals learn to make potentially novel responses that yield rewarding outcomes.

Perhaps the first example of operant conditioning in modern experimental psychology was Thorndike's puzzle box for cats (Thorndike, 1911). This experimental chamber was a small enclosure from which cats could escape given the right sequence of actions. In different permutations of the box, they could either pull a chain, or push a bar, or step on a latch to escape. After repeated exposure to this puzzle box, the cats gradually learned to perform the appropriate actions and escape more and more quickly. This gradual, trial-and-error learning of new actions typifies many of the procedures in the modern study of operant conditioning (see Staddon & Cerutti, 2003, for a review).

In classical conditioning, animals learn to predict the US on the basis of the CS, or, in more traditional psychological terms, the animals learn an association between the CS and the

US. In operant conditioning, when a new response is performed, which predictive relationships animals learn are not nearly as obvious. At least two possibilities present themselves: Animals might learn a link between the stimulus and response (S-R association) or between a response and outcome (R-O association). In Thorndike's puzzle box, the cat may have learned (1) to step on the latch in the box or (2) that stepping on the latch leads to escape. This three-way connection between stimulus, response, and outcome is known as the *three-term contingency* (Skinner, 1938). Initial investigations of operant conditioning tended to focus most strongly on the S-R association. For example, on the basis of the cat puzzle boxes, Thorndike (1911) formulated his famous *Law of Effect:*

> "Of several responses made to the same situation, those which are accompanied by 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..." (p. 244)

Here, he clearly pegs operant conditioning as learning a connection between a situation (stimulus) and response. RL has also mostly adopted this convention, wherein agents try to learn a policy—a mapping from stimuli to responses—on the basis of the previous reward history (see Section 2.3).

More recent work has drawn into question to what degree these simple stimulus-response associations drive behaviour in operant conditioning experiments (Balleine & Dickinson, 1998; Daw, Niv, & Dayan, 2005; Dickinson & Balleine, 1994; for a recent review, see Balleine & O'Doherty, 2009). An important distinction between habitual (S-R) and goal-directed (R-O) systems has been proposed, with separate neural substrates for each. One empirical example in support of this distinction comes from reinforcer devaluation experiments

(e.g., Adams & Dickinson, 1981; Rescorla & Colwill, 1985). In these experiments, animals are trained to perform a particular response, say press a lever, for a food reward. After training, the food reward is devalued typically by either satiating the animal with the food or poisoning the animal following consumption of the food, but in a different context so as not to contaminate the previous training. When those animals are brought back to the original experimental set-up, they press the lever for the now-devalued food less than a comparable group of animals that had a different food reward devalued in the interim.

In loose terms, this decrease in responding indicates that animals know which food reward is upcoming and are not solely reacting reflexively to the stimulus. In more technical terms, these animals are sensitive to the association between response and outcome (R-O), and not only between stimulus and response (S-R). Responding, however, does not entirely disappear in these situations, suggesting that a S-R connection does exist and persist. In addition, this decrease in responding following devaluation depends on the amount of training given to the animals: Highly overtrained animals become insensitive to the devaluation manipulation and continue to press the lever even afterward. Thus, there is evidence for both habitual (S-R) and goal-directed (R-O) responses. From these experiments, it would seem that Thorndike's cats may have learned both associations: to step on the latch in the box and that latch-stepping leads to escape.

For animals to turn this learning into the effective control of future outcomes requires a solution to two significant problems: how much and when to respond? These questions about the rate and timing of learned responding have dominated much of the literature on operant conditioning. The primary tactic for asking and answering these questions empirically has been through the two major schedules of reinforcement: ratio and interval schedules (Ferster &

Skinner, 1957). In a ratio schedule, animals are rewarded after a certain number of responses are emitted; in an interval schedule, animals are rewarded for the first response after a certain amount of time has elapsed. Both ratio and interval schedules come in fixed or variable varieties; in a fixed schedule, the number of responses or time to reward is always the same, whereas in a variable schedule, only the average number or time is specified. Much of the theoretical work in this area has focused on steady-state behaviour—what the animal does after the course of learning is complete (e.g., Gibbon, 1977; Herrnstein, 1961). As with the Rescorla-Wagner model of classical conditioning, these real-time limitations to many models of operant conditioning suggest an opening for future theoretical contributions. Recent computational models based on RL have begun to make in-roads on both these problems with new models of both response vigor (Niv et al., 2005) and response timing (Daw, Courville, & Touretsky, 2006; Ludvig et al., 2008, 2009).

1.3. Evaluating the Future: Choice

Imagine that you are on the game show *Deal or No Deal* and faced with the choice of taking the offer from the banker for a guaranteed $100,000 and going home or continuing onward in the game and gambling for a 50/50 chance between two briefcases with either $1 or $250,000 in them. How would you decide what to do? You might be "rational" and figure out that the *expected value* of the second, risky option is ~$125,000 (50% of $250K), which is higher than the expected value of the safe option and decide to gamble. As it turns out, most people faced with this choice would take the less-"valuable" safe option, acting risk averse, and walk away with the guaranteed $100K (e.g., Kahneman & Tversky, 1979). This question of how people and animals value different outcomes and make decisions between them has been the purview of behavioral economics (Camerer & Loewenstein, 2003; see also Ariely, 2008) and,

more recently, as questions about brain mechanisms have come to the fore, of neuroeconomics (e.g., Glimcher, 2009; Platt & Huettel, 2007; Trepel et al., 2005).

Why do people tend to undervalue the gamble and play it safe? One possible answer comes from *prospect theory* (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981; but back to Bernoulli, 1738), which proposes that people make choices based on the *expected utility* of an outcome rather than the expected value. In this context, the utility can be thought of as the subjective value—how much the $250,000 is worth to the decision maker, rather than what its objective value is. Prospect theory contends that the relationship between objective value and subjective utility is sub-linear for gains: Winning $200 is less than twice as good as winning $100. As a result, people tend to choose the safer option as opposed to an objectively equivalent, but riskier option; they are *risk averse* for gains.

The converse result, however, is seen when the decision involves a sure small loss (losing $100) vs. the chance of a big loss (50/50 chance of losing $200). In this instance, people tend to choose the gamble, making them *risk prone* or *risk seeking* for losses. Prospect theory proposes a similar non-linearity for the negative utility curve: Losing $200 is less than twice as bad as losing $100. As a result, someone trying to minimize their subjective loss would take the gamble. Within prospect theory, this asymmetry between risk sensitivity for decisions about gains and losses means that the way a question is framed or anchored can have a great influence on how people make choices about different potential economic outcomes (e.g., Tversky & Kahneman, 1981).

Animals, too, show varying risk sensitivity profiles based on the types of choices with which they are faced (e.g., Bateson & Kacelnik, 1995; Shafir, 2000). For example, Bateson and Kacelnik (1995) found that starlings, like humans, were risk averse when the reward was the

amount of food. In contrast, when these birds were tested with delays to food, the starlings were risk seeking, preferring variable delays to food over fixed delays. In general, as might be imagined, shorter delays to reward are preferred to longer delays to reward, the result of a phenomenon known as *temporal discounting* (e.g., Green & Myerson, 2004). In this instance, the asymmetry between amounts and delays may be explained by the increase in variance that goes along with estimating larger magnitudes. Larger amounts are good, but larger delays are bad. As a result, the good amounts have more variance in their estimate, but the bad delays have more variance in their estimate. Simply sampling from this remembered distribution of amounts or delays produces this asymmetry in risk sensitivity (e.g., Marsh & Kacelnik, 2002). This risk sensitivity in birds also manifests itself in more naturalistic conditions. In one series of experiments, dark-eyed juncos, a small bird, chose the safe, small food option when the external temperature was warm, but in cold conditions, when they needed a larger meal to survive through the night, the juncos were risk seeking and sought the larger, more variable food source (Caraco, 1981; Caraco et al., 1990). Thus, though variations in risk sensitivity may reflect seemingly sub-optimal non-linearities in subjective utilities, these variations may still have strong adaptive value in an ecological context.

The preferences in the various choice situations described above are typically not pure preferences, but rather a tendency towards picking one option or another. In fact, when animals and humans are confronted with repeated options to which they can allocate varying portions of behaviors, they often show a distinct regularity known as *matching* behaviour (Herrnstein, 1961, 1970; Davison & McCarthy, 1988). In matching behaviour, the degree of preference for different options depends on the rates of reward for those options. So, if a monkey can press Button A and get on average 2 candies or press Button B and get on average 4 candies, the monkey will tend to

press Button B twice as often. Numerous mechanisms have been proposed to explain how animals achieve this matching behaviour (e.g., Jozefowiez et al., 2009; McDowell, 2004; Simen & Cohen, 2009; Sugrue et al., 2005), but only recently have the potential connections to RL models begun to be evaluated (e.g., Lau & Glimcher, 2005; Loewenstein, Prelec, & Seung, 2009; Sakai & Futai, 2008).

In this section, we have reviewed some of the major findings in the psychology of learning and decision making in animals and humans. These represent many of the core behavioral phenomena that reinforcement-learning models attempt to explain. Most of the theoretical work has focused on the simpler learning phenomena of classical conditioning (e.g., Sutton & Barto, 1990; Schultz et al., 1997), but more recent work has made headway on operant conditioning and even more complex decision making (Gureckis & Love, 2009; Niv et al., 2005; Wunderlich, Rangel, & O'Doherty, 2009). One of the challenges for RL researchers in the future will be how to reconcile the simple learning rules that guide performance in classical- and operant-conditioning tasks with the more complex decision making exhibited by humans and animals in behavioral-economics settings.

2. Algorithms for Reinforcement Learning

Reinforcement learning (RL) is a branch of AI that is concerned with the computational study of real-time decision making (Sutton & Barto, 1998). In RL, agents are assumed to interact with an environment while attempting to maximize a reward signal (see Figure 2A). In biological terms, these agents can be conceived as entire organisms or, occasionally, as control centers in the brain that receive filtered input from the external environment. For a rigorous and accessible introduction to RL, see the book by Sutton and Barto (1998). Here, we first introduce the formalisms and goals of RL in the context of broader work in machine learning and then step

through some of the key concepts in the area, including value functions and the temporal-difference (TD) learning algorithm. We conclude with a discussion of some of the key issues in the design and use of RL algorithms, including strategies for action selection and efficient exploration.

2.1. Machine Predictions: Supervised Learning

Although the idea of learning is familiar to young children, the computational mechanisms that drive learning remain largely unknown. With the *Dartmouth Summer Research Conference on Artificial Intelligence*, held in 1956 and organized by the pioneers of the field, computer scientists began to consider how to describe intelligent concepts in machine terms. They conjectured that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1955/2006, p. 12). Thus the field of AI was born—in a quest to discover and understand algorithms that exhibit intelligent behavior. Paramount to the study of AI is learning, which is often investigated in three guises: supervised learning, unsupervised learning, and reinforcement learning (Mitchell, 1997; Sutton & Barto, 1998). We first turn our attention to supervised learning—a framework for describing how machines can learn to predict the future from data—whose limitations will provide computational motivation for the techniques of reinforcement learning.

Supervised learning: Machines making predictions. From a mathematical perspective, learning is about finding input-output mappings that are consistent with some dataset. In supervised learning, this data is usually provided as a set of inputs and desired outputs; these desired outputs act as a supervisory signal, telling the algorithm what it should learn about the data. For example, a popular supervised-learning problem is handwritten-digit recognition (e.g.,

LeCun et al., 1998). In this problem, the inputs are grainy images of handwritten digits, and the desired outputs are corresponding labels between 0 and 9. The goal for supervised-learning algorithms is to learn a mapping from a set of already-labeled images, so that the algorithm can correctly identify the digits in novel images.

One popular technique for solving supervised-learning problems is the family of gradient-descent algorithms. These algorithms are highly analogous to the Rescorla-Wagner model (see Section 2.1), wherein learning occurs in response to prediction errors. These algorithms operate by examining a group of inputs (stimuli) and making a prediction about the corresponding output. The prediction is computed through a set of weights that act in a similar fashion to the associative strengths present in the Rescorla-Wagner model. A prediction error is calculated, which is simply the difference between the prediction and the output. If this error is positive, then the prediction gets adjusted upwards; if the error is negative, then the prediction gets adjusted downwards. In the digit-recognition example, the algorithm might learn the probability that an image corresponds with each digit. If the algorithm predicts the correct label with less than 100% probability, there would be a positive error, and the algorithm would predict the correct label for this image next time with a higher probability. If the algorithm predicts the wrong labels, there would be a negative error, and those predictions would be downgraded. These algorithms are known as gradient-descent algorithms because they adjust their prediction by looking at the *gradient* (slope and direction) of the error.

In this supervised-learning framework, it is generally assumed that the learning algorithms attempt to solve one-shot prediction problems. These ideas have proved enormously useful on a wide variety of problems, from bioinformatics to health care (e.g., Asgarian et al., 2009; Cooper et al., 2005). Animals and humans in the real world, however, face a constant

stream of information, which can only be coarsely represented as a one-shot prediction problem. One major element is missing in the supervised-learning formulation of the prediction problem: time.

2.2. Temporal Predictions: Reinforcement Learning

A major limitation of supervised-learning methods is that they ignore the temporal aspects of decision-making problems. For a supervised learner, there is some data from which a prediction is made. The consequences of that prediction are limited to the congruence with the output; future success is not directly compromised by one bad prediction/decision. In contrast, in real-world systems, both natural and artificial, time plays a crucial role: Every decision made by an agent affects all possible future decisions. Consider the task of going out for lunch at work. Your first choice might involve selecting your lunch mates. The second choice determines the restaurant where you will eat. Pending which restaurant you choose, you would be faced with different menus and thus different options. After ordering, you may then decide whether to eat with your hands, utensils, or chopsticks. This inherent sequentiality of real-world decision making necessitates a different set of learning methods for predicting and behaving in a real-time setting. RL addresses exactly this set of questions.

Figure 2A shows how, in RL, the world is typically divided into an interacting agent and environment. The agent receives observations (stimuli) from the world and emits actions. In biological terms, the agent can be thought of as the whole animal, a small control center in the brain, or even an extended cognitive apparatus (e.g., Clark, 2008). A common assumption in most RL problems is that the environment or outside world consists of different states, and given that state, the future is entirely independent of the history before that state. This assumption, known as the *Markov* property, ensures that knowledge of the state is sufficient information for

predicting anything (rewards or otherwise) that can be known about the future—no other information can help. An entire RL problem can be fully described as a *Markov Decision Process* (MDP), which consists of the set of environmental states, the possible actions available to the agent, the reward function, and the transition function that details how the environment changes from state to state. For RL researchers, the usual task is to develop an algorithm for an agent that best picks actions so as to maximize future rewards in such an MDP. In doing so, useful sub-problems can include learning to predict the future rewards, exploring the environment successfully, inferring the state from given observations, or building a model of the environmental state transitions.

An illustrative example should help make these concepts clearer. Figure 2B shows a schematic of a rat in a fairly simple maze. There are multiple choice points for this rat, with the possibility of cheese or water rewards at some end points, while a big cat awaits in another corner. The reward values for each end point (i.e., the reward function) are presented in blue boxes, and some transition probabilities (p) are presented in the orange boxes. Each decision the rat makes influences possible future decisions. If the rat goes up from the start state S1, then it is faced with the prospect of possibly meeting the cat, getting a big hunk of cheese, or returning back to the start. Describing this problem as an MDP involves detailing the 4 states (the choice points), the 4 possible actions (up, down, left, right), the reward function (0, except where indicated), and the transition probabilities (the probability of each action succeeding). From the RL point-of-view, this abstract description captures the whole problem (see Figure 2C). The research question becomes: How do you learn to maximize rewards in this context?

Predicting Rewards: Value functions. A *value function* is the prediction of future rewards from the different states, and is a fundamental tool in RL for solving MDPs. The value of being

in an environmental state (s) at a particular time (t) can be defined as the weighted sum of all future rewards:

$$V(s_t) = r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \gamma^3\, r_{t+4} \ldots , \tag{5}$$

where t denotes the current time step and $t+1$, $t+2$... are future time steps, making r_{t+1} the reward following the current state, r_{t+2} the reward after the following state, and so on. The parameter γ is a discount factor that (when less than 1) causes distant rewards to matter less than immediate rewards in determining the value of a state. At the extreme, with a discount factor of 0, the value of a state is exactly equal to the immediately ensuing reward. The value is thus the sum of all future rewards from a given state, appropriately discounted. Of course, this "true" value is never directly available to the agent, but must somehow be estimated based on the agent's experience. Approximating this idealized value function can be thought of as the goal for all of RL.

One interesting relationship emerges if we compare the values at successive states. The values of two consecutive states (s_t and s_{t+1}) are respectively equal to:

$$\begin{aligned} V(s_t) &= r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \gamma^3\, r_{t+4} \ldots \\ V(s_{t+1}) &= \quad\; r_{t+2} + \gamma\, r_{t+3} + \gamma^2\, r_{t+4} \ldots \end{aligned} \tag{6}$$

The right half of Equation 6 is intentionally shifted to the right to highlight the similarities between the two equations: the value of the second state (s_{t+1}) is almost the same as the value of the first state (s_t), save the first reward (r_{t+1}) and the degree of discounting. Combining the two equations, we get the following relationship between the value of a state, and the value of its successor:

$$V(s_t) = r_{t+1} + \gamma\, V(s_{t+1}). \tag{7}$$

Because of the Markov property, the value of a state can be fully expressed as the next reward

plus the discounted value of the following state. This relationship will prove important in calculating value functions through the temporal-difference (TD) algorithm below.

Let us return to the example of the rat in the maze for a moment. Figure 2C provides an alternate view of the maze as a tree of possible paths. The circles represent the different states, and the squares represent the rewarding outcomes following different choices. For illustrative purposes, we assume that the rat has had some prior experience with the maze: Whenever the rat was in the top state (S2), the rat ended up meeting the cat with probability (*p*) .1 (i.e., 10% of the time) and finding the big cheese with probability .9 (i.e., 90% of the time). As a result, the estimated value of that top state is -.5, which is equal to -5 (i.e., .1 times the -50 reward for meeting the cat) plus +4.5 (i.e., .9 times the +5 reward for getting the big cheese). Similar calculations can be conducted for the left (value = +1) and right states (value = +2.9), and the results are displayed inside the circles in Figure 2C, representing states. For the value of the start state (S1), the calculation is a little more interesting. There are no immediate rewards following any of the actions from this state. Instead, all actions take the rat to another state; however, the value of these potential future states is known. We can therefore calculate the value of S1 from the values of the three immediately succeeding states. If the agent went up, left, or right with equal probability from S1, the value of this state becomes $(1 + 2.9 - 0.5)/3 \approx 1.1$. Note that this example implicitly assumes a discount factor of 1.

Value functions are closely related to the associative strengths present in the Rescorla-Wagner model. Associative strengths can be thought of the prediction of the upcoming US or reward. Value functions are similarly predictions of upcoming rewards, not only of the immediately ensuing reward, but a function of many future rewards. This subtle difference between associative strength and value functions has a distinct analogy to the difference between

supervised learning for prediction and reinforcement learning. In one case, the target is a timeless entity, and in the latter case, the target is a time-embedded set of future outcomes.

Learning Values: The temporal-difference (TD) algorithm. Pending what information is available to the agent, there are many methods for estimating value functions. For example, if the agent has a model of the environment and therefore knows what the next state will be, then the value function can be computed directly through *dynamic programming* methods (Bertsekas & Tsitsiklis, 1996). When agents do not have such a model of the environment, they must somehow estimate the value function from their stream of experience. The temporal-difference (TD) learning algorithm is a procedure for learning these reward predictions in an efficient manner (Sutton, 1988). The key idea behind the TD algorithm is *bootstrapping*: learning a guess from a guess. The agent improves its estimate for the value of a state by learning from the value of the next state. This incremental improvement capitalizes on the key relationship between the values of successive states (see Eq. 7 and Fig. 2B): The value of a given state depends only on the immediate reward and the value of the next state.

The TD-learning algorithm provides a very simple and elegant way of learning to predict future rewards. The algorithm works in a similar fashion to the Rescorla-Wagner model (see Section 2.2) by learning through a reward-prediction error or TD error. In this case, the reward-prediction error (δ) is the discrepancy between what was expected (the old value) and what actually occurred (a reward plus the new value):

$$\delta_t = [r_t + \gamma V(s_t)] - V(s_{t-1}). \quad (8)$$

TD learning then updates the estimate for the value of the last state based on the TD error and a parameter α that helps determine the speed of learning:

$$V(s_{t-1}) = V(s_{t-1}) + \alpha \delta_t. \quad (9)$$

As a result, through the TD-learning algorithm, the estimated value of the new state directly influences the estimated value of the old state. Through this process, over repeated iterations, the reward prediction can percolate back to earlier and earlier states.

To step through these details of the TD learning algorithm more carefully, let us revisit the maze of Figure 2B. Suppose that the rat is completely naive, never having visited this maze before. For simplicity in calculation in this example, we set the step-size parameter α to .5 and the discount factor γ to 1. On the first trial, let us imagine that the rat goes up from the start state S1 to state S2. At this point, no rewards have been received, and the value of all states is 0, thus no learning occurs. Now, the rat goes to the right and receives a reward of +5. Because the value of S2 is 0, a large prediction error of +5 occurs, and the value of this state is updated to +2.5 (step size of .5 times a reward of +5). On the second trial, the rat again goes up from the start state S1 to state S2. This time, however, state S2 has a non-zero value. This change in estimated value induces a prediction error of +2.5 (value of state S2 minus the value of state S1), and the value of state S1 is now updated to +1.25 (.5 x +2.5), even though a reward has not been encountered yet. This propagation of value back through the states is the *bootstrapping* mechanism through which TD learning achieves efficient learning from the experienced rewards. Finally, imagine the rat again goes to the right and gets a reward of +5—the value of state S2 will again be updated, but this time by a smaller amount as there was already a reward prediction of +2.5 in state S2. The new value for state S2 will be +3.75, which is the original value of +2.5, plus .5 (the learning rate) times the difference between the reward received (+5) and the reward predicted (+2.5). The value of state S1 does not get updated again at this point, unless a memory mechanism known as *eligibility traces* are used—an RL technique we do not discuss here (see Sutton & Barto, 1998). Thus, after only two trials, the agent has gained new information about

the value of these two states, propagating information back through the states as they were encountered.

2.3. From Predictions to Actions

One limitation to TD learning, as discussed above, is that the algorithm does not provide a direct way of learning how to select actions. The TD algorithm only learns the value or predicted future rewards from different states, and thus could not be directly used by real agents that act upon their world and control the rewards they receive. Several solutions to this limitation suggest themselves. One idea would be to learn a separate value for each action leading out of a state, instead of for the state itself. We discuss a pair of such *action-value* methods—SARSA and Q-learning—below (Sutton & Barto, 1998; Watkins, 1989). Another idea would be to separately store a probability for taking each action in each state (known as a *policy*) and then adjust that policy based on the agent's experience, as in the *actor-critic* architecture that is popular in biological circles (e.g., Joel, Niv, & Ruppin, 2002; O'Doherty et al., 2004; Samejima & Doya, 2007). We evaluate the strengths and limitations of these major approaches for action selection in MDPs.

Action Values. We have discussed how a value function encodes the expected future reward from a given state. Value functions, however, cannot directly be used to decide which action to take, without knowledge of the transition function to the next state. The *action-value function* (*Q*), on the other hand, stores the value of taking a certain action from a state. In our maze example, the action-values in state S2 for the left and right actions are correspondingly -50 and 5. In mathematical notation, we would write that as $Q(\text{S2, left}) = -50$ and $Q(\text{S2, right}) = +5$. Similarly, the action-value for the up action in state S1 is -0.5, which corresponds to the expected reward if the rat moves up, based on past experience.

As with the value function for states, the action value for a state-action pair is the expected immediate reward (given that action) plus the sum of all future rewards. To make this clearer, let us denote the reward returned from the environment following action a_t as $r_{t+1}(a_t)$. The definition for the action-value function is quite similar to the equation for the value function of a state (compare Eq. 5):

$$Q(s_t, a_t) = r_{t+1}(a_t) + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} \ldots, \tag{10}$$

with the key difference that the action value does not depend on all the possible rewards following the current state, but only on the reward that follows the action in question. As with the value function (see Eq. 7), this action-value function can be expressed so that it also depends only on the next reward and the value of the next state:

$$Q(s_t, a_t) = r_{t+1}(a_t) + \gamma\, V(s_{t+1}). \tag{11}$$

So, in the maze example, the value of going up from S1 (i.e., Q(S1, up)) depends solely on the immediate reward for going up (0) and the value of the next state S2 (-.5). This relationship between the action value and the value of the next state is critical for the performance of *model-free* learning algorithms, which do not rely on knowing the transition function to learn about the world, but rather learn solely from experienced samples.

Learning about actions. The first action-value algorithm we review, *SARSA* (**S**tate-**A**ction-**R**eward-**S**tate-**A**ction), can be thought of as the natural extension of TD Learning to the decision-making case. The goal of an agent taking actions in an MDP is to maximize the sum of future rewards. When the environmental model is unknown, the agent must learn, from experience, about the value of the actions in each state. A very simple extension of the TD algorithm is to consider bootstrapping (learning a guess from a guess) from the action values, rather than the state values. Once again, we calculate a TD error (see Eq. 8), but this time based

on the difference between the old action value (what was expected) and the new action value plus the reward (what actually happened):

$$\delta_t = [r_t + \gamma Q(s_t, a_t)] - Q(s_{t\text{-}1}, a_{t\text{-}1}). \quad (12)$$

This error can then be used to update the old action value with a similar step-size parameter α that controls the speed of learning:

$$Q(s_{t\text{-}1}, a_{t\text{-}1}) = Q(s_{t\text{-}1}, a_{t\text{-}1}) + \alpha\, \delta_t \quad (13)$$

These two equations form the basis of the SARSA algorithm. Similar to what occurs in the TD algorithm, the action value of the next state-action pair percolates back to influence the estimate of the action value for the previous state-action pair.

To highlight the relationship between SARSA and TD learning, consider a naïve rat back in our example maze. For simplicity in calculation in this example, we again set the step-size parameter α to .5 and the discount factor γ to 1. As before, on the first trial, the rat goes up from state S1 to state S2; at this point, no rewards have been received, and no learning occurs. The rat then goes to the right and gets the large cheese reward (+5). Because the action value of going to the right in state S2 (Q(S2, right)) is 0, there is a large prediction error (+5), and this action value is updated to +2.5, which is equal to the old value of 0 plus the step size .5 times the prediction error of +5. On the next trial, imagine the rat again goes up from state S1 into state S2. At this point, unlike in TD learning, nothing happens yet. State S2 does not directly have a value of its own; another action needs to be taken before the state-action pair can be updated. When the action from state S2 is selected, then the learning occurs. The action value for going up in state S1 (Q(S1, up)) is updated based on the old action value (0), the reward received for going up (0), and the next action value (Q(S2, right) = +2.5), so that this action value is now +1.25 (consult Eqs. 12 and 13). Finally, the rat receives the large reward again (+5), and the action value for

going right in state S2 is duly updated to +3.75, based on the difference between the action value or expected reward (+2.5) and the reward actually received. As with TD learning, through the SARSA algorithm, the action values back up to earlier and earlier state-action pairs.

Exploration vs. exploitation. A new wrinkle is introduced when action selection is incorporated into learning. The agent now controls what experiences it receives. To perform well in any environment, the agent needs to encounter the states and actions that yield high reward. Yet, sampling new actions and states can be fraught with risk, especially when rewarding actions are already available to the agent. This delicate balance between maximizing the reward from known actions and sampling new opportunities is known in RL as the *exploration-exploitation dilemma.*

This dilemma is perhaps best explained through the example of a slot machine (often known as a bandit problem: Robbins, 1952). In this problem, an agent is faced with a number of slot machines, each with unknown payout rates. For example, Machine 1 pays out 4 dollars 10% of the time, while Machine 2 pays out only 1 dollar, but 50% of the time. At every time step, the agent must repeatedly choose between one of the machines. Here, Machine 2 has the higher expected value for each pull than Machine 1 (50 cents versus 40 cents). An agent learning about these slot machines, however, only has access to samples of the payouts; they do not have access to this underlying distribution. Suppose that on the first trial, the agent plays the first machine and receives $4. An agent using a learning rule such as SARSA would then update the action value for selecting the first machine towards $4. If the initial estimates were $0 for both machines, the agent would treat Machine 1 as the better choice, and would never select Machine 2. The more general issue at stake is that relying exclusively on imperfect action-value estimates for action selection can lead to distinctly sub-par behavior. The dilemma is as follows: An agent

must decide at each step whether to collect information to refine its estimates (explore) or take the best (greedy) action with respect to its action-value estimates (exploit).

To remedy this trade-off between exploiting current knowledge and sampling new options, a variety of algorithms to guide exploration have been developed in the RL literature (for overviews, see Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998; for particulars, see Auer, 2003; Brafman & Tenneholz, 2003; Daw et al., 2006; Kolter & Ng, 2009). Perhaps the simplest exploration rule is the *ε-greedy* algorithm, which picks the best action (highest action value) most of the time, but for some small portion ε of actions, picks an action at random. Slightly more sophisticated is the *Softmax* rule, wherein actions are taken at a frequency proportional to their action values. In both these cases, some randomness is inserted into the action selection process to ensure adequate coverage of the potential state space. Newer algorithms tend to direct exploration through optimism in the face of uncertainty (e.g., Auer, 2003; Brafman & Tennenholz, 2003). These important computational issues about balancing reward and knowledge in action selection have only recently begun to be addressed in a biological context (e.g., Daw et al., 2006).

Q-Learning. In the SARSA algorithm described above, the agent computes the value of a state-action pair based on the immediate reward and the next observed state-action pair, independent of whether the second action was exploratory. There is thus an interaction between future action choices and the process for updating the current state-action pair. But, imagine that the next action was a particularly bad exploratory action: Updating from the action value for that state-action pair seems like a very poor idea. As a result of this limitation, SARSA does not learn the best possible policy in certain situations (see Sutton & Barto, 1998).

The Q-Learning algorithm (Watkins, 1989) remedies this problem by learning from what

the agent could have done, rather than from what the agent actually did. To do so, Q-Learning uses a slightly different reward-prediction error for updating the action values (compare Eq. 12):

$$\delta_t = [r_t + \gamma \max_a Q(s_t, a)] - Q(s_{t-1}, a_{t-1}). \quad (14)$$

In Q-Learning, the agent learns from the difference between what was expected (the old action value) and the best possible outcome (maximum action value for that state ($\max_a Q(s_t, a)$), plus the reward). The action value is updated exactly as in SARSA (see Eq. 13). The key difference between Q-learning and the SARSA algorithm is that the agent learns from the estimated best action that it could have taken (Q-Learning), rather than from the action that it actually chose (SARSA).

Let us return again to the naïve rat in our example maze (Fig. 2A). Imagine that on the second trial, instead of going to the big cheese from the top state S2, the rat explored and found the cat (-50). With SARSA, learning is from the actual experience: The rat would compare the action value for going up in S1 (Q(S1, up) = 0) with the reward received for that action (0) and the action value for the next action (Q(S2, left) = 0, when it was taken), and no learning would occur. On future trials, selecting this exploratory action in state S2 would have even worse ramifications, as the action value would be negative, and the action value for going up in state S1 would be updated accordingly. With Q-Learning, however, the update is from the estimated best possible action out of state S2. As a result, the rat would compare the action value for going up in S1 (Q(S1, up) = 0) with the reward received for that action (0) and the action value for the *best possible action in the next state* (Q(S2, right), which after the first trial was updated to +2.5; see above). The action value Q(S1, up) would then be updated to +1.25, despite not taking the right action in state S2. By effectively ignoring the exploratory action, with this action sequence, Q-Learning learns more quickly than SARSA.

In our discussion of decision making, we suggested that a good agent should select actions to maximize the sum of future rewards. This action selection process, or policy, is *optimal* for a given MDP, if the policy maximizes the sum of future rewards from every state. Even with sufficient data, the SARSA algorithm does not learn the optimal policy unless strict conditions are enforced. In contrast, under specific technical conditions, Q-learning can be proven to converge to the value function that will yield the optimal policy (Watkins & Dayan, 1992). There are cases, however, when Q-learning is known to diverge (i.e., its prediction error grows without bounds)—for example, when the value function is only approximated because the state space is too large, as would be the case for most biologically relevant problems (Baird, 1995; for newer, related algorithms that do not diverge, see Maei et al., 2009; Sutton et al., 2009)—limiting the value of the algorithm in many computational settings. In addition, from a biological perspective, the maximization and counterfactual learning that drive Q-Learning may seem less plausible, but the evidence as to what type of action values may be used in the brain is mixed (Morris et al., 2005; Roesch, Calu, & Schoenbaum, 2007; Wunderlich, Rangel, & O'Doherty, 2009; see Section 3.3).

In the two action-selection algorithms discussed thus far (SARSA and Q-Learning), the behavior of the agent is driven by the action-value function. In those cases, modifying the action values immediately leads to changes in behavior. There is no separation between the evaluation system and decision-making system, which is perhaps not ideal for modeling decision making in animals. A more biologically plausible approach might be to explicitly separate the policy evaluation from the action-selection process (e.g., Joel et al., 2002; Samejima & Doya, 2007). The *actor-critic* algorithm is one such approach. This algorithm explicitly defines modules for each of the two mechanisms. The critic module plays the role of the evaluator, receiving the

reward signal and estimating state values through an algorithm like TD learning. The critic also outputs an error signal to the actor, which selects actions based on a set of stored preferences. A positive error from the critic reinforces the actor into taking the same action again, whereas a negative error inhibits such behavior. To compare the actor-critic framework with SARSA and Q-Learning, one may think of the latter two algorithms as implicitly defining the actor module while explicitly representing the value function. The actor-critic framework, on the other hand, defines both modules explicitly.

In this section, we introduced some of the major ideas that characterize the modern study of RL. We started with the concept of prediction from supervised learning and then discussed how RL adds time and sequentiality to the prediction problem. The main RL algorithms reviewed, TD-learning, SARSA, and Q-Learning, all take advantage of this temporality by bootstrapping or learning a guess from a guess. These simple algorithms provide the base for a powerful framework that has had many computational successes and is now being used as a model for learning in animals (see Section 1.2) and the brain (see Section 3.3).

3. Brain Mechanisms for Reinforcement Learning

Learning in animals requires some modification of the neuronal networks within the brain. One challenge for contemporary neuroscience is to determine the mechanisms underlying these modifications and how these mechanisms function in different forms of learning (see Section 2 for a discussion of computational strategies for learning). Substantial progress has been made over the past decade in understanding some of the neuronal events associated with reinforcement learning and value-based decision making as well as linking the computational models with neurobiological findings. This progress has been summarized in many excellent reviews (Daw & Doya, 2006; Dayan & Niv, 2008; Maia, 2009; Niv, 2009; Niv & Schoenbaum,

2008; Platt, 2002; Rangel et al., 2008; Rushworth & Behrens, 2008; Rushworth et al., 2009; Schultz, 2002, 2007; Schultz et al., 2008). Our goal in this section is not to go over the same ground as these reviews (some of which are quite advanced), but rather to describe some of the basic physiological processes underlying reinforcement learning in a manner that is accessible to investigators in the fields of machine learning and behavioral psychology wishing to become familiar with the relevant neurophysiology and brain anatomy.

3.1. Basic concepts in cellular neurobiology

A fundamental requirement for understanding the neurobiology of reinforcement learning is some basic knowledge of the cellular properties of nerve cells, the mechanisms for the transmission of information between nerve cells, and the processes by which the properties of nerve cells, and the networks they form, are modified (an essential requirement for learning). Thus we begin this section by briefly summarizing the key concepts related to this requirement.

Action potentials in nerve cells. Information in the nervous system is transmitted from one region to another in the form of *action potentials*. Action potentials are brief changes in the voltage levels around the membrane of a nerve cell (a *neuron*). They are initiated at or near the cell body and propagate in an all-or-none manner along the axon of the neuron. The amplitude and duration of action potentials are about 100 mV and 1 ms, respectively, and the velocity of conduction along the axon ranges from about 1 to 100 m/s depending on axon diameter. Information about the internal state of the nervous system and external events are often represented by the timing and frequency of action potentials (e.g., Rieke et al., 1999).

Recording the activity of single neurons (i.e., the occurrence of action potentials) during a behavioral task is one of the most powerful methods for gaining an understanding of the neuronal mechanisms underlying the functioning of the nervous system. Indeed, most of what we

know about the connection between reinforcement learning and the brain comes from these types of studies in monkeys and rats (e.g., Schultz et al., 1997). Figure 3A shows how this method involves positioning the tip of a fine microelectrode close to the cell body of a neuron to detect voltage changes generated by currents produced by action potentials in the space immediately outside the neuron. These *extracellularly* recorded potentials (spikes) are relatively small, typically having amplitudes in the range of .1 to .5 mV (Fig. 3B and C). This method is often referred to as *single-unit recording*. Figure 3D shows the activity patterns of single neurons derived from single-unit recording displayed as a *raster plot* in which the occurrence of each spike during a trial is represented by a dot along a time axis; the data for multiple trials are aligned horizontally and separated vertically to form a raster of spike activity. An average pattern of activity across many trials is illustrated as a histogram (Fig. 3D, bottom; see also Fig. 5A), which is the sum of the spikes recorded during multiple trials relative to an event in the behavioral sequence, such as the time of reward delivery in a reinforcement-learning task.

Single-unit recordings have been made during a variety of reinforcement-learning paradigms in rodents and non-human primates (see Section 3.3). Because of its invasive nature, single-unit recording cannot be used routinely in humans. Currently, the study of the neurobiology of reinforcement learning in humans primarily utilizes *function magnetic resonance imaging* (fMRI) to identify brain regions in which changes in blood flow produced by neuronal activity in large numbers of neurons are correlated with specific parameters in reinforcement-learning tasks. The signal detected in fMRI studies is usually referred to as the BOLD signal (**B**lood-**O**xygen-**L**evel-**D**ependent) and is thought to originate in brain regions sending and receiving task-related information (e.g., Logothetis et al., 2001). A major advantage of fMRI is that multiple brain systems can be examined simultaneously, but two drawbacks are

that it has poor spatial and temporal resolution and an inability to distinguish the activity between different classes of neurons.

Synaptic transmission. Most communication between neurons occurs at specialized junctions called *synapses* between the axon terminals of one neuron (the *presynaptic* neuron) and localized sites on the dendrites and/or cell body of another neuron (the *postsynaptic* neuron). Each action potential in the presynaptic neuron causes the release of a chemical transmitter, which binds to receptor molecules embedded in the membrane of the postsynaptic neuron. Depending on the transmitter and the type of postsynaptic receptor, the transmitter can act to either increase (excitatory transmission) or decrease (inhibitory transmission) the activity in the postsynaptic neuron. In the mammalian central nervous system, the most common excitatory and inhibitory transmitters are glutamate and gamma-amino-butyric-acid (GABA), respectively. Significant modification in the activity of a postsynaptic neuron by synaptic input requires the cooperative action of large numbers of synapses because the effect of transmitter release from a single synapse is very small. A single presynaptic neuron may make hundreds of synapses with a single postsynaptic neuron, and each postsynaptic neuron can receive inputs from hundreds of presynaptic neurons.

Neuromodulation. Closely related to synaptic transmission is the phenomenon of neuromodulation. Both neurotransmitters and neuromodulators are released in a similar manner from the axonal terminals, but they exert their actions on other neurons in different ways. Neurotransmitters bind to receptors associated with ion channels and briefly alter the ionic conductance of these channels, directly changing how molecules can get into and out of the neuron. On the other hand, neuromodulators usually exert their action much more slowly (but still in a sub-second range) via a relatively complex signaling pathway starting with the binding

of the neuromodulator to specific receptors coupled to membrane proteins (*G-proteins*). The G-proteins then bind to other membrane molecules that, when activated, increase the level of molecules called *second messengers* inside the postsynaptic neurons and/or axonal terminals. Second messengers have a widespread influence in cells, one of which can be to modify the strength of the synaptic transmission for conventional synapses. For example, a neuromodulator may increase the number of transmitter receptors in the postsynaptic membrane.

The most common neuromodulators are dopamine, noradrenaline, serotonin and acetylcholine. A striking characteristic of these four neuromodulatory systems in the mammalian brain is that they all originate from localized regions within the brain stem and all have diffuse and widespread projections to many regions of the brain. Thus these systems have the capacity to globally alter the functioning of the brain. For example, three of these systems (noradrenergic, serotonergic, and cholinergic) have important roles in regulating the sleep-wake cycle. Neuromodulatory systems can also have more specific actions, such as modification of the level of arousal (noradrenergic system) and the mediation of rewards (dopaminergic system). This global effect on brain function allows the dopamine reward system to potentially affect a wide array of behaviours.

Neuronal plasticity. All forms of learning, and the long-term storage of information in the brain (memory), are associated with long-term modification in the functioning of neuronal networks. The capacity of neuronal networks to be modified by experience (and in response to injury) is termed *neuronal plasticity*. Over the past 20 years, enormous advances have been made in our understanding of the cellular and molecular mechanisms associated with neuronal plasticity. We now know that experiential events can alter the properties of synaptic transmission, change the structure of synapses, and cause the growth of dendritic and axonal

processes. Especially important is that the magnitude of these changes can be strongly influenced by neuromodulators. An example relevant for the neurobiology of reinforcement learning is that the long-terms effects of high-frequency stimulation on the pathways from the cerebral cortex to the striatum (see next section for anatomy) are highly dependent on the level of dopamine (Reynolds et al., 2001). When dopamine levels are low, the transmission in these pathways is depressed, whereas facilitation occurs when dopamine levels are high, providing a potential outlet for a dopamine error signal to modulate long-term changes in the brain (see Section 3.3).

3.2. Anatomy of the brain

General organization. Another essential requirement for understanding the neurobiology of reinforcement learning is some knowledge of the anatomy of the brain. The brain consists of reasonably well-defined major structures that include the cerebral hemispheres, the brain stem, and the cerebellum. The hemispheres themselves are divided into four lobes: frontal, parietal, occipital, and temporal. Neurons in the cerebral hemispheres are primarily confined to a thin layer (the *cerebral cortex*) covering the entire outside surface of the hemispheres. Neurons located in the cerebral cortex make synaptic connections with neurons in other regions of the cortex and to neurons located in regions outside the cortex. Many of the latter are located in the brain stem. Distributed throughout the brain stem, and the junction between the brain stem and cerebral hemispheres are numerous clusters of neurons, called *nuclei.* Some major nuclei are illustrated in Figure 4A. These nuclei are the regions in which synaptic connections between neurons within the nuclei, and onto these neurons from other brain regions, are made. Two of the major structures in the forebrain are the thalamus and basal ganglia. The latter is known to be especially important for reinforcement learning in the brain, so it is necessary to consider the organization of the basal ganglia in detail.

The basal ganglia and dopaminergic neurons. The basal ganglia are situated on both sides of the upper brain stem and consist of an aggregation of anatomically distinct nuclei that are conventionally divided into two groups: the *striatum* consisting of the caudate nucleus, putamen, and nucleus accumbens, and the *globus pallidus* consisting of internal and external segments (Fig. 4B). The basal ganglia are traditionally thought to be important in the control and initiation of movement and are the main area of the brain damaged in certain movement disorders, such as Parkinson's and Huntington's disease.

The striatum receives inputs from the cortex and from numerous nuclei in the brain stem. Cortical inputs terminate largely in dorsal (upper) regions of the striatum (caudate nucleus and putamen), whereas inputs from brain stem nuclei terminate in both the dorsal and ventral (lower) regions of the striatum (especially the nucleus accumbens). This division may be related to the fact that neuronal systems in the dorsal and ventral regions of the striatum are differentially recruited in different reinforcement-learning paradigms (e.g., O'Doherty et al., 2004). Prediction-learning tasks (classical conditioning) recruit neuronal systems in the ventral striatum, while more complex action-choice tasks (instrumental conditioning) recruit networks in both the dorsal and ventral regions. Output from the basal ganglia originates primarily from the globus pallidus, with feedback to the cerebral cortex going via the thalamus.

Although the neuronal circuitry formed by the basal ganglia with other brain regions is quite complex (see Fig. 4B), a number of pathways important for mediating reinforcement learning have been well defined. By far the most intensively investigated are the input pathways to the striatum from the *substantia nigra pars compacta* (SNc) and *ventral tegmental area* (VTA). Neurons in these pathways release the neuromodulator dopamine. Neurons releasing dopamine are termed *dopaminergic* neurons. One critical function of dopaminergic neurons is to

provide information about rewards, which, in turn, modifies neuronal networks in the basal ganglia mediating behavioral actions (Schultz, 2002; Doya, 2007). Two observations provided initial evidence for the importance of dopaminergic neurons in behavioral modification: (1) electrical stimulation at sites close to the axons of dopaminergic neurons can function as the primary reward for instrumental conditioning, and (2) depletion of dopamine disrupts reward-based learning (see Schultz, 2007). Recently, a causal role for dopaminergic neurons in mediating reward has been demonstrated in genetically engineered mice (Tsai et al., 2009). Selective activation of dopaminergic neurons by light pulses in these animals can directly modify behavioral choice. We should also emphasize that the dopaminergic neurons originating in the VTA project in a diffuse manner to many other regions of the brain (not only the striatum) and thus have a role in other aspects of behavior apart from those involving reward, such as memory consolidation (e.g., Rossato et al., 2009) and mood (e.g., Yadid & Friedman, 2008).

3.3. The Neurobiology of Reinforcement Learning

A striking advance in the field of reinforcement learning has been the linking of the computational theory with the neurobiology of the brain. Numerous brain regions have been identified as being involved in different aspects of reinforcement learning and decision making, and strong correlations have been found between the activity of neurons in some of these regions with important variables from the computational models. In this section, we focus on the neuronal representation of only three of these variables: reward prediction error, reward value, and action value. Our goal is to illustrate, with a few examples, the neurobiological approach to reinforcement learning.

The general strategy used in neurobiological studies has been to record neuronal activity in different brain regions during a reinforcement-learning task designed to focus on a specific

variable (or a set of variables) from a computational model. In animals (primarily rats and monkeys), this strategy usually involves recording the activity of single neurons (Figure 3), whereas, in humans, the most common technique is to image the brain using functional magnetic resonance imaging (fMRI).

Representing reward-prediction errors. Most algorithms for reinforcement learning use a reward-prediction error (δ) to either predict a reward (as in classical conditioning) or to modify the probability of choosing different actions (as in instrumental/operant conditioning). Earlier, we detailed one computational use of a reward-prediction error in describing the temporal-difference (TD) algorithm (see Sections 1.1 and 2.2). A neuronal correlate to this TD error has been found in dopaminergic neurons through single-unit recording. Figure 5 illustrates the activity of a dopaminergic neuron at the beginning of a classical conditioning procedure and after conditioning has been established (Schultz et al., 1997; figure from Doya, 2007). This highly influential finding catalyzed the growing body of research looking at the correspondence between reward-based learning in the brain and the algorithms of RL.

We will now step through this important result in some detail (see Fig. 5). In the Schultz et al. (1997) study, at the beginning of training, dopamine neurons discharge briefly to an unpredicted reward (top row in Figure 5A). After conditioning has been established, the neurons discharge only in response to the conditioned stimulus (CS), but not to the reward (middle row). This shift in the timing of this activity burst from the time of reward to the time of the CS is exactly what is expected if the activity is related to the reward-prediction error, as formalized in the TD algorithm. Prior to conditioning, an unpredicted reward leads to a large prediction error because the animal is not predicting any rewards. After conditioning, the CS produces a large prediction error because the animal is not predicting any reward prior to the CS, but the arrival of

the CS leads to an increase in the predicted value of the upcoming reward and a corresponding reward-prediction error. During the interval between the CS and reward, the predicted value of the upcoming reward does not change (ignoring any temporal discounting), so there is no prediction error, corresponding to the low level of activity in dopaminergic neurons. Finally, when the predicted reward arrives, there is no difference between the predicted reward and the reward received, thus there is still no prediction error, which corresponds to the absence of activity in dopaminergic neurons. If, instead, the reward is reduced or omitted when predicted, then there is a negative prediction error because the animal received less reward than expected and, notably, there is a reduction in the activity of the dopaminergic neurons (bottom row in Figure 5). This real-time, moment-to-moment correspondence between the behaviour of dopamine neurons and the reward-prediction error in the TD algorithm is quite remarkable, and provides strong support for a real-time reinforcement-learning model, such as TD, over other trial-based, error-correcting algorithms, such as the Rescorla-Wagner model from psychology (see Sections 1.1 and 2.2).

This reward-prediction error can also serve as a crucial signal in modifying behaviour in instrumental learning tasks (see section 1.2), and the activity of dopaminergic neurons has been found to reflect this error in these tasks (Morris et al., 2005; O'Doherty et al., 2004; Roesch et al., 2007). Reward-prediction error signals in instrumental learning tasks have been postulated to change the probabilities of specific actions by modifying the strengths of pathways between the cortex and the striatum (see Samejima & Doya, 2007).

Representing reward value. Another component that plays a prominent role in many RL algorithms is the reward value (V) or expected reward (illustrated in Figure 5B; see Section 2.2). This value is the reward prediction from which the reward-prediction error (TD error) is

generated. The reward value can be manipulated in decision-making tasks by varying the magnitude and probability of rewards associated with different choice options. For example, in behaving monkeys, Tobler et al. (2005) found that the phasic bursts of activity in dopaminergic neurons following the presentation of the choice options were related to the expected reward value of the choice made by the animal. Larger and more probable rewards produced larger bursts in dopaminergic firing, presumably reflecting a larger reward-prediction error, due to the larger expected reward values. It is important to note that the increased activity occurred before the onset of any overt behavior and was therefore related to the decision-making process and not the motor action performed by the animal. This neuronal representation of expected reward value is not restricted to dopaminergic neurons. For example, single-unit recordings in behaving monkeys and rats have also revealed neuronal responses related to reward value in the striatum (Ito & Doya, 2009), orbitofrontal and prefrontal cortex (Duuren et al., 2009; Kennerley & Wallis, 2009; for a review, see Schoenbaum et al. 2009), and the posterior parietal cortex (Sugrue et al., 2004, 2005). Given the widespread projections of midbrain dopaminergic neurons, it is not surprising that expected reward value would be represented in widespread networks within the brain. This view receives support from fMRI studies in humans in which the magnitude of BOLD signals in the ventral striatum and regions of the prefrontal cortex, both of which receive strong input from dopaminergic neurons, are related to expected reward value (Knutson et al., 2005; Tobler et al., 2007).

Representing action value. During instrumental learning tasks, an action or a sequence of actions must be selected to maximize reward. A fundamental question, therefore, is how does an animal learn the best action to execute from each state? (see discussion in Section 2.3 on computational schemes for action selection). A number of RL algorithms utilize a variable

termed the *action value* (or *Q-value*), where the action value represents the expected reward for a given action. A neurophysiological correlate of these action values has been found in the activity patterns of neurons in the striatum of monkeys (Samejima et al., 2005; see also Kim et al., 2009; Morris et al., 2006; Roesch et al., 2007, 2009). In their experiment, monkeys were trained to choose between two actions (a left or right movement of a handle) in a task in which the probabilities of receiving a large juice reward for each action was varied. The reward probabilities were arranged in blocks, so that each action was reinforced either 10%, 50%, or 90% of the time. From an RL perspective, this probability manipulation would result in different action values for the actions, pending the reward probability currently in place for that action. The main finding was that nearly 50% of the recorded neurons in the striatum (putamen and anterior caudate) significantly changed their activity based on the change in the reward probability of only one action (i.e., the action value). These action-value specific neurons did not change their activity when the reward probability for the other action changed, suggesting that they were not encoding a composite value across both actions. Instead, the response of these neurons is related to a combination of the reward value and action. The action-value algorithms in reinforcement learning (such as Q-learning or SARSA; see Section 2.3) propose that action values are directly involved in the selection of future actions. How the representation of action value in the activity of striatal neurons might be utilized to determine future actions is currently unknown, but the close association of these neurons with other neuronal networks in the basal ganglia involved in motor actions suggests that these neurons may also be elements in the action selection networks.

In this section, we started with a didactic overview of broad issues in neuroscience important for understanding the neural basis of reinforcement learning, specifically intended for

a non-neuroscience audience. We then briefly introduced some of the main neuroscience findings providing evidence that certain areas of the brain may be understood as implementing algorithms from RL. We reviewed findings that some of the major constructs from these RL algorithms, such as predictions errors, reward values, and action values, have strong correlates in the dopaminergic system and related areas of the brain. Together, these ideas from RL are having a transformational effect on this area of neuroscience, providing a well-grounded, normative framework for detailing what these neurons are computing while animals are making value-based decisions.

4. Implications of the Computational Neuroscience of Reinforcement Learning

Throughout this chapter, we have illustrated the remarkable synergy between neuroscience, computer science, and psychology that characterizes the modern multi-disciplinary study of reinforcement learning. Our purpose in this chapter has been two-fold. We primarily attempted to provide an introduction to the key ideas in these three fields that is accessible to those unfamiliar with this literature. In addition, we played out some examples of how reinforcement learning is studied within each of these disciplines—from classical conditioning in psychology to TD learning in computer science to the firing of dopamine neurons in the brain. Along the way, we tried to clarify some of the subtler aspects of the different theories to perhaps enhance understanding of each domain for researchers in the other disciplines. We hope that our readers are now better prepared to learn more about this exciting and rapidly growing field.

One interesting question is how these findings from the computational neuroscience of RL have resonated back into the constitutive disciplines. In neuroscience, the effects have been clearly transformative—the ideas from reinforcement learning are central to all discussions of neural valuation and decision making, and new papers on the subject are published weekly. The

ideas have taken less of a hold in psychology, where a long tradition of theorizing in animal learning has yet to absorb the potential insights from RL, still preferring older formalisms such as the Rescorla-Wagner rule (see Section 1.1). More broadly in cognitive science, the associative mechanisms of RL as a potential account for human cognition face a significant uphill battle in an environment where information processing is often viewed as strictly symbolic (e.g., Gallistel & King, 2009). In the future, as these RL models come to explain more behavioural data, and as neuroscience ideas become more mainstream in psychology, we expect RL to gain a central place in psychology, as well.

Somewhat surprisingly, the least influenced home discipline seems to have been AI and computer science. The transfer of ideas here has been mostly a one-way street, with neuroscience using the formalisms of RL for modeling the brain and behavior, but with little direct feedback. A potential future avenue for this sort of reciprocating feedback may eventually come from more detailed knowledge of the psychological and neural mechanisms that drive reward-based learning and decision making than what exists today. Indeed, historically, the computational study of RL was originally inspired by exactly these sorts of psychological and neural considerations (see Sutton & Barto, 1981, 1998). For example, one of the very first RL algorithms, the associative search network (Barto, Sutton, & Brewer, 1981), drew on the fact that animals do not require target outputs, such as which motor command to execute at each moment, to learn about the world. This algorithm stood in contrast to the existing supervised-learning approaches of the time, and started setting the way for the full development of RL ideas within AI that has followed over the past 25 years (see Section 2.1).

Another angle for potential feedback from the computational neuroscience to AI is suggested by the fact that many of the computational challenges currently facing RL stem from

tasks that the brain seems to handle naturally. For instance, the action-selection mechanisms discussed here (Q-Learning, SARSA, actor-critic, see Section 2.3) all depend on the explicit enumeration of the actions in order to obtain their values—something not possible when many different actions are possible. By studying action selection in the brain more closely, we might gain insight into appropriate algorithms that may be applicable to large-scale problems, such as real-world robots.

Perhaps the most likely source for the transmission of ideas back to AI from this area of computational neuroscience, however, lies not in direct inspiration from the biological substrate, but rather from the new models that have grown up to explain the neuroscientific and psychological data. Though the initial RL models in neuroscience came from AI, newer models have grown from these roots and adapted as they attempt to accommodate more and more data. Variations on well-known RL ideas to deal with issues like temporal discounting (Kurth-Nelson & Redish, 2009), motivational effects (Niv et al., 2005), or response timing (Ludvig et al., 2008) could eventually provide a new source of inspiration for those researchers interested in creating artificial systems with human- or animal-like learning abilities.

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology, 33B*, 109-122.

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces that Shape our Decisions*. New York: Harper.

Asgarian, N., Hu, X., Aktary, Z., Chapman, K. A., Lam, L., Chibbar, R., Mackey, J., Greiner, R., & Pasdar, M. (2009). Learning to predict relapse in invasive ductal carcinomas based on the subcellular localization of junctional proteins. *Breast Cancer Research and Treatment.* Epub: 2009 Sep 29.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research, 3*, 397-422.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. *International Conference on Machine Learning, 12*, 30-37.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*, 407-419.

Balleine, B. W., & O'Doherty, J. P. (2009). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*. Epub: 2009 Sep 23.

Barto, A. G., Sutton, R. S., & Brouwer, P. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics, 40*, 201-211.

Bateson, M., & Kacelnik, A. (1995). Preferences for fixed and variable food sources: Variability in amount and delay. *Journal of the Experimental Analysis of Behavior, 63*, 313-329.

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk.

Econometrica, 22, 23-36.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*, 1020-1022.

Brafman, R. I., & Tennenholtz, M. (2003). R-MAX—A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research, 3,* 213-231.

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology, 25*, 323-332.

Camerer, C., & Loewenstein, G. (2003). Behavioral economics: Past, present, future. In: C. Camerer, G. Loewenstein, & M. Rabin (Eds.) *Advances in Behavioral Economics*. (pp. 3-51). New York and Princeton: Russell Sage Foundation Press and Princeton University Press.

Caraco, T. (1981). Energy budgets, risk and foraging preferences in dark-eyed juncos (*Junco hyemalis*). *Behavioral Ecology and Sociobiology, 8*, 213-217.

Caraco, T., Blanckenhorn, W. U., Gregory, G. M., Newman, J. A., Recer, G. M., & Zwicker, S. M. (1990). Risk-sensitivity: Ambient temperature affects foraging choice. *Animal Behaviour, 39*, 338-345.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York, NY: Oxford University Press.

Colwill, R. C., & Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 120-132.

Cooper G. F., Abraham V., Aliferis, C. F., Aronis, J. M., Buchanan, B. G., Caruana, R., Fine, M. J., Janosky, J. E., Livingston, G., Mitchell, T., Montik, S., & Spirtes, P. (2005). Predicting dire outcomes of patients with community acquired pneumonia. *Journal of Biomedical Informatics*, *38*, 347-366.

Davison, M., & McCarthy, D. (1988). *The Matching Law: A Research Review*. Hillsdale, NJ: Erlbaum.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation, 18*, 1637-77.

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

Daw, N. D., Niv, Y. & 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., Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*, 876-9.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive & Affective Behavioral Neuroscience, 8*, 429-53.

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

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

Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annual Review of Psychology, 56*, 179-206.

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *Human Frontiers Science Program Journal, 1*, 30-40.

Dwyer, D. M., Starns, J., Honey, R. C. (2009). "Causal reasoning" in rats: A reappraisal. *Journal of Experimental Psychology: Animal Behavior Processes, 35*, 578-86.

Gallistel, C. R., & King, A. P. (2009). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Malden, MA: Wiley-Blackwell.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107*, 289-344.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124.

Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (Eds.). (2009). Neuroeconomics: Decision making and the brain. San Diego, CA: Academic Press.

Gottlieb, D. A. (2008). Is the number of trials a primary determinant of conditioned responding? *Journal Experimental Psychology: Animal Behavior Processes, 34*, 185-201.

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

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition, 113*, 293-313.

Guthrie, E. R. (1930). Conditioning as a principle of learning. *Psychological Review, 37*, 412-428.

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.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior, 13*, 243-266.

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

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks, 15*, 535-47.

Jozefowiez, J., Staddon, J. E., & Cerutti, D. T. (2009). The behavioral economics of choice and interval timing. *Psychological Review, 116*, 519-39.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence, 4,* 237-285.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-292.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Kennerley, S. W., & Wallis, J. D. (2009). Evaluating choices by single neurons in the frontal lobe: Outcome value encoded across multiple decision variables. *European Journal of Neuroscience, 29*, 2061-2073.

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

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distrubuted neural representation of expected value. *Journal of Neuroscience, 25*, 4806-4812.

Kolter, J. Z. & Ng, A. Y. (2009). Near-bayesian exploration in polynomial time.

International Conference on Machine Learning, 26, 513-520.

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One, 4*, e7362.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84*, 555-579.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86,* 2278-2324.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature, 412*, 150-157.

Loewenstein, Y., Prelec, D., & Seung, H. S. (2009). Operant matching as a Nash equilibrium of an intertemporal game. *Neural Computation, 21*, 2755-2773.

Ludvig, E. A., & Koop, A. (2008). Learning to generalize through predictive representations: A computational model of mediated conditioning. In *From Animals to Animats: Proceedings of Simulation of Adaptive Behavior*, *10*, 342-351.

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

Ludvig, E. A., Sutton, R. S., Verbeek, E. L., & Kehoe, E. J. (2009). A computational model of hippocampal function in trace conditioning. *Advances in Neural Information Processing Systems, 21*, 993-1000.

Maei, H. R., Szepesvari, C., Bhatnagar, S., Precup, D., Silver, D., & Sutton, R. S. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in Neural Information Processing Systems, 21,* 1609-1616.

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

Marr, D. C. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Marsh, B., & Kacelnik, A. (2002). Framing effects and risky decisions in starlings. *Proceedings of the National Academy of Sciences, USA, 99*, 3352-3355.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955/2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine, 27*, 12-14.

McDowell, J. J. (2004). A computational model of selection by consequences. *Journal of the Experimental Analysis of Behavior*, *81*, 297-317.

Miller, R. R., Barnet, R. C., Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin, 117*, 363-386.

Mitchell, T. (1997). *Machine learning*. Burr Ridge, IL: McGraw Hill.

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

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

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

Niv, Y., Daw, N. D., & Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. *Advances in Neural Information Processing Systems, 18*, 1019-1026.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive*

Science, 12, 265-272.

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

Pavlov, I. P. (1927). Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerberal Cortex (G. V. Anrep Trans.). London: Oxford University Press.

Pearce, J. M. & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52*, 111-139.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532-552.

Platt, M. L. (2002). Neural correlates of decisions. *Current Opinion in Neurobiology, 12*, 141-148.

Platt, M. L. & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature Neuroscience, 11*, 398-403.

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

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Compartive and Physiological Psychology, 66*, 1-5.

Rescorla, R. A. (1980). Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 6*, 207-216.

Rescorla, R. A. (1988) Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151-160.

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

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature, 413*, 67-70.

Rieke, F., Warland, D., van Steveninch, R. d. R., & Bialek, W. (1999). Spikes: Exploring the Neural Code. Cambridge, MA: MIT Press.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527-535.

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

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

Rossato, J. I., Bevilaqua, L. R. M., Izquierdo, I., Medina, J. H., & Cammarota, M. (2009). Dopamine control persistence of long-term memory storage. *Science, 325*, 1017-1020.

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience, 11*, 389-397.

Rushworth, M. F. S., Mars, R. B., & Summerfield, C. (2009). General mechanisms for making decisions? *Current Opinion in Neurobiology, 19*, 75-83.

Sakai, Y., & Fukai, T. (2008). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation, 20*, 227-251.

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

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

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience, 12*, 885-892.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36*, 241-263.

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

Schultz W. (2009). Neuroeconomics: The promise and the profit. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*, 3767-3769.

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

Shafir, S. (2000). Risk-sensitive foraging: The effect of relative variability. *Oikos, 88,* 663-669.

Simen, P., & Cohen, J. D. (2009). Explicit melioration by a neural diffusion model. *Brain Resarch, 1299*, 95-117.

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

Staddon, J. E. R., & Cerutti, D. T. (2003). Operant conditioning. *Annual Review of Psychology, 54*, 115-144.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A

formalization of the comparator hypothesis. *Psychological Review, 114*, 759-83.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science, 304*, 1782-1790.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience, 6*, 363-375.

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

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135-171.

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

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

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvari, C., & Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. *International Conference on Machine Learning*, *26*, 993-1000.

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

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

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of*

Neurophysiology, 97, 1621-1632.

Trepel, C., Fox, C. R., Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Brain Research Cognitive Brain Research, 23*, 34-50.

Tsai, H-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., Deisseroth, K. (2009). Phasic firing of dopaminergic neurons is sufficient for behavioral conditioning. *Science, 324*, 1080-1084.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

van Duuren, E., van der Plasse, G., Lankelma, J., Joosten, R. N. J. M. A., Feenstra, M. G. P., & Pennartz, C. M. A. (2009). Single-cell and population coding of expected reward probability in the orbitofrontal cortex of the rat. *Journal of Neuroscience, 29*, 8965-8976.

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

Wagner, A. R. (1981). SOP: a model of automatic memory processing in animal behavior. In: Spear, N. E., Miller, R. R. (Eds.), *Information Processing in Animals: Memory Mechanisms* (pp. 5-47). Hillsdale, NJ: Erlbaum.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. Ph.D. Thesis. University of Cambridge, England.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning, 8*, 279-292.

Williams, D. A., Lawson, C., Cook, R., Mather, A. A., & Johns, K. W. (2008). Timed excitatory conditioning under zero and negative contingencies. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 94-105.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences, USA, 106*, 17199-17204.

Yadid, G., & Friedman, A. (2008). Dynamics of the dopaminergic system as a key component in the understanding of depression. *Progress in Brain Research, 172*, 265-286.

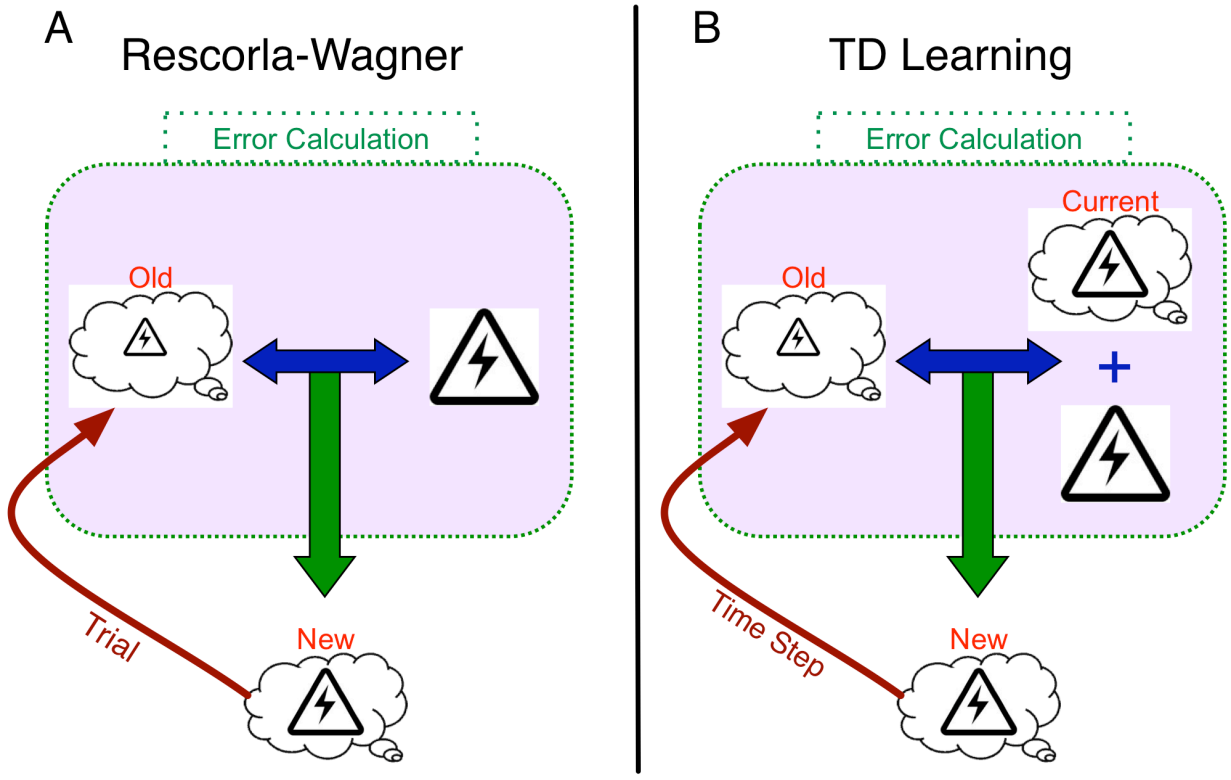

Figure 1. Learning rule schematics for an experiment with a negative shock reward. A. Learning in the Rescorla-Wagner rule is driven by the difference (two-sided arrow) between the reward prediction and the actual reward on a given trial. This reward-prediction error is used to create a new reward prediction (down arrow), which can be used on the next trial (curved arrow). B. Learning in the temporal-difference (TD) algorithm. The *Old* prediction is the reward prediction based on the stimuli that were around on the last time step. The *Current* prediction is the prediction based on the stimuli that are currently available. An error is generated by comparing these two predictions with the reward (two-sided arrow). This error is then used to change the way the algorithm makes its predictions (down arrow). As a result, a *New* prediction can be made based on the stimuli that are still currently available. In some sense, the learning process converts the Current prediction into the New prediction. On the next time step, this New prediction becomes the Old prediction, and the process begins all over again (curved arrow).

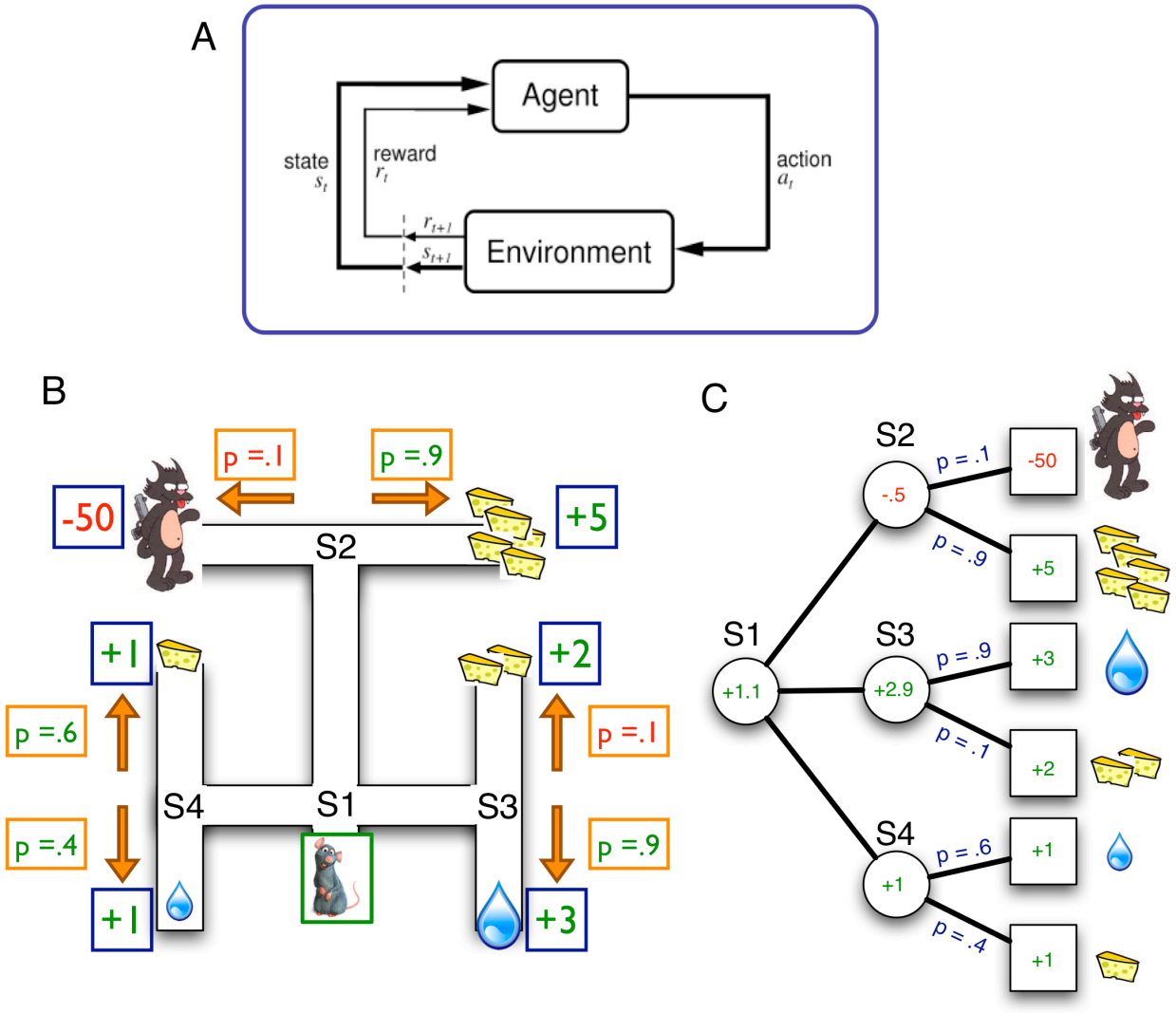

Figure 2. Reinforcement Leaning and Markov Decision Processes. A. The world according to reinforcement learning. There is an agent that interacts with an environment by emitting actions, and receiving states (or observations) and rewards in return. B. A small maze in which the rat at the bottom can navigate and obtain various rewards. Signed values (+ or -) are the reward magnitudes at different end points. Probabilities (*p*) of previous actions at the final choice points are indicated. The states are numbered S1 through S4. C. Same maze problem, but abstracted to a series of connected states. Numbers in the circles indicate the values of various states, given the history.

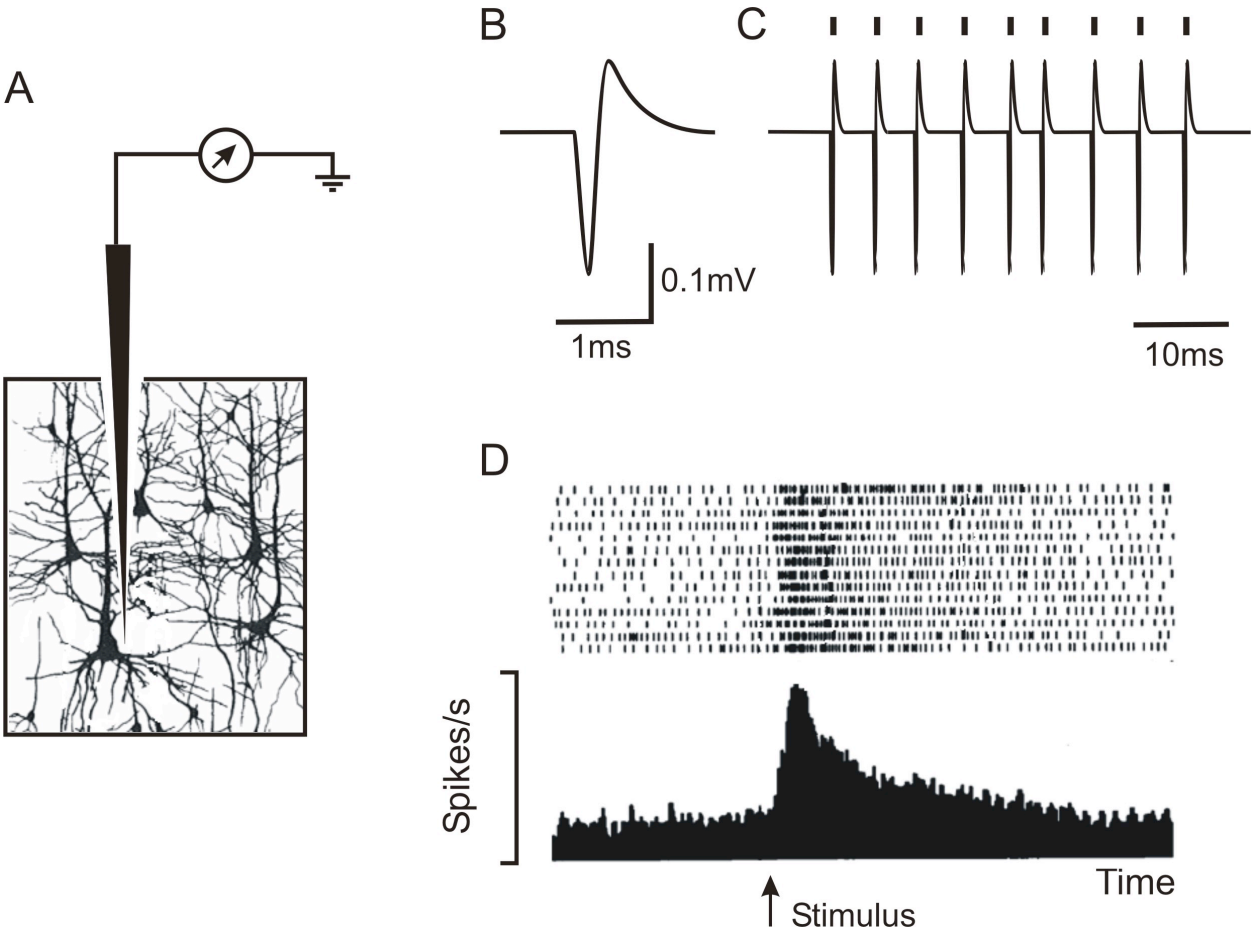

Figure 3. Single unit recording. A. Schematic showing the placement of the tip of an electrode outside but close to the body of a nerve cell. B-C. Drawings of a single spike (action potential) (B) and train of spikes (C) recorded in the extracellular space. Note the small amplitude and the biphasic shape. The time of the occurrence of spikes in a spike train are usually illustrated by a small marker as shown above the spike train. D. *Top*. Raster display of spike trains recorded in response to multiple presentation of a stimulus. Each dash represents a single potential and each row represents a separate trial, each aligned on the time of the stimulus. *Bottom*. Peri-stimulus time histogram showing the average activity across all trials.

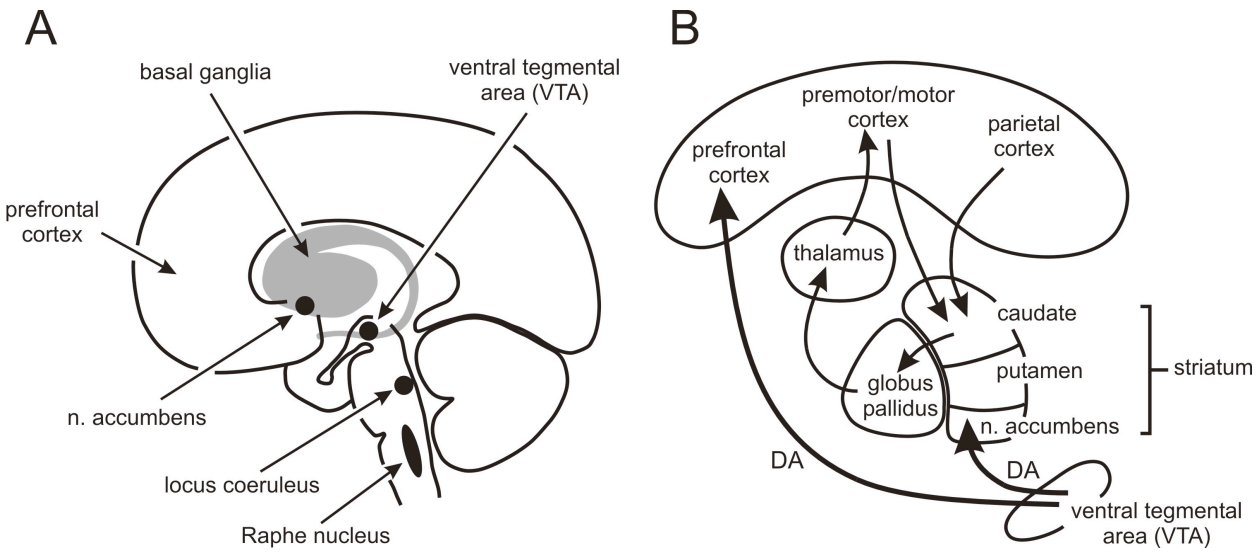

Figure 4. Basic anatomy of brain regions involved in reinforcement learning. A. Drawing of sagittal (side) section through the brain showing the location of the basal ganglia (shaded grey) and some brain stem nuclei (filled black). B. Schematic diagram showing the anatomy of the basal ganglia in more detail (note the striatum is the combination of the caudate nucleus, putamen and nucleus accumbens) and some of the main connections to and from the basal ganglia. Dopaminergic pathways (DA, thick arrows) originate in the ventral tegmental area (VTA) and project densely to the nucleus accumbens in the ventral striatum and to the prefrontal cortex.

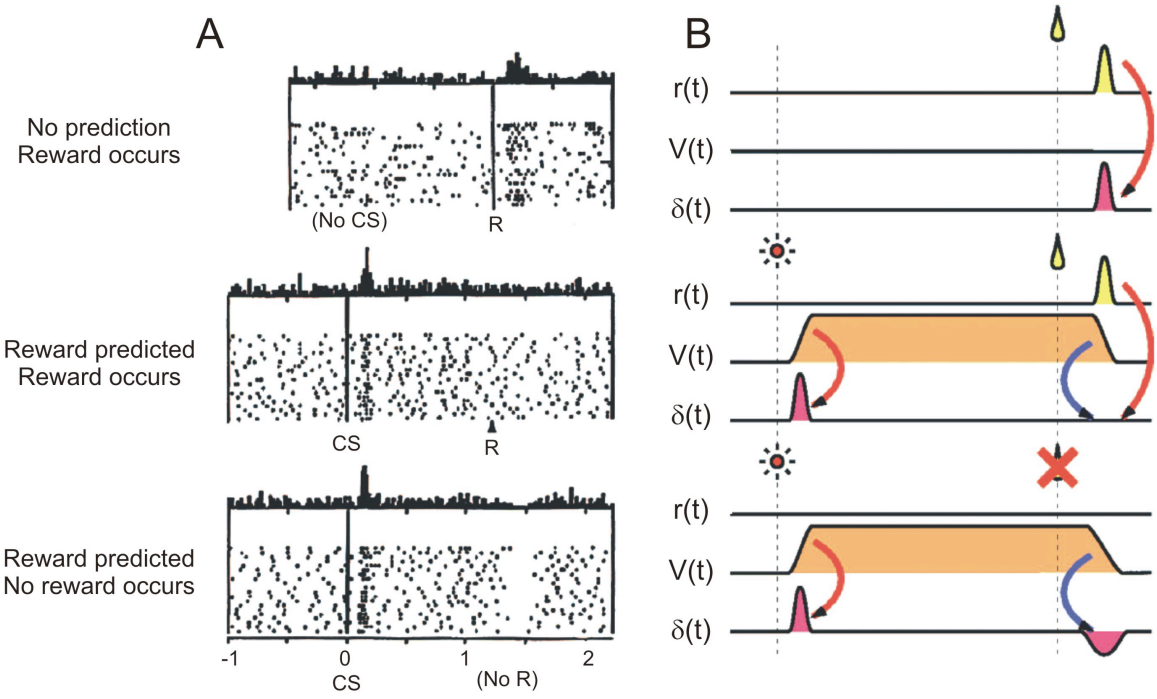

Figure 5. Dopamine neurons and reinforcement learning. A. Results from Schultz et al. (1997) of dopamine neuron activity in three situations. In the first case, an unpredicted reward (R) occurs, and a burst of dopamine firing follows. In the second case, a predicted reward occurs, and a burst follows the onset of the predictor (CS or conditioned stimulus), but there is no firing after the now-predicted reward. Finally, in the bottom case, a predicted reward is omitted, with a corresponding trough in dopamine responding. B. How the various elements of the TD learning algorithm—reward (r), value (V), and error (δ)—change during the time course of the different trials (adapted with permission from Doya, 2007).

Keywords and Definitions

Reinforcement Learning (*RL*). Branch of Artificial Intelligence (AI) that focuses on learning from interactive experience. Also used to describe the collection of processes whereby humans and animals learn through rewards.

Classical Conditioning. Simple learning process whereby humans and animals learn predictive relationships between stimuli and rewards.

Operant Conditioning. Simple learning process whereby humans and animals learn to perform actions based on rewarding experience.

Dopamine. Small molecule that is used in the brain as a neurotransmitter to communicate between neurons. Thought to encode the error in reward predictions.

Temporal-difference (TD) algorithm. Reinforcement-learning technique that learns to predict rewards based on the error between predicted outcomes and actual outcomes.

Striatum. Brain area that receives heavy input from dopamine neurons. Thought to be important for reward valuation and action selection.

Reward. An important outcome, which can be positive or negative. Maximization of reward serves as the goal for reinforcement-learning agents.

Neuroeconomics. New multi-disciplinary enterprise that attempts to explain how value-based decisions are made in the brain.