Chapter 13

Psychology

Among the connections of reinforcement learning to other disciplines, its connections to the study of animal learning by psychologists are among the most extensive. Many of the basic reinforcement learning algorithms were inspired by psychological theories of animal learning, and reinforcement learning algorithms and theory are, in turn, contributing back to psychology. There are detailed models of animal learning that use algorithmic ideas from reinforcement learning, and the theoretical and computational perspectives of reinforcement learning are influencing psychologists in designing new experiments and animal learning models.

In this chapter we discuss how concepts and algorithms from reinforcement learning correspond, in certain ways, to theories of animal learning from psychology. These correspondences should not be surprising. Of all the paradigms of machine learning, we regard reinforcement learning as the closest to the kind of learning that humans and other animals do. All machine learning paradigms are abstractions of situations in which humans and other animals learn, but some are more faithful to these situations than others. The supervised and unsupervised learning paradigms are abstractions of important components of animal learning, but in isolation they do not do justice to the fact that animals learn goal-directed behavior while interacting with dynamic environments. It is essential to keep in mind, however, that reinforcement learning as developed here explores idealized situations from the perspective of an artificial intelligence researcher or engineer—not from the perspective of an animal learning researcher.

Our position outside of psychology makes it possible for us to sidestep the many enduring controversies that have influenced its history. We can be selective in connecting with psychology because our goal is not to replicate animal behavior. For the same reason not every feature of computational reinforcement learning corresponds to a psychological finding or theory. Some findings from psychology have proven valuable given our focus on computational effectiveness, while others have not. The correspondences we describe mainly involve learning theories derived from laboratory experiments with animals. Despite the fact that its influence in contemporary psychology has been overshadowed by emphasis on more cognitive aspects of intelligence, the study of animal learning has established principles that are compelling from a computational perspective due to their combination of precision and generality. We think it would be unwise to neglect these principles in designing systems that use learning methods in solving engineering problems.

This chapter is far too short to include all of the points of correspondence between the theory presented in this book and even this subarea of psychology. For the most part, the correspondences we describe are those of particular significance because they connect ideas or mechanisms that arose independently in their respective fields. We believe these points of correspondence improve our understanding of both computational and psychological learning principles. While we provide numerous references in the final section of this chapter, many of the connections between reinforcement learning and psychology remain beyond our treatment, including extensive research that has borrowed from computational reinforcement learning with the psychological goal of accounting for subtle experimental findings about animal learning and decision making. We hope this chapter provides a useful context for the reader who wishes to probe the subject more deeply.

13.1 Prediction and Control

The algorithms we describe in this book fall into two broad categories: algorithms for *prediction* and algorithms for *control*. These categories arise naturally in solution methods for the reinforcement learning problem presented in Chapter 3. In many ways these categories respectively correspond to categories of learning extensively studied by psychologists: classical, or Pavlovian, conditioning and instrumental conditioning. Although not completely accidental because of psychology's influence on reinforcement learning, these correspondences are nevertheless striking because they are the result of independent objectives.

The prediction algorithms presented in this book estimate quantities that depend on how features of an agent's environment are expected to unfold over the future. We specifically focused on estimating the amount of reward an agent can expect to receive over the future while it interacts with its environment. In this role, these are *policy evaluation algorithms*, which are integral components of algorithms for improving policies. Chapter 4 presents policy evaluation in the context of dynamic programming, and Chapters 5 and 6 respectively present Monte Carlo and Temporal Difference (TD) policy evaluation methods. But prediction algorithms are not limited to predicting future reward; they can predict any numerical-valued feature of the environment, and they can be studied without considering their role in policy improvement. The correspondence between prediction algorithms and classical conditioning rests on the property they have in common of predicting upcoming stimuli, where the stimuli are not necessarily rewards or penalties earned by previous actions.

Classical, or Pavlovian, conditioning experiments address how learning causes reflexes to be triggered by stimuli that reliably predict the natural triggering stimuli of those reflexes. Roughly speaking, a stimulus, call it A, is a reliable predictor of another stimulus, call it B, the predicted stimulus, if B regularly occurs shortly after A and rarely occurs otherwise. A classical conditioning experiment exposes an animal to stimuli in a predictive relationship, where the predicted stimulus reflexively triggers a response. The animal learns to respond to the predicting stimulus in a manner similar to how it responds to the predicted stimulus, thereby acting in anticipation of the predicted stimulus. We discuss classical conditioning in more detail below and here just point out two critical features. First, detecting predictive relationships among events is at the core of this form of learning. Second, classical conditioning experiments are set up to make this relationship independent of the animal's behavior. This means that the predicted stimulus follows the predicting stimulus no matter what the animal does in response to the predictor. Although an animal's response may affect the impact on the animal of the predicted stimulus, the animal does not control whether or not the predicted stimulus occurs.

The situation in an instrumental conditioning experiment is different. Here, the experimental apparatus is set up so that an animal is given something it likes (a reward) or something it dislikes (a penalty) depending on what the animal does. The animal learns to increase its tendency to produce rewarded behavior, and to decrease its tendency to produce penalized behavior. The reinforcing stimulus is said to be *contingent* on the animal's behavior, whereas in classical conditioning it is not. Instrumental conditioning experiments are like those that inspired Thorndike's Law of Effect that we briefly discuss in Chapter 1. *Control* is at the core of this form of learning, which corresponds to the operation of reinforcement learning's policy-improvement algorithms.

At this point, we should follow psychologists in pointing out that the distinction between classical and instrumental conditioning is one between the *experimental setups* (whether or not the experimental apparatus makes the reinforcing stimulus contingent on the animal's behavior). It is not necessarily a distinction between different *learning mechanisms.* In practice, it is very difficult to remove all response contingencies from an experiment, and the extent to which these types of experiments engage different learning mechanisms is a complicated issue about which animal learning theorists have differing views. Our engineering and artificial intelligence perspective may shed some light on this issue. Algorithms for prediction clearly differ from those for control, but many of the reinforcement learning methods we present involve closely linked combinations of both. Animal learning mechanisms likely follow this pattern as well.

We now take a closer look at classical conditioning and details of a particularly close correspondence between animal behavior in these experiments and temporal-difference prediction.

13.2 Classical Conditioning

The celebrated Russian physiologist and Nobel laureate Ivan Pavlov studied how reflexes can come to be triggered by stimuli other than their innate triggers:

> It is pretty evident that under natural conditions the normal animal must

> respond not only to stimuli which themselves bring immediate benefit or harm, but also to other physical or chemical agencies—waves of sound, light, and the like—which in themselves only *signal* the approach of these stimuli; though it is not the sight and sound of the beast of prey which is in itself harmful to the smaller animal, but its teeth and claws. (Pavlov, 1927, p. 14)

Pavlov (or more exactly, his translators) called inborn reflexes "unconditioned reflexes" and new reflexes triggered by predictive stimuli "conditioned reflexes." This terminology persists in describing classical conditioning experiments, where conditioned stimuli (CSs), which are initially neutral in the sense that they do not normally elicit a strong response, are set up to predict biologically significant events (such as a taste of food, a shock, etc.), called unconditioned stimuli (USs), that reflexively produce unconditioned responses (URs), such as salivation or an eye blink.

URs are often protective in some way, like an eye blink in response to something irritating to the eye, or "freezing" in response to seeing a predator. Experiencing the CS-US predictive relationship over a series of trials causes the animal to learn to respond to the CS with a conditioned response (CR) that better protects the animal from, or better prepares it for, the US. The CR tends to be similar to the UR but begins earlier and sometimes differs in ways that increase its effectiveness. For example in one intensively studied type of experiment, a tone CS reliably predicts a puff of air to a rabbit's eye (the US), triggering closure of a protective membrane (the UR). With one or more trials, the tone comes to trigger a CR consisting of membrane closure that begins before the air puff and eventually becomes timed so that peak closure occurs just when the air puff is likely to occur. This CR, being initiated in anticipation of the air puff and appropriately timed, offers better protection than simply initiating closure in reaction to the irritating US. The ability to act in anticipation of important events by learning about predictive relationships among stimuli is so beneficial that it is widely present across the animal kingdom.

Figure 12.1 shows the arrangement of stimuli in two types of classical conditioning experiments: in delay conditioning, the US occurs while the CS is still present, whereas in trace conditioning, the US begins after the CS ends. In trace conditioning, the time interval between CS offset and US onset is called the trace interval. The interstimulus interval, or ISI, is the time interval between CS onset and US onset.

The understanding that CRs anticipate USs eventually led to the development of an influential model based on TD learning. This model, called the *TD model of classical conditioning*, or just the *TD model*, extends what is arguably the most widely-known and most influential model of classical conditioning: the Rescorla-Wagner model (Rescorla and Wagner, 1972). Rescorla and Wagner created their model in order to provide an account of what happens in classical conditioning with compound CSs, that is, CSs that consist of several component stimuli, such as a tone and a flashing light occurring together, where the animal's history of experience with each stimulus component can be manipulated in various ways. Experiments like these demonstrate, for example, that if an animal has already learned to produce a CR in response to a stimulus component that predicts a US, then learning to produce

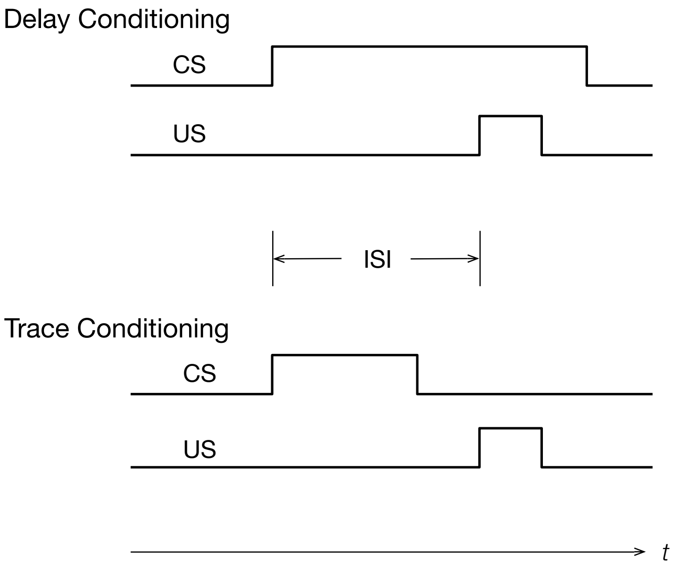

Figure 13.1: Arrangement of stimuli in two types of classical conditioning experiments. In delay conditioning, the US occurs while the CS is present. In trace conditioning, there is a time interval between CS offset and US onset. The interstimulus interval (ISI) is the interval between CS onset and US onset.

a CR in response to a newly-added second stimulus is much reduced. This is called *blocking*. Results like this challenge the idea that conditioning depends only on simple temporal contiguity, that is, that the only requirement for conditioning is that a US frequently follows a CS closely in time. In contrast, the core idea of the Rescorla-Wagner model is that an organism only learns when events violate its expectations, in other words, only when the organism is surprised (although without implying any conscious expectation or emotion).

Here is how Rescorla and Wagner described their model. The model adjusts the "associative strength" of each stimulus, which is a number representing the amplitude of the CR elicited by the stimulus, or how reliably the stimulus elicits the CR. Associative strengths can be positive or negative, with negative values meaning that the stimulus inhibits the CR. When a compound CS consisting of several component stimuli is presented on a trial of a classical conditioning experiment, the associative strength of each component stimulus changes in a way that depends on an associative strength associated with the entire stimulus compound, called the "aggregate associative strength," and not just on the current associate strength of each component itself.

Rescorla and Wagner considered a stimulus compound AX, where the animal may have already experienced stimulus A, and stimulus X might be new to the animal. Let V_{A}, V_{X}, and V_{AX} respectively denote the associative strengths of stimuli A, X, and the compound AX. Suppose that on a trial the compound CS AX is followed by a US, which we label stimulus Y. Then the associative strengths of the stimulus components change according to these expressions:

$$\begin{aligned}\Delta V_{\mathrm{A}} &= \alpha_{\mathrm{A}}\beta_{\mathrm{Y}}(\lambda_{\mathrm{Y}} - V_{\mathrm{AX}})\\ \Delta V_{\mathrm{X}} &= \alpha_{\mathrm{X}}\beta_{\mathrm{Y}}(\lambda_{\mathrm{Y}} - V_{\mathrm{AX}}),\end{aligned}$$

where $\alpha_{\mathrm{A}}\beta_{\mathrm{Y}}$ and $\alpha_{\mathrm{X}}\beta_{\mathrm{Y}}$ determine the learning rate, which depends on both the CS and US, and λ_{Y} is the asymptotic level of associative strength that the US Y can support. (This λ is Rescorla and Wagner's notation and does *not* correspond to the λ of the TD(λ) family.) The model makes the further important assumption that $V_{\mathrm{AX}} = V_{\mathrm{A}} + V_{\mathrm{X}}$.

To complete the model one needs to define a way of mapping values of V to CRs. Since such a mapping would depend on details of the experimental situation, Rescorla and Wagner did not specify a mapping but simply assumed that it would be order preserving and that negative Vs would generally correspond to the absence of a CR.

To connect this model to TD algorithms, think of the conditioning process as one of learning to predict the "magnitude of the US" on a trial on the basis of the stimulus compound present on that trial, where the magnitude of a US Y is the λ_{Y} of the Rescorla-Wagner model as given above. Suppose the stimulus compound on a trial t consists of up to n component stimuli and is represented by a vector $\mathbf{x}_t$ with binary coordinates $x_t(i)$, $i = 1, \ldots, n$, where there is a one in coordinate i if stimulus component i is present on trial t and a zero if that stimulus component is not present. Denote the respective associative strengths of these stimulus components by weights $v_t(i)$, $i = 1, \ldots, n$. Then the aggregate associative strength on trial t is

$$V_t = \sum_{i=1}^{n} v_t(i)x_t(i). \tag{13.1}$$

This corresponds to a *value estimate* of reinforcement learning and is thought of it as the *US prediction*.

As a result of a conditioning trial t, the weight vector, $\mathbf{v}_t$, is updated as follows:

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \alpha\delta_t\mathbf{x}_t, \tag{13.2}$$

where α is the step-size parameter. For the Rescorla-Wagner model δ_t is the *prediction error*

$$\delta_t = \lambda_t - V_t. \tag{13.3}$$

Here, λ_t is the 'target' of the prediction on trial t, that is, the magnitude of the US, or in Rescorla and Wagner's terms, the associative strength that the US on the trial can support (where, again, this does not correspond to the λ of TD(λ) family). Note that the term $\mathbf{x}_t$ in Equation 12.2 implies that only the associative strengths of stimuli present on a trial are adjusted as a result of that trial. The prediction error is considered to be a measure of surprise, with the aggregate associative strength representing the animal's expectation that is violated when it does not match the target US magnitude.

In this form, the Rescorla-Wagner model is recognizable as an error-correcting supervised learning rule identical to the Widrow-Hoff Least Mean Square (LMS) learning rule (Widrow and Hoff, 1960), with the exceptions that for LMS the input vectors $\mathbf{x}_t$ can have any real numbers as coordinates and the scalar step-size parameter α does not depend on both the input vector and target value. The latter is a minor deviation that more elaborate forms of the LMS rule can incorporate. Error correction provides a ready explanation for many of the phenomena observed in classical conditioning with compound stimuli. For example, in a blocking experiment when a new component is added to a stimulus compound to which the animal has already been conditioned, further conditioning with the augmented compound produces little or no increase in the associative strength of the added stimulus. Prior learning blocks learning to the added stimulus component because the error has already been reduced to zero, or to a low value. Because the occurrence of the US is already predicted, no new surprise is introduced by adding the new stimulus component.

Although the Rescorla-Wagner model provides a simple and compelling account of blocking and other features of behavior in classical conditioning experiments, it is far from being a perfect or complete model. Different ideas account for a variety of other observed effects, and progress is still being made toward understanding the many counterintuitive subtleties of classical conditioning. One direction of extension concerns the timing of stimuli. A single time step in the above formulation of Rescorla and Wagner's model represents an entire conditioning trial. The model does not apply to details about what happens during the time a trial is taking place. Within each trial an animal might experience various stimuli whose onsets occur at particular times and that have particular durations. These timing relationships strongly influence learning. For example, one of Pavlov's most reliable observations was that the CS must begin before the US begins for learning to occur; when it is the other way around, he observed no learning at all (although later researchers reported a slight amount of learning for negative ISIs in certain cases).

The model of classical conditioning based on TD learning is a generalization of the Rescorla-Wagner model. It accounts for all of the behavior accounted for by that model but goes beyond it to account for how within-trial and between-trial timing relationships among stimuli influence learning. The TD model of classical conditioning is called a *real-time* model, as opposed to a *trial-level* model like the Rescorla-Wagner model.

To describe the TD model we can begin with the formulation of the Rescorla-Wagner model above, but we now interpret the time step t as representing a small interval of real time, say, 0.01 seconds, and we have to be a bit more precise about the aggregate associative strength. Let t and t' be two possibly different time steps. Then

$$V_t(\mathbf{x}_{t'}) = \sum_{i=1}^{n} v_t(i) x_{t'}(i) \tag{13.4}$$

is the aggregate associate strength at time step t due to the stimulus compound present at time step t'. Then learning occurs according to this update:

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \alpha \delta_t \mathbf{e}_t, \tag{13.5}$$

where α is the step-size parameter, δ_t is the TD error defined below, and $\mathbf{e}_t$ is a vector of eligibility traces at time t that accumulate according to the presence of stimulus components and decay according to $\gamma\lambda$:

$$\mathbf{e}_{t+1} = \gamma\lambda \mathbf{e}_t + \mathbf{x}_t. \tag{13.6}$$

Here γ is the discount factor (between 0 and 1) and λ is the eligibility trace-decay parameter—*not* the Rescorla-Wagner λ in Equation 12.3.

Instead of the Rescorla-Wagner δ_t of Equation 12.3, the TD model uses the TD prediction error to update the US prediction:

$$\delta_t = r_t + \gamma V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1}), \tag{13.7}$$

where γ is the discount factor, and r_t indicates the US strength at time t instead of the trial-level λ_t of the Rescorla-Wagner model. Note that if $\gamma = 0$, the TD model prediction error is identical to the prediction error of the Rescorla-Wagner model (except for the single time-step delay of the stimulus compound).

Like the Rescorla-Wagner model, the TD model does not specify a particular response generation mechanism that converts the US prediction into a behavioral response that can be compared to an animal's CR. One can simply let the time course of the US prediction V directly represent the time course of the CR, as has been done in a number of modeling studies. With this assumption, one can see by comparing the description above with our account of TD learning with linear function approximation in Chapter 9 that the TD model of classical conditioning is the backward view of the gradient-descent TD(λ) algorithm for the case of linear function approximation. The only difference is that in modeling classical conditioning, r_t represents the US strength at time step t instead of the reward signal at time step t as it does when the algorithm is used as a component of a policy-improvement algorithm. The TD model with more complicated response generation mechanisms have also been studied. One example is a thresholded leaky integrator used by Ludvig, Sutton, and Kehoe (2012).

Missing from this description of the TD model is a description of how the various stimuli should be represented to best account for animal data. The TD model is a real-time model, so the representation used in the Rescorla-Wagner model—a one if a stimulus component is present on a trial and a zero otherwise—is not adequate. One has to specify how stimuli appear as extended over time within trials. Moreover, since the convergence of TD learning relies on the idea of states, in fact, on Markovian states, one has to make assumptions about the sequence of states an animal's nervous system passes through during the multiple trials of a classical conditioning experiment.

It would be ideal to base these assumptions on what is known about the activity of neural circuits during conditioning, but since knowledge of this activity is not sufficiently detailed, the usual practice is to investigate the behavior predicted by the model under a variety of assumptions. This allows researchers both to explore the nature of the model's operation and to provide varying levels of support for different hypotheses about what the neural activity might actually be like. Ludvig et al. (2012) describe three representations (Figure 12.2) that make different predictions when combined with the TD model together with their thresholded leaky integrator response generator: the *presence* representation, the *complete serial compound* (CSC) representation, and the *microstimulus* (MS) representation. These differ along a temporal generalization gradient, referring to the degree to which they allow generalization among nearby time points during which a stimulus is present. The presence representation allows complete generalization, the complete serial compound representation allows no generalization, and the generalization allowed by the microstimulus representation falls between the other two. In modeling an experiment involving more than one stimulus, a light and a tone for example, each stimulus is given its own representation (each of which is usually of the same type).

The presence representation is like the one used in the Rescorla-Wagner model but where the representation has value one throughout the time period during which a stimulus is present and value zero when it is absent (the third column of Figure 12.2). Despite its simplicity, simulations show that the TD model with this stimulus rep-

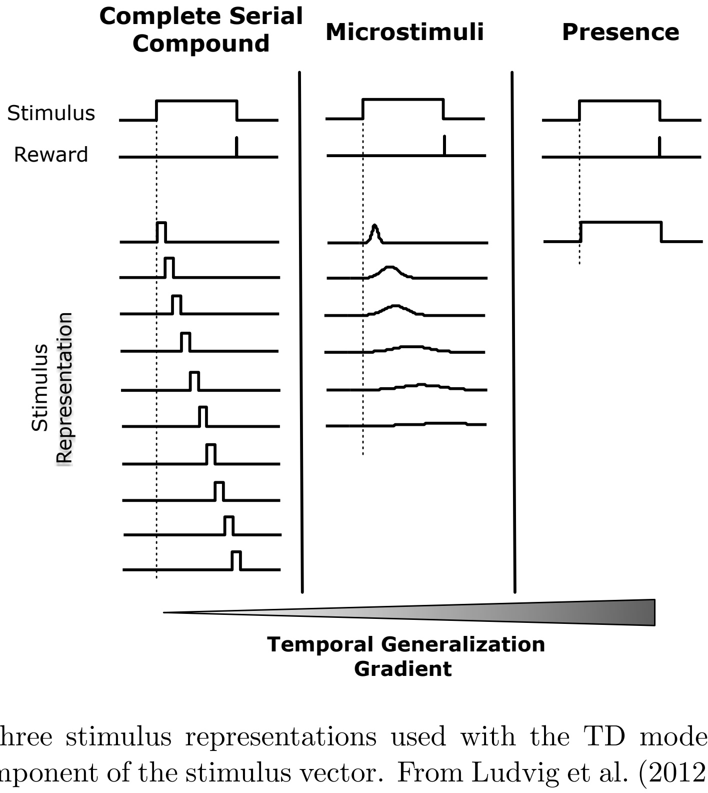

Figure 13.2: Three stimulus representations used with the TD model. Each row represents a component of the stimulus vector. From Ludvig et al. (2012), permission pending.

resentation can produce qualitatively good facsimiles of a wide range of phenomena observed in classical conditioning experiments. These include all of the phenomena produced by the Rescorla-Wagner model, such as blocking, which we described above, and overshadowing, which is when simultaneous conditioning with two stimuli results in the more salient stimulus producing the stronger response.

In addition to the features of classical conditioning produced by the Rescorla-Wagner model, the TD model with the presence representation (and other representations as well) produces facsimiles of phenomena that involve the relative timing of stimuli. Foremost among these is a conspicuous feature of classical conditioning that the US generally must begin *after* the onset of a neutral stimulus for conditioning to occur. In other words, conditioning generally requires a positive ISI. This follows from the fact that the model's behavior parallels animal behavior in how the asymptotic level of conditioning (e.g., the percentage of CRs elicited by a CS) depends on the ISI. The overall shape of this dependency varies substantially across species and response systems, but it is nearly always zero for zero or negative ISIs, increases to a maximum at a positive ISI where conditioning is most effective (often some fraction of a second), and then decreases to zero after an interval that varies widely with response systems. The precise shape of this dependency for the model depends on its parameter values and the details of its input representation, but these basic features are core properties of the TD model.

Another feature of the TD model's behavior involving stimulus timing deserves attention because the model correctly predicted a feature of animal learning that had not been observed at the time of the model's introduction. The TD model with a presence representation (as well as with more complex representations) predicts that blocking is reversed if in a third stage of a blocking experiment the blocked stimulus is moved earlier in time so that its onset occurs before the onset of the blocking stimulus. The simulated three-stage procedure involving two CS stimuli, A and B, and a US is shown in Figure 12.3. For the first 10 trials CS A is presented alone followed by the US, and its associative strength, V_A, increases as shown in the graph. Trials 11–20 correspond to the second stage of a blocking experiment: A has already been conditioned and now B is introduced with the same time course a A, followed by the US. The model shows complete blocking, with V_B remaining at its initial value of zero during these trials. For the third stage, trials 21–35, B is extended so that its onset precedes A's onset. B is now an earlier predictor of the US than A. Not only does B's associative strength increase over these trials, A actually loses associate strength. The model's prediction led Kehoe, Scheurs, and Graham (1987) to conduct the experiment using the well-studied rabbit eye-blink conditioning paradigm (actually the rabbit learns to retract its eyeball in anticipation of an air puff or small shock near its eye so that a protective membrane, the nictitating membrane, moves over the eye.) They observed the key features of the model's prediction and noted that other theories have considerable difficulty explaining their data.

The TD model with a presence representation (and other representations as well) can also generate a facsimile of *second-order conditioning*—something that the Rescorla-Wagner model cannot do. This is the phenomenon in which a previously-conditioned

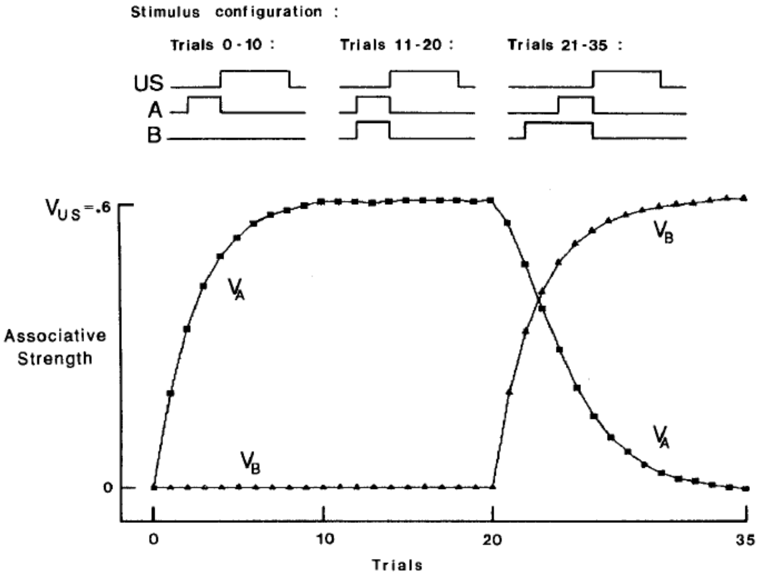

Figure 13.3: Temporal primacy overrides blocking. From Sutton and Barto (1990), permission pending.

CS can act as if it were a US in conditioning another initially neutral stimulus. Pavlov described an experiment in which his assistant conditioned a dog to salivate to the sound of a metronome that predicted a food US. Then a number of trials were conducted in which a black square, to which the dog was initially indifferent, was placed in the dog's line of vision followed by the sound of the metronome—and this was *not* followed by food. In just ten trials, the dog began to salivate just upon seeing the back square, despite the fact that it had never been paired with food. The sound of the metronome itself acted as reinforcement for the salivation response to the black square. Learning analogous to second-order conditioning occurs in instrumental tasks, where a stimulus that consistently predicts reward (or penalty) becomes rewarding (or penalizing) itself, producing what is called secondary, or conditioned, reinforcement. This happens whether the predicted reinforcing stimulus is a primary reinforcing stimulus or another secondary reinforcing stimulus.

The presence of $V_t(\mathbf{x}_t) - V_t(\mathbf{x}_{t-1})$ in the TD error (Equation 12.7) means that the error can be non-zero as a result of previous learning that made $V_t(\mathbf{x}_t)$ differ from $V_t(\mathbf{x}_{t-1})$ (a temporal difference). This difference has the same status as the reward signal in the TD error, implying that as far as learning is concerned there is no difference between a temporal difference and a reward signal. In fact, this feature of the TD algorithm is one of the major reasons for its development, which we now understand through its connection to dynamic programming as described in Chapter 6. Backing-up values is intimately related to second-order, and higher-order, conditioning.

Where the presence representation has a single element for each stimulus, the CSC representation has a separate element for each moment of time during which a stimulus is present (the first column of Figure 12.2). This means that a single stimulus is represented—in the discrete-time case—by a vector having as many elements as there are time steps in the duration of the stimulus, where each element is 'on' for exactly one of those time steps. If multiple stimuli are involved in a simulation, each is represented by a separate vector. The CSC representation completely differentiates every moment of a stimulus's duration so that there is no temporal generalization. This representation was devised to allow different versions of the TD model to 'show off' their behavior while being as unconstrained as possible by the stimulus representation. It is as if a sequence of distinct states is passed through during a conditioning trial, with the resulting US predictions (comprising a value function) being stored in what is essentially a lookup table whose contents are the associative strengths of the various stimulus elements.

The name *serial compound* comes from classical conditioning experiments in which an animal is exposed to multiple external stimuli strung out over time like the elements of a CSC representation. Here, though, we think of the elements of a CSC vector as being a collection of *internal microstimuli* generated by the nervous system in response to the presentation of a single external stimulus, such as the stimulus shown in the top row of Figure 12.2. Although not particularly plausible from a neural perspective, the CSC representation has been widely used in studying TD models of both classical conditioning and the activity of dopamine producing neurons, the latter discussed in Chapter 13. Ludvig et al. (2012) call the CSC representation a 'useful fiction'.

The second column of Figure 12.2 illustrates the MS representation which, like the CSC representation, consists of a vector whose elements are also thought of as internal microstimuli, but in this case they are not of such limited and non-overlapping form. Several examples of MS representations have been studied in the literature, and their details need not concern us here. The important point is that this kind of representation, by being situated between the presence representation—which permits complete generalization among the different moments of a CS's presence—and the CSC representation—which permits no such generalization—produces a limited form of temporal generalization that allows the behavior of the TD model to be related to a broader collection of phenomena observed in animal experiments. Notable examples of these phenomena involve the timing and shape of CRs and how these change throughout the course of learning.

The TD model of classical conditioning, when combined with a particular response-generation mechanism and stimulus representation, is able to account for a surprisingly wide range of phenomena observed in classical conditioning experiments, but it is far from being a perfect model. To generate other details of classical conditioning the model needs to be extended by, for example, adding model-based elements and mechanisms for adaptively altering some of its parameters. Other approaches to modeling classical conditioning depart significantly from the Rescorla-Wagner style error-correction process. Bayesian models, for example, work within a probabilistic framework in which experience revises probability estimates. All of these models usefully contribute to our understanding of classical conditioning.

Perhaps the most notable feature of the TD model is that it is based on a theory—the theory we have described in this book—that suggests an account of what an animal's nervous system is *trying to do* while undergoing conditioning: it is trying to form accurate predictions of the *long-term return* over the animal's future, consistent with the limitations imposed by the way stimuli are represented and how the nervous system works. In other words, it suggests a *normative account* of classical conditioning in which long-term, instead of immediate, prediction is a key feature.

The development of the TD model of classical conditioning is one instance in which the explicit goal was to model details of animal learning behavior. In addition to its standing as an *algorithm*, then, TD learning is also the basis of this *model* of aspects of biological learning. As we discuss in Chapter 13, TD learning has also turned out to underlie an influential model of the activity of dopamine producing neurons in the brain. These are instances in which reinforcement learning theory makes detailed contact with animal behavioral or neurophysiological data.

We now turn to considering correspondences between reinforcement learning and animal behavior in instrumental conditioning experiments, the other major type of laboratory learning experiment studied by psychologists.

13.3 Instrumental Conditioning

Instrumental conditioning experiments differ from classical conditioning experiments in that the delivery of a reinforcing stimulus depends on the animal's behavior, whereas in a classical conditioning experiment the reinforcing stimulus—the US—is delivered independently of what the animal does. The roots of instrumental conditioning go back to experiments performed by the American psychologist Edward Thorndike one hundred years before publication of the first edition of this book.

Thorndike observed the behavior of cats when they were placed in "puzzle boxes" from which they could escape by appropriate actions (Figure 12.4). For example, a cat could open the door of one box by performing a sequence of three separate actions: depressing a platform at the back of the box, pulling a string by clawing at it, and pushing a bar up or down. When first placed in such a box, with food visible outside, all but a few of Thorndike's cats displayed "evident signs of discomfort" and extraordinarily vigorous activity "to strive instinctively to escape from confinement" (Thorndike, 1898).

In experiments with different cats and boxes with different escape mechanisms, Thorndike recorded the amounts of time each cat took to escape over multiple experiences in each box. He observed that the time almost invariably decreased with successive experiences, for example, from 300 seconds to 6 or 7 seconds. He described cats' behavior in a box (with a simpler escape mechanism) like this:

> The cat that is clawing all over the box in her impulsive struggle will

> probably claw the string or loop or button so as to open the door. And gradually all the other non-successful impulses will be stamped out and the particular impulse leading to the successful act will be stamped in by the resulting pleasure, until, after many trials, the cat will, when put in the box, immediately claw the button or loop in a definite way. (Thorndike 1898, p. 13)

These and other experiments (some with dogs, chicks, monkeys, and even fish) led Thorndike to formulate a number of "laws" of learning, the most influential being the *Law of Effect* that we quoted in Chapter 1. This law describes what is generally known as learning by trial and error. As mentioned in Chapter 1, many aspects of the Law of Effect have generated controversy and its details have been modified over the years. Still the law—in one form or another—expresses an enduring principle of learning.

Essential features of reinforcement learning algorithms correspond to features of animal learning described by the Law of Effect. First, reinforcement learning algorithms are *selectional*, meaning that these algorithms try alternatives and select among them by comparing their consequences. Second, reinforcement learning algorithms are *associative*, meaning that the alternatives found by selection are associated with particular situations, or states, to form the agent's policy. Like learning described by the Law of Effect, reinforcement learning is not just the process for *finding* actions that produce a lot of reward, but also for *connecting* them to situations or states. Thorndike used the phrase learning by "selecting and connecting" (Hilgard, 1956). Natural selection in evolution is a prime example of a selectional process, but it is not associative (at least as it is commonly understood); supervised learning is associative, but it is not selectional because it relies on instructions that directly tell the agent how to change its behavior.

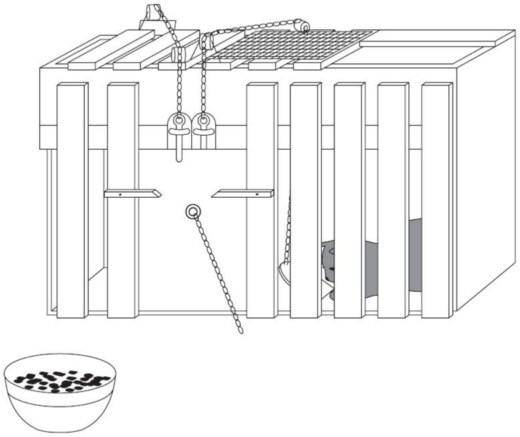

Figure 13.4: One of Thorndike's puzzle boxes. Copyright 2001 Psychology Press Ltd. permission pending.

In computational terms, the Law of Effect describes an elementary way of combining *search* and *memory*: search in the form of trying and selecting among many actions in each situation, and memory in the form of associations linking situations with the actions found to work best in those situations. Search and memory are essential components of all reinforcement learning algorithms, whether memory takes the form of an agent's policy, value function, or environment model.

A reinforcement learning algorithm's need to search means that it has to *explore* in some way. Animal's clearly explore as well, and early animal learning researchers disagreed about the degree of guidance an animal uses in selecting its actions in situations like Thorndike's puzzle boxes. Are actions the result of "absolutely random, blind groping" (Woodworth, 1938, p. 777), or is there some degree of guidance, either from prior learning, reasoning, or other means? Although some thinkers, including Thorndike, seem to have taken the former position, others disagreed. In fact, in some problem-solving experiments animals were said to demonstrate *insight* because the animals found a solution rather suddenly, sometimes after periods not involving physical exploratory activity during which the animal seemed to "figure out" the solution. Reinforcement learning algorithms allow wide latitude for how much guidance an agent can employ in selecting actions. The forms of exploration we have used in the algorithms presented in this book, such as ϵ-greedy and upper-confidence-bound action selection, are merely among the simplest. More sophisticated methods are possible, with the only stipulation being that there has to be *some* form of exploration for the algorithms to work effectively.

The feature of our treatment of reinforcement learning allowing the set of actions available at any time to depend on the environment's current state also echoes something Thorndike observed in his cats' puzzle-box behavior. They selected actions from those that they instinctively perform in their current situation, which Thorndike called their "instinctual impulses." First placed in a puzzle box, a cat instinctively scratches, claws, and bites with great energy: a cat's instinctual responses to finding itself in a confined space. Successful actions are selected from these and not from every possible action or activity. This corresponds to the feature of our formalism where the action selected from a state s belongs to a set of available actions, $A(s)$. Specifying these sets is an important aspect of reinforcement learning because it can radically simplify learning. They are like an animal's instinctual impulses.

Among the most prominent animal learning researchers influenced by the Law of Effect were Cark Hull (e.g., Hull, 1943) and B. F. Skinner (e.g., Skinner, 1938). At the center of their research was the idea of selecting behavior on the basis of its consequences. Reinforcement learning has features in common with Hull's theory, which included eligibility-like mechanisms and secondary reinforcement to account for the ability to learn when there is a significant time interval between an action and the consequent reinforcing stimulus (see Section 12.4). Randomness also played a role in Hull's theory through what he called "behavioral oscillation" to introduce exploratory behavior.

Skinner did not fully subscribe to the memory aspect of the Law of Effect, being averse to the idea of associative linkage and emphasizing instead selection from spontaneously-emitted behavior. He introduced the term “operant” to emphasize the key role of an action’s effects on an animal’s environment. Unlike the experiments of Thorndike and others, which consisted of sequences of separate trials, Skinner’s *operant conditioning* experiments allowed animal subjects to behave for extended periods of time without interruption. He invented the operant conditioning chamber, now called a “Skinner box,” the most basic version of which contains a lever or key that an animal can press to obtain a reward, such as food or water, which would be delivered according to a well-defined rule, called a reinforcement schedule. By recording the cumulative number of lever presses as a function of time, Skinner and his followers could investigate the effect of different reinforcement schedules on the animal’s rate of lever-pressing. Operant conditioning is often regarded as being the same as instrumental conditioning, but Skinner’s original intent was to study the effects of reinforcement on behavior in environments more like animals’ natural environments. It is fair to say that an animal in an instrumental conditioning experiment faces what we call an episodic task, whereas in an operant conditioning experiment, it faces what we call a continuing task.

Another of Skinner’s notable contributions was his recognition of the effectiveness and importance of training an animal by reinforcing successive approximations of the desired behavior, a process he called *shaping.* Although this technique had been used by others, including Skinner himself, its significance was impressed upon him when he and colleagues were attempting to train a pigeon to bowl by swiping a wooden ball with its beak. After waiting for a long time without seeing any swipe that they could reinforce, they

> ... decided to reinforce any response that had the slightest resemblance to a swipe—perhaps, at first, merely the behavior of looking at the ball—and then to select responses which more closely approximated the final form. The result amazed us. In a few minutes, the ball was caroming off the walls of the box as if the pigeon had been a champion squash player. (Skinner, 1958, p. 94)

Not only did the pigeon learn a behavior that is unusual for pigeons, it learned quickly through an interactive process in which its behavior and the reinforcement contingencies changed in response to each other. Skinner compared the process of altering reinforcement contingencies to the work of a sculptor shaping clay into a desired form. Shaping is a powerful technique for computational reinforcement learning systems as well. When it is difficult for an agent to receive any non-zero reward signal sat all, either due to sparseness of rewarding situations or their inaccessibility given initial behavior, starting with an easier problem and incrementally increasing its difficulty as the agent learns can be an effective, and sometimes indispensable, strategy.

Being able to shape behavior to be uncharacteristic of an animal’s natural behavior appears to be at odds with Thorndike’s observation that his cats’ puzzle box activity was selected from their “instinctual impulses.” This discrepancy is more apparent than real because the process of shaping successively changes the situations the animal experiences. Activity that is instinctual in situations occurring early in learning leads to new situations in which different collections of activities are instinctual. In some situations, the set of actions upon which selection can work may shrink to a single reflexive response, as has been observed in animal training where even with careful shaping some behaviors seemed impossible to obtain. (Breland and Breland, 1961, provide a famous account of some of their failed animal training experiences.)

A concept emphasized by psychologists that is especially relevant in the context of instrumental conditioning is *motivation*, which refers to processes that influence the direction and strength, or vigor, of behavior. Thorndike's cats, for example, were motivated to escape from puzzle boxes because they wanted the food that was sitting just outside. Obtaining this goal was rewarding to them and reinforced their escapes. We do not use the term motivation in presenting reinforcement learning, but it clearly corresponds to elements of the theory. For most of the reinforcement agents we have discussed, value functions are the main driving force determining the agent's direction of behavior: one might say that an agent is motivated to *ascend the gradient of the its value function.*

Psychologists say that the effect of reward depends on an animal's *motivational state.* For example, an animal will be more rewarded by eating (as measured by its rate of learning) when it is hungry than when it has just finished a satisfying meal. In reinforcement learning, the generation of reward signals depends on the state of a reinforcement learning agent's environment in addition to the agent's actions, and this environment is everything outside of the reinforcement agent, which can include information analogous to an animal's motivational state. The concept of state dependence is broad enough to allow for many types of modulating influences on the generation of reward signals.

We turn now to the subject of learning when reinforcing stimuli occur well after the events they reinforce. The mechanisms used by reinforcement learning algorithms to enable learning with delayed reinforcement—eligibility traces and TD learning—closely correspond to psychologists' hypotheses about the means animals use to learn under these conditions.

13.4 Delayed Reinforcement

The Law of Effect requires a backward effect on connections, and some early critics of the law could not conceive of how the present could affect something that was past. This concern is amplified by the fact that learning can even occur when there is a considerable delay between an action and the consequent reward or penalty. Similarly, in classical conditioning, learning can occur when US onset follows CS offset by a non-negligible time interval. We call this the problem of delayed reinforcement, which is related to what Minsky (1961) called the "credit-assignment problem for learning systems." The reinforcement learning algorithms presented in this book include two basic mechanisms for addressing this problem. The first is the use of eligibility traces, and the second is the use of TD methods to learn value functions from which nearly immediate evaluations of actions can be extracted. Both of these methods correspond to similar mechanisms proposed in theories of animal learning.

Pavlov pointed out that every stimulus must leave a trace in the nervous system that persists for some time after the stimulus ends. He regarded learning when there is a temporal gap between the CS offset and the US onset as dependent on stimulus traces, and to this day conditioning under these conditions is called *trace conditioning* (Figure 12.1). Assuming a trace of the CS remains when the US arrives, learning occurs through the simultaneous presence of the trace and the US. We discuss some proposals for trace mechanisms in the nervous system in Chapter 13.

Stimulus traces were also proposed as a means for bridging the time interval between actions and consequent rewards or penalties in instrumental conditioning. In Hull's influential learning theory, for example, "molar stimulus traces" accounted for what he called an animal's *goal gradient*, a description of how the maximum strength of an instrumentally conditioned response decreases with increasing delay of reinforcement (Hull, 1932, 1943). Hull hypothesized that an animal's actions left internal stimuli whose traces decayed exponentially as functions of time since an action was taken. Looking at the animal learning data available at the time, he hypothesized that the traces effectively reach zero after 30 to 40 seconds.

The eligibility traces used in the algorithms described in this book are like Hull's traces: they are decaying traces of past state visitations, or of past state-action pairs. Eligibility traces were Introduced by Klopf (1972) in his neuronal theory in which they are temporally-extended traces of past activity at synapses, the connections between neurons. Klopf's traces are more complex than the exponentially-decaying traces our algorithms use, and we discuss this more when we take up his neuronal theory in Chapter 13.

To account for goal gradients that extend over longer longer time periods than spanned by stimulus traces, Hull (1943) proposed that longer gradients result from secondary reinforcement passing backwards from the goal, a process acting in conjunction with his molar stimulus traces. Animal experiments showed that if conditions favor the presence of secondary reinforcement during a delay period, learning does not decrease with increased delay as much as it does under conditions that obstruct secondary reinforcement. The more favorable the conditions are for secondary reinforcement, the slower this decrease will be and the longer it will take the gradient to fall to zero. He therefore envisioned that there is a primary gradient based on the delay of the primary reinforcement mediated by stimulus traces, and that this is progressively modified, and lengthened, by secondary reinforcement.

Algorithms presented in this book that use both eligibility traces and value functions to enable learning with delayed reinforcement correspond to Hull's hypothesis about how animals are able to learn under these conditions. The actor-critic architecture discussed in Section 10.1 illustrates this correspondence most clearly. The critic uses a TD algorithm to learn a value function associated with the system's current behavior, that is, to predict the current policy's return. The actor updates the current policy based on the critic's predictions, or more exactly, in changes in the critic's predictions. The TD error produced by the critic acts as a secondary reward signal for the actor, providing an immediate evaluation of performance even when the primary reward signal itself is considerably delayed. Algorithms that estimate action-value functions, such as Q-learning and Sarsa, similarly use TD learning principles to enable learning with delayed reinforcement by means of secondary reinforcement. The close parallel between TD learning and the activity of dopamine producing neurons that we discuss in Chapter 13 lends additional strength to the consistency between reinforcement learning algorithms and this aspect of Hull's influential learning theory.

13.5 Cognitive Maps

Model-based reinforcement learning algorithms use environment models that have elements in common with what psychologists call *cognitive maps.* Recall from our discussion of planning and learning in Chapter 8 that by an environment model we mean anything that an agent can use to predict how its environment will respond to its actions in terms of state transitions and rewards, and by planning we mean any process that computes a policy from such a model. Environment models consist of two parts: the state-transition part encodes knowledge about the effect of actions on state changes, and the reward model part encodes knowledge about the reward signals expected for each state (or, more generally, for each state-action-next state triple). To decide on an action a model-based algorithm uses the model to predict the consequences of actions in terms of future states and the rewards expected from them. The simplest kind of planning is to compare the predicted consequences of a collection of "imagined" sequences of decisions.

Questions about whether or not animals use environment models, and if so, how these models are learned, have played influential roles in the history of animal learning research. Some researchers challenged the then-prevailing stimulus-response (S–R) view of learning and behavior, which corresponds to the simplest model-free way to learn policies, by demonstrating *latent learning.* In the earliest latent learning experiment, two groups of rats were run in a maze. For the experimental group, there was no reward during the first stage of the experiment, but food was suddenly introduced into the goal-box of the maze at the start of the second stage. For the control group food was in the goal-box throughout both stages. The interest was in whether or not rats in the experimental group had learned anything during the first stage in the absence of food reward. Although these rats did not *appear* to learn much during the first, unrewarded, stage, as soon as food was introduced in the second stage, they rapidly caught up with the rats in the control group. It was concluded that "during the non-reward period, the rats [in the experimental group] were developing a latent learning of the maze which they were able to utilize as soon as reward was introduced" (Blodgett, 1929).

Latent learning is most closely associated with the psychologist Edward Tolman, who interpreted this and results like it as showing that animals could learn a cognitive map in the absence of rewards or penalties, and that they used the map later when they were motivated to reach a goal (Tolman, 1948). A cognitive map could also allow a rat to plan a route to the goal that was different from the one it it had used in its initial exploration. Explanations of results like these led to the enduring controversy lying at the heart the behaviorist/cognitive dichotomy in psychology. In modern terms, cognitive maps are not restricted to models of spatial layouts but are more generally environment models, or models of an animal's "task space" (e.g., Wilson, Takahashi, Schoenbaum, and Niv, 2014). The cognitive map explanation of latent learning experiments is analogous to the claim that animals use model-based algorithms and that environment models can be learned without explicit rewards or penalties. Models are then used for planning when the animal is motivated by the appearance of rewards or penalties.

Tolman's account of how animals learn cognitive maps was that they learn stimulus-stimulus, or S–S, associations by experiencing successions of stimuli as they explore an environment in the absence of reward. In psychology this is called *expectancy theory*: given S–S associations, the occurrence of a stimulus generates an expectation about the stimulus to come next. This is much like what control engineers call *system identification*, in which a model of a system with unknown dynamics is learned from labeled training examples. In the simplest discrete-time versions, training examples are S–S$'$ pairs, where S is a state and S$'$, the subsequent state, is the label. When S is observed, the model creates the "expectation" that S$'$ will be observed next. Models more useful for planning involve actions as well, so that examples look like SA–S$'$, where S$'$ is expected when action A is executed in state S. It is also useful to learn how the environment generates rewards. In this case, examples are of the form S–r or SA–r, where r is a reward signal associated with S or the SA pair. These are all forms of supervised learning by which an agent can acquire cognitive-like maps in the absence of reward signals. Researchers have also proposed that environment models can be learned via Bayesian methods that extract environment structure from the statistics of varied experiences through what is more like an unsupervised learning process.

13.6 Habitual and Goal-Directed Behavior

The distinction between model-free and model-based reinforcement learning algorithms corresponds to the distinction psychologists make between *habitual* and *goal-directed* control of behavior. Habits are behavior patterns triggered by appropriate stimuli and then performed more-or-less automatically, whereas goal-directed behavior is purposeful in the sense that it is controlled by knowledge of the value of goals and the relationship between actions and their consequences. Habits are said to be controlled by antecedent stimuli, whereas goal-directed behavior is said to be controlled by its consequences. Goal-directed control has the advantage that it can rapidly change an animal's behavior when the environment changes how it responds to the animal's actions. While habitual behavior can produce rapid responses, it is unable to quickly adjust when the environmental contingencies change. The development of goal-directed behavioral control was likely a major advance in the evolution of animal intelligence.

Figure 12.5 illustrates the difference between model-free and model-based decision strategies in a hypothetical task requiring a rat to navigate a maze with distinctive goal boxes, each having an associated reward of the magnitude shown (Panel a). Starting at S_1, the rat has to first select left (L) or right (R) and then has to choose between L and R again at S_2 or S_3 to reach one of the goal boxes. The goal boxes are the terminal states of the rat's episodic task. A model-free strategy (Panel b) relies on stored (cached) values for state-action pairs. These action values (Q-values) are estimates of the return expected for each action taken from each (nonterminal) state. They are obtained over many trials of running the maze from start to finish. To make decisions the rat just has to select at each state the action with the largest action value for that state. In this case, when the action-value estimates become accurate enough, the rat selects L from S_1 and R from S_2 to obtain the maximum return of 4. Alternatively, a model-free strategy might simply rely on a cached policy instead of action values, making direct links from S_1 to L and from S_2 to R. In neither case do decisions rely on an environment model. There is no need to consult a state-transition model, and no connection is required between the features of the goal boxes and the rewards they deliver.

Figure 12.5 Panel (c) illustrates a model-based strategy. It uses an environment model consisting of a state-transition model and a reward model. The state-transition model is shown as a decision tree, and the reward model associates the distinctive features of the goal boxes with the rewards to be found in each. (The rewards associated with states S_1, S_2, and S_3 are also part of the reward model, but here they are zero and are not shown.) A model-based agent can decide which way to turn at each state by using the model to simulate possible action choices to find a path yielding the highest return. In this case the return is the reward obtained from the outcome at the end of the path. Here with a sufficiently accurate model, the rat would select L and then R to obtain reward of 4. Comparing the predicted returns of simulated paths is a simple form of planning, which can be done in a variety of ways as discussed in Chapter 8.

For a model-free agent to change its behavior when its environment changes how it responds to actions, the agent has to acquire new experience in the changed environment during which it can update its policy and/or value function. In the model-free strategy shown in Panel (b) of Figure 12.5, for example, if one of the goal boxes were to somehow shift to delivering a different reward, the rat would have to traverse the maze, possibly many times, to experience the new reward upon reaching that goal box, while updating either its policy or its action-value function (or both) based on this experience. The key point is that for a model-free agent to change the action its policy specifies for a state, or to change an action value associated with a state, it has to act—possibly many times—in that state and experience its actions' consequences.

A model-based agent can accommodate changes in its environment without this kind of 'personal experience' with the states and actions affected by the change. A change in its model automatically (through planning) changes its policy. While new experience in an altered environment is one way that a model can change, it is not the only way. Just observing the environment, or observing the activities of

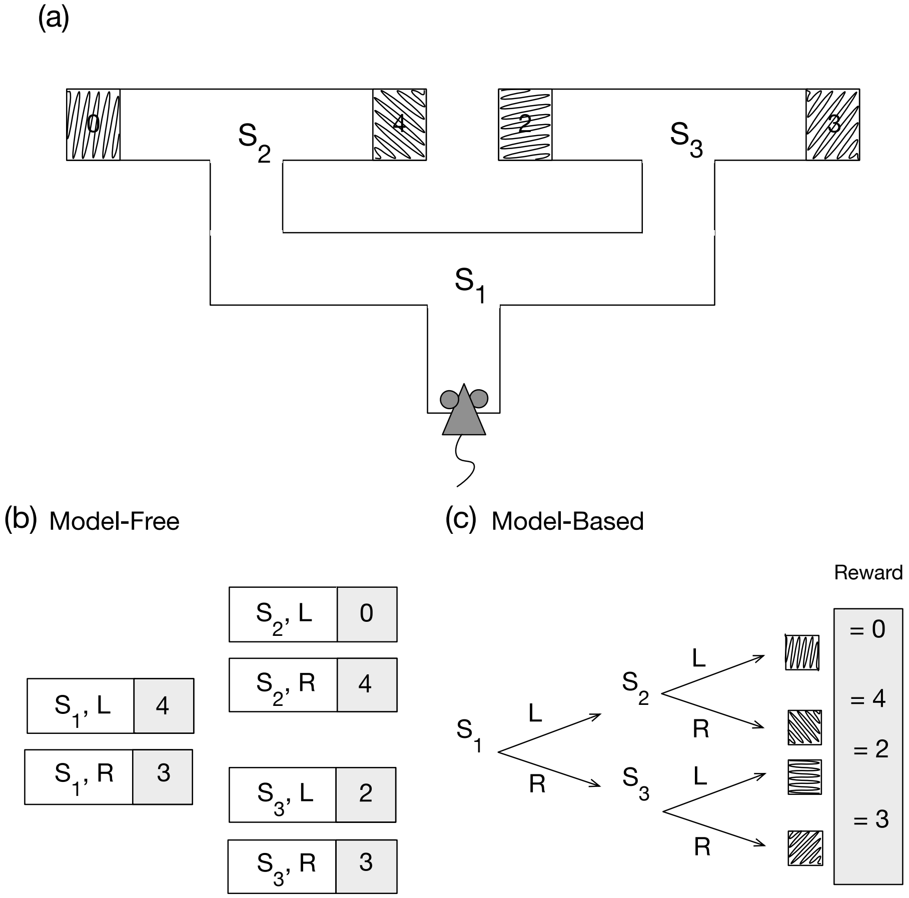

Figure 13.5: Model-based and model-free strategies to solve a hypothetical sequential action-selection problem. (a) A rat navigates a maze with distinctive goal boxes, each associated with a reward having the value shown. (b) A model-free strategy relies on stored (cached) action values for all the state-action pairs obtained over many learning trials. These are estimates of the return to be obtained if that action is taken from that state. To make decisions the rat just has to select at each state the action with the largest action value for that state. (c) In a model-based strategy, the rat learns an environmental model, consisting of knowledge of state-action-next state transitions and a reward model consisting of knowledge of the the reward associated with each distinctive goal box. The rat can decide which way to turn at each state by using the model to simulate possible action choices to find a path yielding the highest return. Adapted from Figure 1 of Niv, Joel, and Dayan (2006), permission pending..

other agents in the environment, is enough for a model-learning process to revise a model to account for these observations. Moreover, planning can determine the consequences of environmental changes that have never been linked together in the agent's own experience. For example, again referring to the maze task of Figure 12.5, the rat's reward model will change if it learns to associate one of the goal boxes with a different reward in circumstances that do not involve the action choices required to find that goal box in the maze. The planning process will bring knowledge of the new reward to bear on maze running without the need for additional experience in the maze.

Exactly this logic is the basis of *goal devaluation experiments* with animals. Results from these experiments provide insight into whether an animal has learned a habit or if its behavior is under goal-directed control. Reward-devaluation experiments are like latent-learning experiments in that the reward signaling changes from one stage to the next, but they are designed to provide finer-grained information about the animal's mode of behavior. After an initial rewarded stage of learning, the reward value of an outcome is decreased, including being shifted to zero or even to a negative value.

The first experiment of this type was conducted by Adams and Dickinson (1981). They trained rats via instrumental conditioning until the rats energetically pressed a lever for food pellets in a training chamber. The rats were then placed in the same chamber with the lever retracted and allowed non-contingent food, meaning that pellets were made available to them independently of their actions. After 15-minutes of this free-access to the food, rats in one group were injected with the nausea-inducing poison lithium chloride. This was repeated for three sessions, in the last of which none of the injected rats consumed any of the non-contingent pellets, indicating that the reward value of the pellets had been decreased—they had been devalued. In the next stage taking place a day later, the rats were again placed in the chamber and given a session of extinction training, meaning that the response lever was back in place but disconnected from the pellet dispenser so that pressing it did not release pellets. Finally, the lever was reconnected to the pellet dispenser to compare how the rats in the two groups would reacquire lever pressing. The results were that the injected rats had significantly lower response rates than the non-injected rats *right from the start of the extinction trials*, and unlike the non-injected rats, they did not reacquire lever pressing in the last stage.

Adams and Dickinson concluded that the injected rats associated lever pressing and consequent nausea, and hence in the extinction trials they "knew" that the consequences of pressing the lever would be something they did not want, and so they reduced their lever pressing right from the start. The important point is that they reduced lever pressing without ever having experienced lever pressing directly followed by being sick: no lever was present when they were made sick. They seemed able to combine knowledge of the outcome of a behavioral choice (pressing the level will be followed by getting a pellet) with the reward value of the outcome (pellets are to be avoided) and hence could alter their behavior accordingly. Not every psychologist agrees with this "cognitive" account of experiments like this, and it is not the only possible way to explain these results, but the model-based planning explanation is widely accepted.

Nothing prevents an agent from using both model-free and model-based algorithms, and there are good reasons for doing this. We know from our own experience that with enough repetition, goal-directed behavior tends to turn into habitual behavior. Experiments show that this happens for rats too. In one recent example Smith and Graybiel (2013) conducted a reward-devaluation experiment in which, simplifying a bit, two groups of rats were trained to run a T-maze, each rat starting at the start gate and turning right or left depending on an auditory cue. If they followed the cue correctly they found food reward at the end of the arm, otherwise there was no reward. Rats in one group were trained until they made the correct turn about 75% of the time, while rats in another group—the overtrained group—received additional training until they were correct about 90% of the time. After this training, the reward was devalued by allowing each rat access to the reward in its home cage and then making it sick by an injection of lithium chloride. After devaluation, the rats were again placed at the start gate of the maze to see what they would do when they heard the instruction cue, but now there was no reward at the end of either arm. The rats that were not overtrained reduced their running to the devalued arm by about 50%. Overtrained rats, on the other hand, continued running to the devalued arm as they had done before undergoing the devaluation procedure. This result suggests that while the non-overtrained rats were acting in a goal-directed manner sensitive to their knowledge of the outcome of their actions, the overtrained rats had developed a habit of running to the instructed arm: their behavior had become insensitive to the reward devaluation.

Viewing this and other results like it from a computational perspective provides insight as to why one might expect animals to behave habitually in some circumstances, in a goal-directed way in others, and why they shift from one mode of control to another as they continue to learn. While animals undoubtedly use algorithms that do not exactly match those we have presented in this book, one can gain insight into animal behavior by considering the tradeoffs that various reinforcement learning algorithms imply.

An idea developed by computational neuroscientists Daw, Niv, and Dayan (2005) is that animals use both model-free and model-based processes. Each process proposes an action, and the action chosen for execution is the one proposed by the process judged to be the more trustworthy of the two as determined by measures of confidence that are maintained throughout learning. Early in learning the planning process of a model-based system is more trustworthy because it chains together short-term predictions which can become accurate with less experience than cached long-term predictions of the model-free process. But with continued experience, the model-free process becomes more trustworthy because planning is prone to making mistakes due to model inaccuracies and short-cuts necessary to make planning feasible, such as various forms of tree-pruning. According to this idea one would expect a shift from goal-directed behavior to habitual behavior as more experience accumulates. Other ideas have been proposed for how animals arbitrate between goal-directed and habitual control, and both behavioral and neuroscience research continues to examine this and related questions.

The distinction between model-free and model-based algorithms is proving to be useful for this research. One can examine the computational implications of these algorithms in abstract settings that expose basic advantages and limitations of each type. This serves both to suggest and to sharpen questions that guide the design of experiments necessary for increasing psychologists' understanding of habitual and goal-directed behavior.

13.7 Extrinsic and Intrinsic Motivation

Psychologists distinguish between *extrinsic motivation*, which means doing something because of some specific rewarding outcome, and *intrinsic motivation*, which refers to doing something "for its own sake." Intrinsic motivation leads organisms to engage in exploration, play, and other behavior driven by curiosity in the absence of externally-supplied rewards. Something like this distinction exists for reinforcement learning systems.

The usual way to apply reinforcement learning to problem solving is to provide the agent with a reward signal determined by the degree of success it has in solving the problem. For example, in the tic-tac-toe illustration in Chapter 1, the agent receives a reward of 1 when it wins a game and otherwise receives a reward of 0. This way of generating a reward signal makes the agent behave like an extrinsically motivated animal. Reward signaling is set up to get the agent to solve a particular problem. This reward signal is like giving a shiny sticker to a student or a tasty treat to a pet when they perform well at something you want them to do.

In contrast to this are ways of generating reward signals that make an agent behave more like it is intrinsically motivated. One example is the "exploration bonus" described in Chapter 8. Instead of being tied to a specific task, this kind of reward signal encourages exploration in general, that is, out of the context of any specific task. Another example is the proposal by Schmidhuber (1991a, b) for how something like curiosity would result from defining reward signals in a certain way. He proposed a reinforcement learning agent that contains a module using supervised learning to learn a model of the agent's environment. Prediction errors both drive model learning and generate reward signals that the reinforcement learning module uses to learn a policy for directing the agent's actions. In particular, positive reward signals are generated to the extent that prediction errors decrease over time. This means that the agent will prefer experiences that enable it to improve its environment model, which implies that it will try to avoid regions of the state space where learning to predict is difficult or where it has already learned to make accurate predictions. As a consequence, the preferences of this "curious agent" will continue to change as it improves its predictive model, attempting to remain in regions where this improvement is most rapid. One might say that the agent is intrinsically motivated to efficiently learn an environment model. Here, again, this is not a particular task posed by an outside entity but rather a general one that, in some sense, the agent is posing for itself.

Reinforcement learning algorithms “don’t care” how reward signals are generated. They do not have to be the result of something in the agent’s external environment that approves of, or disapproves of, the agent’s behavior. They can depend on a wide range of information maintained by the device whose behavior the agent is controlling, including memories of past sensations, actions, and rewards; representations of goals and the state of progress in achieving them; and even the agent’s current policy, value function estimate, and environment model.

Psychologists view Intrinsically motivated behavior as an essential part of an animal’s developmental process. The benefits an animal derives from intrinsically motivated behavior unfold over the long term of its lifetime. When young animals, including humans, play and explore, they are learning skills they will need in many different contexts throughout their lives. The same is true for reinforcement learning systems that are expected to face many different problems over extended periods of time. Considerable research is being devoted to developing analogs of intrinsic motivation for reinforcement learning agents, some of which we cite in Section 12.9 below.

13.8 Summary

Our goal in this chapter has been to discuss correspondences between reinforcement learning and the experimental study of animal learning in psychology. We emphasized at the outset that reinforcement learning as described in this book is not intended to model details of animal behavior. It is an abstract computational framework that explores idealized situations from the perspective of artificial intelligence and engineering. But many of the basic reinforcement learning algorithms were inspired by psychological theories, and in some cases, these algorithms have in fact formed the basis of animal learning models. This chapter describes the most conspicuous of these correspondences.

The distinction in reinforcement learning between algorithms for prediction and algorithms for control parallels animal learning theory’s distinction between classical, or Pavlovian, conditioning and instrumental conditioning. The key difference between instrumental and classical conditioning experiments is that in the former the reinforcing stimulus is contingent upon the animal’s behavior, whereas in the latter it is not. Learning to predict via a TD algorithm corresponds to classical conditioning, and we described the *TD model of classical conditioning* as one instance in which reinforcement learning principles account for some details of animal learning behavior. This model generalizes the influential Rescoral-Wagner model by including the temporal dimension where events within individual trials influence learning, and it provides an account of secondary conditioning, where predictors of reinforcing stimuli become reinforcing themselves. It also is the basis of an influential view of the activity of dopamine neurons in the brain, something we take up in Chapter 13.

Learning by trial and error is at the base of the control aspect of reinforcement learning. We gave some details about Thorndike's experiments with cats and other animals that led to his *Law of Effect*, which we discussed here and in Chapter 1. We pointed out that in reinforcement learning exploration need not be limited to "blind groping"; trials can be generated by sophisticated methods using innate and previously learned knowledge as long as there is *some* exploration. The sets $A(s)$ specifying the actions available in a state s, correspond to the an animal's repertoire of responses to a given situation, what Thorndike called"instinctual impulses." We discussed the training method B. F. Skinner called *shaping* in which reward contingencies are progressively altered to train an animal to successively approximate a desired behavior. Shaping is not only indispensable for animal training, it is also an effective tool for training reinforcement learning agents. There is also a connection to the idea of an animal's motivational state, which influences what an animal will approach or avoid and what it learns from the experience. Both the state of a reinforcement learning agent's environment and the agent's actions influence reward signaling, and the state can include information analogous to an animal's motivational state.

The reinforcement learning algorithms presented in this book include two basic mechanisms for addressing the problem of delayed reinforcement: eligibility traces and value functions learned via TD algorithms. Both mechanisms have antecedents in theories of animal learning. Eligibility traces are similar to stimulus traces of early theories, and value functions correspond to the role of secondary reinforcement in providing nearly immediate evaluative feedback.

The next correspondence the chapter addressed is that between reinforcement learning's *environment models* what and psychologists sometimes call "cognitive maps." Experiments in the mid 20th century challenged the then prevailing S–R view of animal learning by purporting to demonstrate *latent learning*: learning in the absence of rewards or penalties a cognitive map which the animal later uses to guide its behavior when rewards or penalties are introduced. Environment models in reinforcement learning are like cognitive maps in that they can be learned by supervised learning methods without the need for reward signals, and then they can be used to plan behavior in order to obtain reward.

Reinforcement learning's distinction between *model-free* and *model-based* algorithms corresponds to the distinction in psychology between *habitual* and *goal-directed* behavior. Model-free algorithms make decisions by accessing information that has been cached in a policy or an action-value function, whereas model-based methods select actions as the result of planning ahead using a model of the agent's environment. Goal-devaluation experiments provide information about whether an animal's behavior is habitual or under goal-directed control. Reinforcement learning theory has helped clarify thinking about these issues.

The final correspondence addressed in this chapter centers on psychologists' distinction between *extrinsic motivation*, which refers to doing something to achieve some specific rewarding outcome, and *intrinsic motivation*, which refers to doing something "for its own sake." A distinction like this exists for reinforcement learning systems. Exploration bonuses and reward signals linked to improvements in an environmental model are examples of reward signals that make an agent behave as if it were intrinsically motivated to explore or to model its environment. Unlike reward signals based on the success, or lack thereof, in performing a specific task, these kinds of signals encourage behavior that is useful for many different tasks the agent may encounter during its "lifetime." Psychologists view Intrinsically motivated behavior as essential to an animal's development, and the same reasoning is relevant to efforts to develop artificial learning systems that are expected to operate over long periods of time in varying situations.

13.9 Conclusion

In this chapter we discussed how some concepts and algorithms from reinforcement learning correspond to theories of animal learning from psychology. We emphasized that reinforcement learning as developed here explores idealized situations from the perspective of an artificial intelligence researcher or engineer—not from the perspective of an animal learning researcher. As a type of machine learning, the goal of reinforcement learning is not to replicate animal behavior but to design and understand effective learning algorithms. Animal learning clearly informs much of our perspective, but we have been selective in connecting with psychological studies of animal learning. We placed priority on aspects of animal learning that relate in clear ways to methods for solving prediction and control problems. As a result of this selectivity, we did not venture into many of the behavioral details and controversies that have occupied the attention of animal learning researchers. As research in computational reinforcement learning continues, it is likely that further development and refinement of theory and algorithms will be inspired by some of these details, but only to the extent that their computational significance becomes apparent.

Despite this purposeful detachment from psychology, reinforcement learning as developed here is giving back to psychology. The TD model of classical conditioning is one instance where computational principles led to a theory of some details of animal behavior that other theories have considerable difficulty explaining. The examination of habitual and goal-directed behavior in terms of model-free and model-based reinforcement learning algorithms is another case where computational principles from reinforcement learning are informing psychological theories. There are other instances—too numerous to have covered here—where reinforcement learning has suggested fresh ways to think about animal decision making and learning.

We have not been surprised by the fruitful two-way flow of ideas between reinforcement learning and psychology. Of all the paradigms of machine learning, reinforcement learning addresses problems that are the most like those that animals have to face in their natural environments. Without a doubt, the supervised and unsupervised learning paradigms are abstractions of important aspects of animal learning, but they do not encompass the whole problem of learning how to interact with a dynamic environment in order to achieve goals. For its part, reinforcement learning also does not encompass this *whole* problem, but it includes some of its essential elements that are missing from other paradigms. Reinforcement learning's focus on these essential elements of animals' natural problems is the main reason behind the many correspondences we have discussed in this chapter. If the problems that animals face are well modeled as stochastic sequential decision problems—as we think they are—it would be surprising if effective algorithms bore no relationship to the methods that have evolved enabling animals to deal with the problems they face over their lifetimes.

As a final comment, we note that the correspondences between reinforcement learning and animal learning described in this chapter center on behavior observed in laboratory settings rather than in the "wild" of an animal's natural environment. Most of animal learning theory addresses data from laboratory experiments because this setting allows intricate control of conditions and relative ease of observation. Learning principles uncovered in the laboratory hold in natural settings as well, but ethologists and behavioral ecologists focus on ecological and evolutionary aspects of behavior: how animals relate to one another and to their physical surroundings, and how their behavior contributes to evolutionary fitness. Many features of our treatment of reinforcement learning also correspond to the perspective and methods of ethology and behavioral ecology. Optimization, MDPs, and dynamic programming figure prominently in these fields, and our emphasis on agent interaction with dynamic environments connects to the study of agent behavior in complex "ecologies." We do not address agent-agent interaction because we do not cover multi-agent reinforcement learning in this book, but this area has connections to how animals interact with one another. Furthermore, reinforcement learning should by no means be interpreted as dismissing evolutionary perspectives. Nothing about reinforcement learning implies a *tabula rasa* view of learning and behavior. Indeed, experience with engineering applications has highlighted the importance of building into reinforcement learning systems knowledge that is analogous to what evolution provides to animals.

Bibliographical and Historical Remarks

Shah (2012) discusses connections between psychology and reinforcement learning in a review that is a useful companion of this chapter.

The idea built into the Rescorla-Wagner model that learning occurs when animals are surprised is derived from Kamin (1969). Other models of classical conditioning include the models of Klopf (1988), Grossberg (1975), Mackintosh (1975), Moore and Stickney (1980), and Pearce and Hall (1980). Courville, Daw, and Touretzky (2006) present a Bayesian perspective of classical conditioning, and Gershman and Niv (2010) review experimental and theoretical research relating Bayesian structure learning to classical conditioning. An excellent overview of computational models of classical conditioning is provided by Schmajuk (2008).

Blocking in classical conditioning was first reported by Kamin (1968) and is commonly known as Kamin blocking. Moore and Schmajuk (2008) provide an excellent summary of the blocking phenomenon, the research it stimulated, and its lasting influence on animal learning theory.

An early version of the TD model of classical conditioning appeared in Sutton and Barto (1981), which also included the early model's prediction that temporal primacy overrides blocking, later shown by Kehoe, Scheurs, and Graham (1987) to occur in the rabbit nictitating membrane preparation. Sutton and Barto (1981) contains the earliest recognition of the near identity between the Rescorla-Wagner model and the Least-Mean-Square (LMS), or Widrow-Hoff, learning rule (Widrow and Hoff, 1960). This early model was revised following Sutton's development of the TD algorithm (Sutton, 1984, 1988) and was first presented as the TD model in Sutton and Barto (1987) and more completely in Sutton and Barto (1990).

Additional exploration of the TD model and its possible neural implementation was conducted by Moore and colleagues (Moore, Desmond, Berthier, Blazis, Sutton, and Barto, 1986; Moore and Blazis, 1989; Moore, Choi, and Brunzell, 1998; Moore, Marks, Castagna, and Polewan, 2001). Klopf's (1988) drive-reinforcement theory of classical conditioning extends the TD model to address additional experimental details, such as the S-shape of acquisition curves. Ludvig, Sutton, and Kehoe (2012) evaluated the performance of the TD model in previously unexplored tasks involving classical conditioning and examined the influence of various stimulus representations, including the microstimulus representation that they introduced earlier (Ludvig, Sutton, and Kehoe, 2008). In some of these publications TD is taken to mean Time Derivative instead of Temporal Difference.

Section 1.7 includes comments on the history of trial-and-error learning and the Law of Effect. Peterson's essay on Skinner's discovery of shaping highlights the influence this discovery had on his subsequent research (Peterson, 2004). Selfridge, Sutton, and Barto (1985) illustrated the effectiveness of shaping in the pole-balancing reinforcement learning task. Because a long pole is easier to balance than a short pole, their system started with a long pole whose length was incrementally decreased while learning took place until it was the shorter length required by the task. Overall learning time was significantly less than when the pole length was fixed at this shorter length throughout learning. Other examples of shaping in reinforcement learning are provided by Mahadevan and Connell (1992), Mataric (1994), Dorigo and Colombette (1994), Saksida, Raymond, and Touretzky (1997), and Randløv and Alstrøm (1998). Ng (2003) and Ng, Harada, and Russell (1999) use the term shaping in a sense somewhat different from Skinner's, focussing not on successive approximation but on the problem of how to alter the reward signal without altering the set of optimal policies.

Spence, Hull's student and collaborator at Yale, elaborated the role of secondary reinforcement in addressing the problem of delayed reinforcement (Spence, 1947). Learning over very long delays under conditions that rule out secondary reinforcement, as in taste-aversion conditioning with delays up to several hours, led to interference theories, described by Revusky and Garcia (1970). According to these theories credit can be assigned to actions taken very far in the past if there were no, or few, relevant intervening stimuli or actions to interfere with the process. Boakes and

Costa (2014) thoroughly review data related to the delay-of-reinforcement problem, particularly data supporting interference theories. Johanson, Killeen, Russell, Tripp, Wickens, Tannock, Williams, and Sagvoldenet (2009) discuss interference theories in the context of a hypothesis that attention deficit hyperactivity disorder (ADHD) results from disruption of the ability to assign credit for delayed reinforcement. Seo, Barraclough, and Lee (2007) report that the activity of neurons in the prefrontal cortices of rhesus monkeys is modulated by previous action choices, leading them to suggest that working memory is involved in handling delayed reinforcement

Thistlethwaite (1951) is an extensive review of latent learning experiments up to the time of its publication. Ljung (1998) provides an overview of model learning, or system identification, techniques in engineering, and Gopnik, Glymour, Sobel, Schulz, Kushnir, and Danks (2004) discuss a Bayesian theory about how children learn models.

Connecting habitual and goal-directed behavior respectively to model-free and model-based reinforcement learning was first proposed by Daw, Niv, and Dayan (2005). Dolan and Dayan (2013) provide a comprehensive review of four generations of experimental research related to this issue and discuss how it can move forward on the basis of the model-free/model-based distinction. Dickinson (1980, 1985) and Dickinson and Balleine (2002) discuss in detail experimental evidence related to this distinction. Donahoe and Burgos (2000) alternatively argue that model-free processes can account for the results of goal revaluation experiments.

In addition to discussing habitual and goal-directed modes of behavioral control, Dayan (2008) uses the term Pavlovian control to refer to behavior programmed by evolution to deal with particular appetitive or aversive outcomes. He also introduces the term episodic control to refer to a primitive process of simply repeating sequences of actions that have been successful in the past. Dayan and Berridge (2014) argue that classical conditioning involves model-based processes. Rangel, Camerer, and Montague (2008) review many of the outstanding issues involving habitual, goal-directed, and Pavlovian modes of control.

Dickinson and Balleine (2002) discuss the relationship between learning and motivation, revealing the complex nature of the interaction. Though focussing on neuroscience, Wise (2004) provides an overview of reinforcement learning and its relation to motivation. Daw and Shohamy (2008), while also addressing neuroscience issues, link motivation and learning to aspects of reinforcement learning theory. Niv, Joel, and Dayan (2006) proposed that motivation can be thought of as the mapping that assigns numerical reward signals to objects or events observed in an animal's external environment. As input to this mapping, an animal's motivational state is an 'index' of different ways of assigning reward signals to experiences. McClure, Daw, and Montague (2003) suggest that *incentive salience* in the theory of Berridge and Robinson (1998) is the expected future reward, that is, that it is given by a value function.

McClure et al. (2003) present a theory of behavioral vigor in which the time taken for an animal to select an action depends on the distribution of action-values: an animal will behave more vigorously to the extent that possible actions have high values; low vigor is the result of an absence of actions with high values. Niv, Daw, and Dayan (2005) use continuous-time MDPs to suggest a normative account of behavioral vigor where decisions involve choosing an action together with a latency with which it is executed. See also Niv et al. (2007).

The influence of internal state on valuation is discussed by Rangel et al. (2008) and Dayan and Berridge (2014). In reviewing two neuroscience studies of the influence of internal state on dopamine signaling, Burke, Dreher, Seymour, and Tobler (2014) mention the influence of financial status and expectations on valuation, referring to Bernoulli's 1738 paper that laid the foundation for utility theory (Bernoulli, 1954, is an English translation) and to prospect theory (Kahneman and Tversky, 1979).

Ryan and Deci (2000) provide a psychological introduction to intrinsic motivation. Barto, Singh, and Chentanez (2004) and Singh, Barto, and Chentanez (2004) introduced the term intrinsically motivated reinforcement learning in the context of hierarchical reinforcement learning. The first instance we know of that would be characterized as intrinsically motivated reinforcement learning is Schmidhuber's proposal for endowing an agent with a kind of curiosity (Schmidhuber, 1991a, 1991b; Schmidhuber, Storck, and Hochreiter, 1994; Storck, Hochreiter, and Schmidhuber, 1995; Schmidhuber, 2009). This is an active area of research with a growing literature. Baldassarre and Mirolli (2013) is a collection that widely covers the topic, including a chapter by Barto (2013) that explicitly discusses intrinsic motivation in the context of reinforcement learning. Among other notable expositions of these ideas are those of Oudeyer and Kaplan (2007a, 2007b).

Singh, Lewis, Barto, and Sorg (2010) provide an evolutionary perspective on intrinsic motivation by considering what kinds of reward signaling confer analogs of high evolutionary fitness to agents whose learning is guided by those reward functions (see also Singh, Lewis, and Barto, 2009). Sorg, Singh, and Lewis (2010) and Sorg (2011) used this approach to argue that good reward functions for reinforcement learning systems can mitigate a variety bounds under which the system must operate. The argument implies that the designer of a reinforcement learning system should not necessarily make his or her own objective the objective of the learning system.