Reinforcement learning and causal models

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

December 6, 2015

Abstract

This chapter reviews the diverse roles that causal knowledge plays in reinforcement learning. The first half of the chapter contrasts a "model-free" system that learns to repeat actions that lead to reward with a "model-based" system that learns a probabilistic causal model of the environment which it then uses to plan action sequences. Evidence suggests that these two systems coexist in the brain, both competing and cooperating with each other. The interplay of two systems allows the brain to negotiate a balance between cognitively cheap but inaccurate model-free algorithms and accurate but expensive model-based algorithms. The second half of the chapter reviews research on hidden state inference in reinforcement learning. The problem of inferring hidden states can be construed in terms of inferring the latent causes that give rise to sensory data and rewards. Because hidden state inference affects both model-based and model-free reinforcement learning, causal knowledge impinges upon both systems.

KEYWORDS: habits, goals, Markov decision process, structure learning

Introduction

Reinforcement learning (RL) is the study of how an agent (human, animal or machine) can learn to choose actions that maximize its future rewards (Sutton & Barto, 1998). Two strong constraints have shaped the evolution of RL in the brain. On the one hand, the world is complex, favoring the development of rich causal models that can be used to accurately predict future reward. On the other hand, building and using causal models is computationally costly. If an agent needs to act quickly and energy-efficiently, cheaper but less accurate predictions may be required. Algorithms that directly estimate future reward without building an explicit causal model are known as *model-free*, in contrast to *model-based* algorithms that employ a causal model.

To make this distinction concrete, imagine how you might navigate from home to your office. If you just moved to a new house, the route may be unfamiliar and so you rely upon a map to figure out a step-by-step plan. The map is a causal model: It tells you that taking an action (i.e., moving in a particular direction) *causes* a change in your state (i.e., location). Constructing a plan amounts to designing a causal chain that terminates at your intended goal. For this reason, the map-based strategy is a form of model-based control. As you become more familiar with the route, you may

Reward	state, action → reward
Transition	state, action → state
Hidden state	hidden state → observation

Table 1: Summary of causal relationships in reinforcement learning.

find yourself relying less on maps—you simply "know" what direction to go in a particular location. One way this might happen is that you have learned to cache the values of actions in different states, so that you can determine where to go simply by inspecting these cached values. A causal model is not required for this navigation strategy. This is the essence of model-free control.

Experimental work has revealed that humans and animals use a combination of model-based and model-free algorithms, implicating the co-existence of two "systems" in the brain that are at least partially dissociable (Balleine & Dickinson, 1998; Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013). These systems compete for control of behavior, but may also cooperate with each other, as I will discuss later.

The aim of this chapter is to highlight the diverse roles that causal knowledge plays in model-based and model-free RL. I begin with a brief summary of the historical background, and then review the modern computational synthesis of model-based and model-free RL. While it is tempting to view causal knowledge as falling strictly within the purview of model-based RL, this is not the case. Agents must perpetually contend with *partial observability*: sensory data provide imperfect information about the underlying "state" of the environment. It is the hidden state, rather than sensory data, that is causally related to reward. For example, if one smells food cooking at a restaurant and sits down to eat, it is not the smell (sensory data) that caused the food (reward) to appear, but rather the cook who made the food (the hidden state). Both model-based and model-free learning systems employ causal knowledge to form a belief about the hidden state. The second half of this chapter is devoted to a review of research on the role of causal models in dealing with partial observability.

The concept of causality appears in various forms throughout this chapter. Table 1 provides a summary of the three forms of causality that play key roles in RL: Taking an action in a state causes both a reward and a transition to a new state, and in partially observable environments the state generates perceptual signatures (observations). In later sections, I will formalize these ideas and discuss how they have been studied experimentally.

Historical background

The early study of RL was dominated by behaviorism, which explicitly rejected any notion of an internal model. The behaviorist view of learning is succinctly summarized by Thorndike's *law of effect*: if an action leads to reward, it will become more likely to be repeated in the future (Thorndike, 1911). While later computational models posited more complex rules governing behavior, virtually all of them embodied the law of effect (e.g., Mackintosh, 1975; Pearce, 1980; Rescorla & Wagner, 1972). As reviewed in the next section, this characterization also applies to contemporary theories of model-free RL.

Nonetheless, a variety of behavioral phenomena suggest that there exist powerful determinants of responding that cannot be reduced to simple reinforcement. Tolman (1948) described a number of ingenious experiments whose results are perplexing from a behaviorist perspective. For example, Tolman showed that rats could take shortcuts or plan detours around obstacles without ever being reinforced for these actions. Another example described by Tolman is *latent learning*: a rat allowed to explore a maze without reinforcement was subsequently faster at learning to navigate to a reward. Since the rat was not reinforced for its actions during the exploratory phase, this behavior cannot be explained by the law of effect. Later research on contextual fear conditioning revealed a similar phenomenon: brief pre-exposure to a context enhanced the acquisition of contextual fear (Fanselow, 1990; Kiernan & Westbrook, 1993).

Tolman interpreted latent learning and other findings as evidence for a "cognitive map"—an internal model of the environment that encodes information about spatial layout, object attributes, and relations between objects. Several decades after Tolman's pioneering work, the idea of a cognitive map received direct support from recordings in the hippocampus that revealed neurons tuned to an animal's location in space (O'Keefe & Nadel, 1978). Subsequent research showed that the hippocampal cognitive map is replete with representations of landmarks, boundaries, sequences, and relations (Eichenbaum, 2004; Hasselmo, 2012).

Another line of assault on the law of effect was pursued by Dickinson and his colleagues in the early 1980s (Dickinson, 1985). These studies mapped out the conditions under which instrumental behavior is controlled by goals, overriding the actions prescribed by an animal's reinforcement history. For example, rats trained to press a lever for sucrose would subsequently cease lever pressing in an extinction test after the sucrose was separately paired with illness (thereby devaluing the sucrose reinforcer), demonstrating outcome sensitivity consistent with a cognitive map or goal-directed view of instrumental behavior (Adams, 1982). Since the instrumental action (lever pressing) was never directly paired with illness, the law of effect predicts no reduction of responding under these circumstances. Importantly, goal-directed control of behavior could be superseded by stimulus-response habits given enough training. In particular, rats overtrained with the sucrose reinforcer continued to press the lever after the devaluation treatment, demonstrating outcome insensitivity more consistent with a habit learning system governed purely by the law of effect. These observations led to the idea of multiple competing learning systems, which will be discussed further in the next section.

The important point to take away from these studies is that causality is central to a complete understanding of RL in the brain. The cognitive map encodes information about how actions cause changes in state, and the goal-directed nature of instrumental behavior suggests that animals understand the causal effects of their actions on subsequent rewards. These correspond to the first two causal relationships listed in Table 1.

Reinforcement learning theory

Contemporary RL theory has its origins in a family of engineering techniques developed to deal with complex planning and control problems (Bellman, 1957). This section introduces these techniques formally (see Sutton & Barto, 1998, for a thorough introduction), and describes how they have

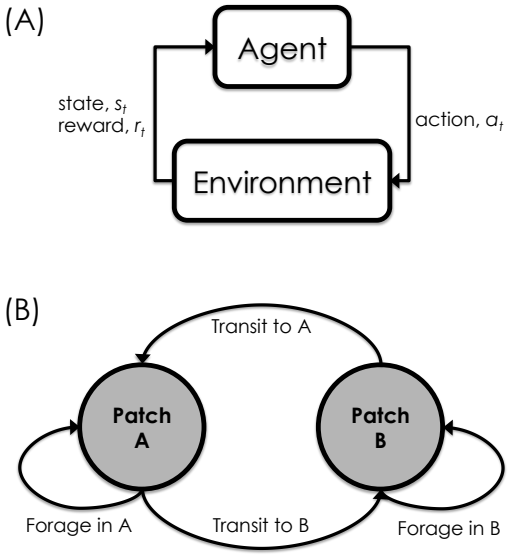

Figure 1: **Reinforcement learning**. (A) The agent-environment interface. (B) Example of a Markov decision process. Circles denote states, and arrows denote deterministic state transitions caused by particular actions.

been used to explain experimental findings from psychology and neuroscience. I will first review some basic notation and concepts, and then discuss several important algorithms for solving the RL problem.

Formalization of the problem

The basic RL problem is summarized in Figure 1A. An agent at time t occupies state s_t, takes action a_t from a policy $\pi(a_t|s_t)$, receives a reward r_t with expected value $\mathcal{R}(s_t, a_t)$ and transitions to a new state s_{t+1} according to a transition distribution $\mathcal{T}(s_{t+1}|s_t, a_t)$. The agent continues interacting with the environment *ad infinitum* (or until it reaches a terminal state), accumulating rewards. For example, consider the simple foraging environment shown in Figure 1B. At each point in time, the forager collects resources at one of two patches and continually chooses whether to stay at one patch or transit to the other patch. In this example, the patches correspond to states, the resources correspond to rewards, and the actions are stay/switch decisions.

One criterion of optimality is the maximization of cumulative reward, or *return*, $\sum_{t=0}^{\infty} r_t$. However, this does not take into account the fact that most biological agents prefer rewards sooner rather than later (Frederick, Loewenstein, & O'donoghue, 2002). This can be captured by assuming that future rewards are discounted exponentially, leading to the following definition of discounted return:

$$R = \sum_{t=0}^{\infty} \gamma^t r_t. \tag{1}$$

The discount factor $\gamma \in [0, 1]$ represents the agent's preference for immediate rewards: lower values of γ indicate a steeper discounting of future rewards.

Because rewards and transitions may be stochastic, and hence R is a random variable, we take the goal of the agent to be maximizing *expected* discounted return, or *value*, defined as:

$$Q(s,a) = \mathbb{E}\left[R \mid s_0 = s, a_0 = a\right], \tag{2}$$

where $\mathbb{E}[\cdot]$ is the expectation operator, returning the average of its arguments (in this case averaging over randomness in states, actions and rewards under a particular policy).To understand this equation, imagine an agent who takes action a in state s and then pursues policy π over an infinitely long trajectory through the state space, meanwhile recording the discounted return. We can imagine the agent restarting this trajectory many times, and then averaging the discounted return recorded on each trajectory (this is known as a "Monte Carlo" approximation). The resulting value $Q(s,a)$ is equivalent to the average of discounted returns over all possible trajectories, weighted by the probability of each trajectory under policy π. The optimal action in state s maximizes $Q(s,a)$:

$$a_s^* = \underset{a}{\operatorname{argmax}}\, Q(s,a). \tag{3}$$

We say that a policy π^* is optimal if it maximizes $Q(s,a)$ for all states. While policies may in general be probabilistic, the optimal policy is always deterministic, with $\pi^*(s,a_s^*) = 1$ and 0 for all other actions. In the foraging example given above, suppose that the expected reward in Patch A is larger than in Patch B; provided γ is sufficiently large, the optimal policy is to always take the "stay" action in Patch A and the "switch" action in Patch B. If γ gets small enough, however, the optimal policy is to stay in Patch B, since switching will result in delayed reward.

The environment described above is known as a *Markov decision process* (MDP) because it obeys the *Markov property*: state transitions and rewards are independent of the agent's history conditional on the current state and action. In the patch foraging example used above, the Markov property says that the probability of transit to another patch depends only on the current patch and the agent's stay/switch decision (likewise for the resource collection at the current patch). The Markov property enables the value function to be expressed recursively:

$$Q(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \sum_{a'} \pi(a'|s') Q(s',a'). \tag{4}$$

This expression is known as the *Bellman equation* (Bellman, 1957). Intuitively, the Bellman equations shows that the value function can be broken down into the immediate reward (first term) and the expected future reward $\mathbb{E}[Q(s',a')]$ (second term). The sum over future states and actions in the second term reflects the agent's uncertainty; in probability theory, this is known as *marginalization.*The optimal Q-value (i.e., the Q-value under the optimal policy π^*) can correspondingly be written as:

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} Q^*(s',a'). \tag{5}$$

Here we have simply substituted $\pi^*(a'|s') = \operatorname{argmax}_{a'} Q(s',a')$ into Eq. 4. The Bellman equation serves as the basis of efficient learning and planning algorithms, which we discuss next.

Algorithmic solutions

Model-based and model-free algorithms can, loosely speaking, be seen as working on different sides of the Bellman equation. Model-based algorithms operate on the right-hand side of the Bellman equation, in the sense that they compute $Q(s, a)$ by directly applying the Bellman equation to the learned reward and transition functions. For example, the value iteration algorithm (Sutton & Barto, 1998) initializes the Q-values randomly and then repeatedly applies Eq. 5 to compute new Q-values for each state-action pair. Value iteration is guaranteed to converge to the optimal Q-value. However, it is intractable for large state and action spaces. For this reason, the most successful modern techniques use some form of local tree search (Browne et al., 2012). These algorithms employ the model as a means of simulating trajectories though the state space around the current state, and estimate Q-values on the basis of these trajectories. While there is evidence that humans carry out something resembling tree search (e.g., De Groot, 1978; Holding & Pfau, 1985; Huys et al., 2012, 2015), our current knowledge about model-based planning in the brain is very limited (see Daw & Dayan, 2014, for further discussion).

Model-free algorithms operate on the left-hand side of the Bellman equation: Instead of learning a model, they directly estimate $Q(s, a)$ from experience and cache these estimates in a look-up table.[1] The most influential class of model-free algorithms is known as *temporal difference (TD) learning* (Sutton, 1988). All TD algorithms have in common the idea that learning is driven by the discrepancy between observed and predicted reward (the prediction error). To understand how TD learning is connected to the Bellman equation, notice that Eq. 5 can be written as an expectation:

$$Q^*(s_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q^*(s_{t+1}, a')\right], \tag{6}$$

where we have replaced the reward and transition functions with sampled rewards (r_t) and states (s_{t+1}) inside the expectation. The expectation can always be approximated by averaging many such samples (cf. the Monte Carlo approximation described in the previous section). This equation implies a consistency condition: If we have appropriately estimated the optimal Q-values, then the difference between $r_t + \gamma \max_{a'} Q^*(s_{t+1}, a')$ and $Q^*(s_t, a_t)$ should, on average, be zero:

$$\delta_t = r_t + \gamma \max_{a'} Q^*(s_{t+1}, a') - Q^*(s_t, a_t), \tag{7}$$

$$\mathbb{E}[\delta_t] = 0. \tag{8}$$

The variable δ_t is precisely the prediction error mentioned above, because it reflects the difference between observed and predicted rewards. What happens if we do not have an accurate estimate of the optimal Q-values (or we are following a sub-optimal policy)? Then the prediction error will, on average, be non-zero. In fact, the direction of the prediction error tells you something important about how to update the Q-values. When the prediction error is positive, the value function has underestimated the expected future reward and therefore the Q-value should be increased; likewise, when the prediction error is negative, the value function has overestimated the expected future reward and therefore the Q-value should be decreased.

[1] In practice, storing values in a look-up table for MDPs with many states is inefficient. For this reason, most algorithms use some form of function approximation (Sutton & Barto, 1998).

This is the essential idea underlying one of the most important TD algorithms, Q-learning (Watkins & Dayan, 1992), which updates an estimate of the optimal value function according to:

$$\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) + \alpha\delta_t \tag{9}$$

$$\delta_t = r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t) \tag{10}$$

where $\alpha \in [0, 1]$ is a learning rate parameter. Although it is still a matter of debate what particular form of TD learning is used by the brain (Niv, 2009), all TD algorithms embody the basic prediction error logic laid out above.

The main reason that TD learning has figured so prominently in neuroscience is that the phasic firing of midbrain dopamine neurons appears to correspond closely with the theoretical prediction error (Bayer & Glimcher, 2005; Glimcher, 2011; Niv & Schoenbaum, 2008; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000). Some of the key evidence comes from Pavlovian conditioning tasks (Schultz et al., 1997), where dopamine neurons fire in response to unexpected reward (e.g., early in learning) but not to expected reward (e.g., late in learning). Furthermore, dopamine neurons fire below baseline when an expected reward is omitted. The prediction error interpretation of dopamine has received support from a wide range of studies, too numerous to review here (see Glimcher, 2011).

The TD model has also played an important role in the development of animal learning theory (Ludvig, Sutton, & Kehoe, 2012; Sutton & Barto, 1990). It can be seen as a "real-time" generalization of the Rescorla-Wagner model (which does not make predictions about intra-trial events), allowing the TD model to explain various phenomena outside the scope of the Rescorla-Wagner model (Rescorla & Wagner, 1972). For example, in trace conditioning, reward is delivered following an unfilled delay after the offset of cue A. Acquisition of a conditioned response is facilitated if another cue (B) is presented during the delay interval (Kehoe, 1982). According to the TD model, this facilitation occurs because cue B acquires positive value, which generates a large positive prediction error at the offset of cue A, thereby providing an amplified learning signal. TD learning provides a similar account of second-order conditioning: when cue A is paired with reward, and subsequently cue B is paired with cue A, cue B acquires the ability to elicit a conditioned response. According to the TD model, the prediction error is positive when cue B is paired with cue A (since cue A has a positive value), and this error signal drives learning of a positive value for cue B (Sutton & Barto, 1990).[2]

Despite these successes, the TD model is still essentially an implementation of Thorndike's law of effect, and hence fails to explain the phenomena discussed in the previous section, such as latent learning and goal-directed control. What is needed, as Tolman pointed out, is a "cognitive map." Model-based RL provides one possible formalization of how a cognitive map can be used to support goal-directed control (see also Reid & Staddon, 1998). Because model-based RL computes values on the fly, rather than retrieving cached estimates, it can immediately and flexibly respond to changes in rewards or transition probabilities, without having to back-propagate the TD error along an unbroken sequence of states.

It is worth noting here that some authors have proposed mechanisms for goal-directed control that

[2]Note that this analysis assumes that the association between B and the absence of reward is not encoded (see Gershman, Blei, & Niv, 2010, for more discussion of this point).

are associative rather than model-based (de Wit & Dickinson, 2009; Elsner & Hommel, 2001). According to these theories, goal-directed control arises from associative links between stimuli, actions, and outcomes. Supporting evidence comes from studies showing that outcomes can activate the representations of actions that have caused the outcomes in the past (e.g., Elsner & Hommel, 2001). While these associative theories are not grounded in the formalism of RL, more recent ideas have begun to bridge the gap. In particular, Stachenfeld, Botvinick, and Gershman (2014) showed that one way to construct a cognitive map is to learn a predictive representation (Dayan, 1993), which is, in essence, an association between current and future states. This predictive representation can then be combined with a reward function to efficiently compute action values. In addition to reproducing some of the behavior typically attributed a model-based system, the predictive representation can capture many aspects of the hippocampal cognitive map. Importantly, the predictive representation is not a causal model of the environment, in the sense that it cannot be given a causal Bayes net interpretation—it does not encode the transition function that governs the causal effect of actions on the environment. Rather, it can be understood as a kind of summary representation of the underlying causal system. Thus, it remains an open question whether goal-directed control requires a system that learns a causal model of the environment and uses it to formulate plans.

Transitions and interactions between the systems

The transition from goal-directed to habitual behavior has been rationalized in terms of uncertainty-based arbitration between model-free and model-based RL (Daw et al., 2005). The idea is that each learning system keeps track of its uncertainty via Bayesian estimation of its values, and the system with lower uncertainty is given control of behavior. In the case of the model-free system, the uncertainty is dominated by the stochasticity of transitions, rewards and actions (all sources of "statistical noise"). In the case of the model-based system, the uncertainty is dominated by "computational noise" induced by finite cognitive resources (e.g., truncation of tree search). Generally speaking, the model-free system requires considerably more experience to suppress its uncertainty to the level of the model-based system. On the other hand, the model-free system is much more computationally efficient, since values can be computed merely by inspecting the look-up table. Thus, the model-based system controls behavior early in learning, when the model-free values are mostly useless; later in learning, the model-free system takes control, when its values become more accurate (statistical noise is reduced through averaging).[3]

The devaluation experiments described above (Adams, 1982; Dickinson, 1985) provide examples of this transition. For an animal that has been moderately trained on an instrumental learning task, the model-based system retains control of behavior (because its values are more accurate than those of the model-free system), and hence instrumental responding is sensitive to reinforcer devaluation. For an extensively trained animal, the model-free system (whose value estimates are now sufficiently accurate) assumes control of behavior, rendering instrumental control insensitive to devaluation.

Various factors can shift the balance between the two learning systems. For example, environments

[3] According to a related account, the transition from model-based to model-free control can be understood in terms of a speed/accuracy tradeoff (Gershman, Horvitz, & Tenenbaum, 2015; Keramati, Dezfouli, & Piray, 2011).

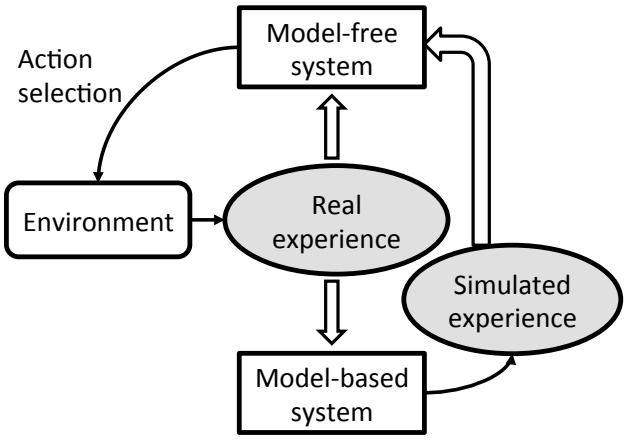

Figure 2: **The Dyna architecture**.

in which the reward and transition probabilities change quickly favor the model-free system (Simon & Daw, 2011). Placing people under working memory load also shifts control to the model-free system, presumably by diverting some of the cognitive resources upon which the model-based system depends (Otto, Gershman, Markman, & Daw, 2013). Concomitantly, working memory capacity predicts the degree to which behavior appears model-based (Otto, Raio, Chiang, Phelps, & Daw, 2013).

So far, the two learning systems have been treated as largely independent, interacting only in their competition for control of behavior. However, competition may not be their only form of interaction. Sutton (1990) proposed that the systems could also interact cooperatively; in this architecture, called *Dyna* (Figure 2), the model-based system was used to produce simulated experience from which the model-free system could then learn. Recently, behavioral evidence for this form of interaction has begun to emerge (Gershman, Markman, & Otto, 2014). For example, Gershman et al. (2014) showed that human subjects can make choices on the basis of model-based knowledge under conditions where the model-free system is ostensibly in control of behavior. The utilization of model-based knowledge by the model-free system can be enhanced by a brief period of quiescence (listening to a piece of classical music), consistent with the idea that the model-based system simulates experience "offline" in the service of model-free learning.

The Dyna architecture also sits well with the observation that the hippocampus appears to simulate spatial trajectories, leading a corresponding simulation in the striatum, the putative seat of model-free learning (Lansink, Goltstein, Lankelma, McNaughton, & Pennartz, 2009). If the model-free and model-based systems interact in this way, it may explain why model-based knowledge infiltrates reward prediction errors measured in the striatum (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), a finding which is perplexing from the perspective of a competitive architecture. Various other possibilities for interactions between the two systems are discussed further in Daw and Dayan (2014).

Causal knowledge and partial observability

Both model-free and model-based learning rely on a representation of state. However, the state representation that is relevant for obtaining rewards is often not the representation furnished by early sensory processing. Rather, the state must be *inferred* from sensory data. Formally speaking, this is a case of *partial observability* (Kaelbling, Littman, & Cassandra, 1998), where an agent only has access to the hidden state via noisy sensory data. If the hidden state obeys the Markov property, then we can call this environment a *partially observable Markov decision process* (POMDP). Bayes' rule can be employed to infer the posterior distribution over hidden states given sensory data, and this posterior distribution functions as a "belief state" in a fully observable MDP over which learning can operate (albeit in a higher-dimensional space). The belief state MDP has the appealing property that all the machinery of the previous section can be applied to this representation.

From hidden states to latent causes

One way to think about hidden state inference is in terms of latent causes: an MDP corresponds to a probabilistic causal model in which states and actions jointly cause rewards, transitions, and sensory data. In the foraging example used above, choosing the "switch" action in Patch A causes a transition to Patch B and the receipt of reward; in a partially observable setting, the action would also cause the observation of sensory information (e.g., entering the patch causes the appearance of a prey type that is informative about which patch has just been entered). Hidden state inference is a form of causal reasoning in this model, and thus shares much in common with causal reasoning in other domains. For present purposes, the important point is that even the ostensibly "model-free" system utilizes these inferential computations, thus further blurring the sense in which such a system is truly model-free. One plausible possibility, suggested by several authors (Daw, Courville, & Touretzky, 2006; Rao, 2010), is that the belief state is computed by cortical circuitry late in the sensory processing stream, and then fed into subcortical circuits responsible for RL. Both model-based and model-free systems, in this scheme, rely on the same belief state representation.

Rao (2010) has offered one neurobiologically detailed proposal for how this might work in the case of simple perceptual decisions about random dot motion. In the reaction-time version of the random dots task (Roitman & Shadlen, 2002), a subject must make a rapid binary decision (left/right) about the motion direction of randomly moving dots, where some fraction (the *coherence*) of the dots are moving in the same direction. By changing the coherence of the dot motion, the experimenter can parametrically adjust the perceived motion strength, and this produces corresponding changes in discrimination accuracy (lower accuracy for low coherence) and response time (longer response times for low coherence). While on the surface the random dots task may not appear like a problem of latent causal inference, it resembles an ecologically valid problem faced by many animals. Imagine, for example, a lion moving through the savannah brush; its camouflage induces a noisy, fluctuating percept, with different points along the surface of the lion bound together by their common motion. The visual system must integrate the noisy motion information to discern the lion's direction of movement, the latent cause generating the sensory information.

According to Rao (2010), motion selective neurons in area MT report the momentary likelihood of sensory data (transmitted from early visual cortex) under different motion directions. The

likelihoods are integrated over time in area LIP to compute the belief state (i.e., the posterior over motion directions), producing a ramping of activity as evidence accumulates (Gold & Shadlen, 2002). The striatum (a part of the basal ganglia) receives inputs from cortical regions (including LIP) and computes the Q-value, which then gets fed into midbrain circuits that compute the prediction error, reported in the form of dopamine release. The dopamine signal drives updating of the value function by modulating plasticity at cortico-striatal synapses (Reynolds & Wickens, 2002). Here the value function is defined over belief states and actions (motion direction judgments, typically registered by a saccadic response).

In addition to explaining how animals could learn to solve the random dots task, Rao's POMDP model offers a functional explanation of dopaminergic responses in the task. Nomoto, Schultz, Watanabe, and Sakagami (2010) found that when dot coherence is 60%, dopamine neurons ramped up their activity, peaking at the time of response. In the POMDP model, this occurs because the value is lowest at the highest entropy belief state (i.e., when the animal is completely uncertain), and increases rapidly as perceptual information reduces the entropy; because the prediction error tracks temporal differences in the value function, this results in the observe ramping pattern.[4]

Structure learning

Any RL system operating in a real-world environment must not only perform hidden state inference, but must also *discover* the hidden states underlying its observations. This is a form of *latent structure learning* (Courville, Daw, & Touretzky, 2006; Gershman & Niv, 2010). In the rest of this section, I will describe several case studies illustrating how structure learning can explain various empirical lacunae that have troubled RL theories.

Consider a Pavlovian fear conditioning experiment, in which a cue is repeatedly paired with an aversive outcome (e.g., a shock). Over the course of training, the cue will come to elicit an innate fear response (freezing, in the case of rat subjects). If the cue is subsequently extinguished, by presenting it repeatedly without a shock, the fear response will subside. If states have a one-to-one mapping with cues, then standard RL theory predicts that extinction produces unlearning of the (negative) value acquired during training. However, this prediction is problematic, because a variety of assays demonstrate that the fear memory persists despite extinction, and will reemerge under certain circumstances (Bouton, 2004). Pavlov (1927) demonstrated that simply presenting the cue again after a retention interval was sufficient to elicit conditioned responding, a phenomenon known as *spontaneous recovery* (Rescorla, 2004). In another procedure, known as *reinstatement*, exposing the subject to an isolated shock before testing can lead to conditioned responding to the subsequently presented cue (Rescorla & Heth, 1975). These phenomena indicate that the states are *not* identical with cues—rather, states are latent and must be inferred. The problem is made difficult by the fact that nothing tells the observer how many states exist or what their properties are, hence these must be inferred as well.

A principled approach to this problem can be derived by appealing to ideas from Bayesian nonparametrics, a field of statistics that deals with inference over latent structures with unbounded

[4]More precisely, ramps will occur when the value function is a convex function of the state representation (Gershman, 2014).

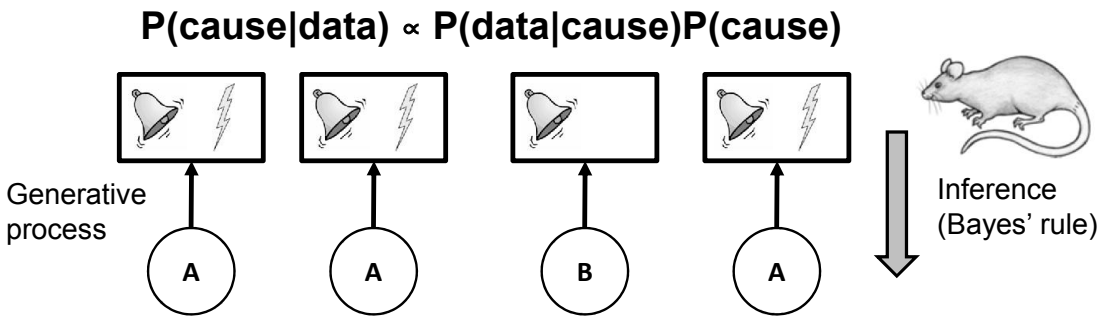

Figure 3: **The latent cause theory**. Each box represents the animal's observations on a single trial. The circles represent latent causes, labeled to distinguish different causes. The upward arrows denote probabilistic dependencies: observations are assumed to be *generated* by latent causes. The animal does not get to observe the latent causes; it must infer these by inverting the generative model using Bayes' rule, as indicated by the downward arrow. As shown at the top of the schematic, Bayes' rule defines the probability of latent causes conditional on observations, which is obtained (up to a normalization constant) by multiplying the probability of observations given hypothetical causes (the likelihood) and the probability of the hypothetical latent causes (the prior).

complexity (Gershman & Blei, 2012). Recent work has developed models of Pavlovian conditioning that use Bayesian nonparametric priors over latent causes, allowing the model to simultaneously infer the number and properties of the latent causes (Gershman et al., 2010; Gershman & Niv, 2012; Soto, Gershman, & Niv, 2014). Interested readers are referred to these papers for more details; here I will simply convey a few examples of how these models are applied (see also Redish, Jensen, Johnson, & Kurth-Nelson, 2007, for a related, non-probabilistic approach).

To a first approximation, a latent cause model is a good representation of the true causal structure underlying Pavlovian conditioning experiments. Cues do not cause outcomes—the experimenter causes both cues and outcomes. That is, the experimenter is a latent cause. This shows why it is useful to think about hidden states in terms of latent causes rather than simply as expedient mental constructs. As in other domains of cognition, rational analysis leads us to hypothesize that the mind has evolved the capacity to learn about and represent the underlying causal structure of the environment (Anderson, 1990).

Gershman et al. (2010) argued that memory recovery following extinction occurs because training and extinction trials are assigned to separate latent causes. This partition of trials into latent causes prevents unlearning of the fear memory during extinction, allowing it to return later. The theory predicts that performing training and extinction in different contexts will increase the probability of assigning them to separate latent causes. Bouton and Bolles (1979) confirmed this prediction, showing that returning the subject to the training context increases conditioned responding (an effect known as *renewal*).

One can also reverse the order of training and extinction, so that the extinction phase becomes a "preexposure" phase, causing a retardation of learning during training known as *latent inhibition* (Lubow, 1973). This phenomenon is interesting because there is no reward prediction error during

the preexposure phase (assuming that values are initialized to 0), and hence no learning signal according to the TD model. The latent cause model, on the other hand, naturally explains latent inhibition in terms of changes in the joint probability of cues and outcomes (Gershman et al., 2010). Latent inhibition is also context sensitive: performing preexposure and training in different contexts attenuates the latent inhibition effect (Hall & Honey, 1989). Differential context, according to the latent cause model, increases the posterior probability that the two phases were generated by separate latent causes (Gershman et al., 2010).

One might object that positing latent causes is superfluous when the different contexts are distinguished by observable stimuli. Context-dependency could therefore be captured by assuming that context acts as another cue, so that context effects are a form of compound conditioning. However, this assumption runs into the problem that contexts do not act like punctate cues such as tones and lights. Contexts do not summate with other cues: Pairing a previously conditioned context with a cue does not enhance responding compared to a condition in which the cue is presented alone (Bouton & Swartzentruber, 1986), and pairing an extinguished context with a cue does not suppress conditioning to the cue (Bouton & Bolles, 1979). In a similar vein, contexts do not excite conditioned responding on their own (Bouton & Swartzentruber, 1986). These findings support the proposal that contexts are modulatory in nature (Swartzentruber, 1995). At present, it is not clear that existing latent cause theories can adequately account for the modulatory role of context, but the findings at least cast doubt on a simple compound conditioning account.

The context-dependency of renewal and latent inhibition both rely on an intact hippocampus (Honey & Good, 1993; Ji & Maren, 2005), leading Gershman et al. (2010) to suggest that the ability to flexibly infer new latent causes depends crucially on the hippocampus. This suggestion fits with the work (reviewed above) characterizing the hippocampus as the seat of the "cognitive map," but in this case the inferred latent causes might feed into both model-based and model-free RL. Young rats also appear to lack context-dependent renewal and latent inhibition (Yap & Richardson, 2005, 2007), possibly due to immature hippocampal development.

Another factor that influences the assignment of trials to latent causes is reinforcement rate. A classic finding in Pavlovian conditioning is the *partial reinforcement extinction effect*: partially reinforcing the cue during training results in slower extinction (Capaldi, 1957; Wagner, Siegel, Thomas, & Ellison, 1964). This is surprising because standard RL models predict that partial reinforcement will produce a weaker value estimate that can be extinguished more easily. The latent cause model, in contrast, offers an intuitive explanation: slower extinction occurs because similar reinforcement rates during training and extinction provide evidence that the two phases were generated by the same latent cause (Courville et al., 2006; Gershman & Niv, 2012).

Gershman, Jones, Norman, Monfils, and Niv (2013) took this idea one step further and examined the effects of manipulating the reinforcement *sequence*. The logic of these studies was that large prediction errors during extinction induce the inference of a new latent cause. Thus, extinguishing gradually (by incrementally reducing the frequency with which a cue was paired with shock) should prevent the prediction errors from being large enough to induce the inference of a new latent cause, while being small enough to drive unlearning of the fear memory. The gradual extinction procedure was compared to a standard extinction procedure and a "gradual reverse" control, in which the cue and shock were paired with the same probability as in the gradual extinction condition but in reverse order (i.e., gradually increasing). All the conditions had a buffer of 8 unreinforced trials at the end

of extinction to ensure that conditioned responding fell to the same level across groups. Despite similar responding at the end of extinction, the groups differed strikingly in their recovery: while both the standard and gradual reverse groups showed spontaneous recovery and reinstatement, the gradual extinction group showed no evidence of recovery. This finding is consistent with the interpretation that gradual extinction led to a single latent cause assignment for both training and extinction.

These are a few examples of how latent cause models can address the problem of latent structure learning in partially observable domains. Undoubtedly, the models reviewed here are simplistic in a number of ways, and other versions attempt to address these shortcomings. For example, both Courville et al. (2006) and Soto et al. (2014) explored versions allowing multiple latent causes to be simultaneously active. Lloyd and Leslie (2013) have developed a version of the latent cause model that deals with a variety of complex instrumental learning phenomena. An important open question is how these approaches can be more tightly integrated into the RL formalism reviewed above, and ideally furnished with detailed neurobiological correlates.

Conclusions

In this chapter, I have argued that causal knowledge plays several roles in RL. First, model-based RL involves building a causal model of the environment and using this model to compute values. Second, both model-based and model-free RL rely upon inferences about latent causes in partially observable domains.

For many cognitive psychologists, RL has the inescapable odor of behaviorist ideology, and indeed traditional model-free RL enshrines this ideology by embracing Thorndike's law of effect. However, my hope is that this chapter conveys some of the ways in which theoretical ideas about RL have evolved beyond the law of effect. Moreover, some of the same formalisms invoked above appear throughout cognitive psychology. In particular, the probabilistic approach to causal learning and structure discovery has played a prominent role in the "rational analysis" of cognition (Anderson, 1990; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Modern theories of RL are now firmly ensconced in the cognitive fold.

Acknowledgments

I am grateful to the many collaborators who have influenced my thinking about these topics, in particular Nathaniel Daw, Yael Niv, Peter Dayan, Fabian Soto, and Ross Otto. This research was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL), under contract FA8650-14-C-7358. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *34*, 77–98.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

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

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

Bellman, R. (1957). Dynamic programming. *Princeton University Press*.

Bouton, M. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*, 485–494.

Bouton, M., & Bolles, R. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*, 445–466.

Bouton, M., & Swartzentruber, D. (1986). Analysis of the associative and occasion-setting properties of contexts participating in a Pavlovian discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *12*, 333–350.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., ... Colton, S. (2012). A survey of Monte Carlo tree search methods. *Computational Intelligence and AI in Games, IEEE Transactions on*, *4*, 1–43.

Capaldi, E. (1957). The effect of different amounts of alternating partial reinforcement on resistance to extinction. *The American Journal of Psychology*, *70*, 451–452.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300.

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

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130478.

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

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

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*, 613–624.

De Groot, A. D. (1978). *Thought and choice in chess*. The Hague: Mouton Publishers.

de Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: a case for animal–human translational models. *Psychological Research*, *73*, 463–476.

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

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312–325.

Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*, 109–120.

Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 229–240.

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Animal Learning & Behavior*, *18*, 264–270.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Gershman, S. J. (2014). Dopamine ramps are a consequence of reward prediction errors. *Neural Computation*, *26*, 467–471.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*, 197–209.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*, 273–278.

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., & Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, *7*.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*, 182–194.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*, 251–256.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*, 255–268.

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

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*, 299–308.

Hall, G., & Honey, R. C. (1989). Contextual effects in conditioning, latent inhibition, and habituation: Associative and retrieval functions of contextual cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*, 232–241.

Hasselmo, M. E. (2012). *How we remember: Brain mechanisms of episodic memory*. MIT press.

Holding, D. H., & Pfau, H. D. (1985). Thinking ahead in chess. *The American journal of psychology*, 271–282.

Honey, R. C., & Good, M. (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. *Behavioral Neuroscience*, *107*, 23–33.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*, e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*, 3098–3103.

Ji, J., & Maren, S. (2005). Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. *Learning and Memory*, *12*, 270–276.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kehoe, E. J. (1982). Conditioning with serial compound stimuli: Theoretical and empirical issues. *Experimental Behavior*, *1*, 30–65.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*, e1002055.

Kiernan, M., & Westbrook, R. (1993). Effects of exposure to a to-be-shocked environment upon the rat's freezing response: Evidence for facilitation, latent inhibition, and perceptual learning. *The Quarterly Journal of Experimental Psychology*, *46*, 271–288.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, *7*, e1000173.

Lloyd, K., & Leslie, D. S. (2013). Context-dependent decision-making: a simple bayesian model. *Journal of The Royal Society Interface*, *10*, 20130069.

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin*, *79*, 398–407.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & behavior*, *40*, 305–319.

Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–98.

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

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*, 265–272.

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

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press Oxford.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*, 751–761.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*, 20941–20946.

Pavlov, I. (1927). *Conditioned reflexes*. Oxford University Press.

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

Rao, R. P. (2010). Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, *4*.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*, 784–805.

Reid, A. K., & Staddon, J. (1998). A dynamic route finder for the cognitive map. *Psychological Review*, *105*, 585–601.

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory*, *11*, 501–509.

Rescorla, R. A., & Heth, C. D. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*, 88–96.

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

Century-Crofts.

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

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, *22*, 9475–9489.

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

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500.

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. In *Advances in neural information processing systems* (pp. 127–135).

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*, *121*, 526–558.

Stachenfeld, K. L., Botvinick, M., & Gershman, S. J. (2014). Design principles of the hippocampal cognitive map. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2528–2536). Curran Associates, Inc.

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

Sutton, R. S. (1990). Integrated architecture for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the seventh international conference (1990) on machine learning* (pp. 216–224).

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (p. 497-537).

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

Swartzentruber, D. (1995). Modulatory mechanisms in Pavlovian conditioning. *Animal Learning & Behavior*, *23*, 123–143.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*, 1279–1285.

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

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

Wagner, A., Siegel, S., Thomas, E., & Ellison, G. (1964). Reinforcement history and the extinction of conditioned salivary response. *Journal of Comparative and Physiological Psychology*, *58*, 354–358.

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

Yap, C. S., & Richardson, R. (2005). Latent inhibition in the developing rat: an examination of context-specific effects. *Developmental Psychobiology*, *47*, 55–65.

Yap, C. S., & Richardson, R. (2007). Extinction in the developing rat: an examination of renewal effects. *Developmental Psychobiology*, *49*, 565–575.