Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework

Samuel J. Gershman[1] and Nathaniel D. Daw[2]

[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138; email: gershman@fas.harvard.edu

[2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, New Jersey 08544

Annu. Rev. Psychol. 2017. 68:101–28

First published online as a Review in Advance on September 2, 2016

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi:
10.1146/annurev-psych-122414-033625

Keywords

reinforcement learning, memory, decision making

Abstract

We review the psychology and neuroscience of reinforcement learning (RL), which has experienced significant progress in the past two decades, enabled by the comprehensive experimental study of simple learning and decision-making tasks. However, one challenge in the study of RL is computational: The simplicity of these tasks ignores important aspects of reinforcement learning in the real world: (*a*) State spaces are high-dimensional, continuous, and partially observable; this implies that (*b*) data are relatively sparse and, indeed, precisely the same situation may never be encountered twice; furthermore, (*c*) rewards depend on the long-term consequences of actions in ways that violate the classical assumptions that make RL tractable. A seemingly distinct challenge is that, cognitively, theories of RL have largely involved procedural and semantic memory, the way in which knowledge about action values or world models extracted gradually from many experiences can drive choice. This focus on semantic memory leaves out many aspects of memory, such as episodic memory, related to the traces of individual events. We suggest that these two challenges are related. The computational challenge can be dealt with, in part, by endowing RL systems with episodic memory, allowing them to (*a*) efficiently approximate value functions over complex state spaces, (*b*) learn with very little data, and (*c*) bridge long-term dependencies between actions and rewards. We review the computational theory underlying this proposal and the empirical evidence to support it. Our proposal suggests that the ubiquitous and diverse roles of memory in RL may function as part of an integrated learning system.

Contents

INTRODUCTION

Reinforcement learning (RL) is the process by which organisms learn by trial and error to predict and acquire reward. This is challenging from a computational point of view because actions have long-term effects on future reward (e.g., failing to save may lead to penury later in life, drinking stagnant water may slake thirst at the expense of later illness). Furthermore, these deferred consequences may depend critically on other, subsequent actions and events: for example, getting admitted to college pays off only if one manages to graduate. This sequential dependency greatly compounds the classic curse of dimensionality (Bellman 1957) by extending it over time. Clearly, biological organisms cannot try every possible sequence of actions. By making certain simplifying assumptions about the structure of the environment, computer scientists have designed efficient algorithms that are guaranteed to find the optimal behavioral policy. The discovery that the brain itself uses one (indeed several) of these algorithms is one of the great success stories of modern cognitive and computational neuroscience.

Two decades of research have buttressed this account with converging evidence from behavioral, neural, and computational studies. Beginning in the 1990s, celebrated work on RL focused on a dopaminergic and striatal system for simple, incremental learning of action values, known as model-free learning (Houk et al. 1995, Montague et al. 1996, Schultz et al. 1997). Later work has extended this view to encompass additional processes for more deliberative, so-called model-based evaluation (Daw et al. 2005, Dolan & Dayan 2013). This expansion of our understanding of RL increases the computational capability of the theories—allowing them, for instance, to choose more effectively in novel or changed circumstances—and also situates them in relation to a broader framework of research in the cognitive neuroscience of memory. Model-based learning formalizes how organisms employ knowledge about the world—maps or models of task contingencies—in the service of evaluating actions. These results parallel research on multiple memory systems, e.g., distinguishing a striatal procedural learning system from a hippocampal declarative one (Eichenbaum & Cohen 2004, Poldrack et al. 2001), each of which possesses several properties that echo their decision-making counterparts. The emerging relationship between RL and the memory systems that likely subserve it has been illuminating for the study of both.

Despite this success, we are still far from understanding how real-world RL works, either cognitively or computationally. In this review, we suggest that these two questions have a common answer.

Cognitively, research on RL has long embraced procedural learning and semantic memory (in the sense of knowledge of facts about the world that are typically viewed as abstracted from many experiences, such as the map of a well-explored maze). However, this research has had limited contact with another prominent sort of memory: episodic memories connecting different aspects of individual events that the organism experienced at a particular time and place (Tulving 1972). Such traces seem, in principle, relevant to decisions; one goal of this review is to clarify what specific advantages they might confer.

Computationally, biological RL is still greatly hobbled by the restrictive formal assumptions that underpin it. With few exceptions, the kinds of experimental tasks that have been used to study RL are quintessentially toy problems: They are designed to isolate certain computations in a well-controlled setting but do not grapple with the complexity of many decision problems faced by organisms in their natural environments. In particular, (*a*) real state spaces are high-dimensional, continuous, and partially observable; this implies that (*b*) data are relatively sparse and, indeed, precisely the same situation will never be encountered twice; furthermore, (*c*) rewards depend on the long-term consequences of actions in ways that violate the probabilistic independence assumptions that make RL tractable.

Intuitively, these implications can be understood by considering the problem of investing in the stock market. The state of the stock market is high-dimensional and continuous, such that any given state is unlikely to be repeated (i.e., the market history sparsely samples the state space). Furthermore, the long-term consequences of an investment decision depend on forces that are only partially observable (e.g., the strategies of other investors). The kinds of algorithms that have been imputed to the brain break down when confronted with this sort of real-world complexity. Because organisms clearly find a way to cope with this complexity, we are left with the conundrum that much of our understanding about RL in the brain may in fact be irrelevant to important aspects of how organisms naturally behave.

In this review, we suggest that one computational answer to this conundrum is to use a set of algorithmic approaches—those based on nonparametric (kernel- or instance-based) estimation methods—that are different from and complementary to those typically examined in cognitive neuroscience. Nonparametric methods are statistically well-suited for dealing with sparse, arbitrarily structured, trial-unique data. Moreover, because they ultimately base their estimates on records of individual events, they may also clarify the missing links between decision making and episodic memory. These links are relatively underexplored, though they relate to a number of other ideas and make connections with other empirical literatures, which therefore form the balance of our review. The most important idea [building on the work of Lengyel & Dayan (2007)] is that episodic memory could provide detailed and temporally extended snapshots of the interdependency of the actions and outcomes originating from individual experiences; this information may be a reliable guide to decision making precisely in situations in which classical algorithms break down. Episodic memory may thus enable organisms to (*a*) efficiently approximate value functions over complex state spaces, (*b*) learn with very little data, and (*c*) bridge long-term dependencies between actions and rewards.

In the following sections, we review the current conception of RL in neuroscience and psychology and lay out the main arguments, both theoretical and empirical, that make this conception at best incomplete. We then describe a theoretical framework for augmenting RL with additional systems based on nonparametric estimation, which we tentatively identify with episodic memory. We consider the computational implications of this approach and review the available evidence related to this framework and its connection to earlier ideas.

REINFORCEMENT LEARNING: THE CURRENT PICTURE

We begin this section with a brief overview of the standard algorithmic solutions to the RL problem and then review behavioral and neural evidence that the brain implements these algorithms (for more extensive reviews of this material, see Dolan & Dayan 2013, Niv 2009; see also a trilogy of textbook chapters, Daw 2013, Daw & O'Doherty 2013, Daw & Tobler 2013). Our goal in this review is mainly to motivate a more prospective review of the possible connections with additional areas of research.

Markov Decision Processes

In machine learning, RL concerns the study of learned optimal control, primarily in multistep (sequential) decision problems (Bertsekas & Tsitsiklis 1996, Sutton & Barto 1998). Most classic work on this subject concerns a class of tasks known as Markov decision processes (MDPs). MDPs are formal models of multistep decision tasks, including spatial navigation, games such as Tetris, and scheduling problems such as those in factories; if some game-theoretic aspects of the opponent's behavior are neglected, then they can also roughly model multistep multiplayer games such as chess. The goal of RL is typically to learn, by trial and error, to make optimal choices in an initially unknown MDP.

Formally, MDPs are expressed in terms of discrete states s, actions a, and numeric rewards r. Much of the research in psychology and neuroscience surrounding these models concerns the tricky relationship between these formal objects and real-world situations, behaviors, and outcomes. Informally, states are like situations in a task (e.g., locations in a spatial maze), actions are like behavioral choices (turn left or right), and rewards are a measure of the utility obtained in some state (a high value for food obtained at some location, if one is hungry).

An MDP consists of a series of discrete time steps, in which the agent observes some state s_t of the environment, receives some reward r_t, and chooses some action a_t. The agent's goal is to choose actions at each step so as to maximize the expected cumulative future rewards, discounted (exponentially by decay factor $\gamma < 1$) for delay, i.e., the sum $r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$ of future rewards.

Thus, the goal is to maximize not the immediate reward of an action but instead the cumulative reward (the return), summed over all future time steps. Actions influence longer-run reward expectancy because, in an MDP, each successor state s_{t+1} is drawn from a probability distribution $P(s_{t+1} \mid s_t, a_t)$ that depends on the current state and action; rewards at each step are generated according to a probability distribution $P(r_t \mid s_t)$ that depends on the current state. Informally, this means that the agent navigates the states (like positions in a maze) and harvests rewards by choosing actions. Each action not only affects the current reward but, by affecting the next state, also sets the stage for subsequent rewards. Conversely, because the consequences of an action for cumulative reward depend also on subsequent states and actions, choosing optimally can be quite involved.

What makes these problems nevertheless tractably solvable is the eponymous feature of MDPs, the Markov conditional independence property: At any time step t, all future states and rewards depend only on the current state and action via the probability distributions given above. Thus, importantly, conditional only on the present state and action, all future events are independent of all preceding events. This permits a recursive expression for the state-action value function (the sum of future rewards expected for taking some action in some state, the quantity that is the goal of optimization):

$$Q_\pi(s_t, a_t) = r_t + \gamma \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_\pi(s_{t+1}, \pi(s_{t+1})). \tag{1}$$

Equation 1 is a form of the Bellman equation (Bellman 1957), versions of which underlie most classical RL algorithms. Here, it says that the expected future reward for taking action a_t in state s_t (then following some policy π thereafter) is given by the sum of two terms, the current reward and the second term, which stands in for all the remaining rewards $\gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$. The insight is that this sum is itself just the value Q of the subsequent state, averaged over possible successors according to their probability.

One of the chief problems of RL is how to choose advantageously given the deferred consequences of one's actions. One way to solve this problem is to focus on predicting those consequences via learning to estimate $Q_\pi(s_t, a_t)$ (or some closely related quantity) from experience with rewards, states, and actions in the MDP. Given a good estimate of the value function, you can choose the action with the best return simply by comparing values across candidate actions. Many RL algorithms rely on variations on this basic logic. (We omit some details related to the dependence of Q on the continuation policy π; for our purposes, imagine that by learning Q and choosing according to it, we gradually improve our prevailing action selection policy, which in turn drives an updated Q until we arrive at the best policy.)

Model-Based and Model-Free Algorithms

There are two main classes of algorithms for RL based on Equation 1; these classes focus on either the left- or right-hand side of the equal sign in that equation. The first approach is based on estimating the one-step reward and state transition distributions $P(r_t \mid s_t)$ and $P(s_{t+1} \mid s_t, a_t)$, which together are known as an internal model of the MDP. Notably, these concern only immediate events: which rewards or states directly follow other states and are thus easy to learn from local experience, essentially by counting. Given these, it is possible to iteratively expand the right-hand side of Equation 1 to compute the state-action value for any state and candidate action. Algorithms for doing this, such as value iteration, essentially work by mental simulation, enumerating the possible sequences of states that are expected to follow a starting state and action, summing the rewards expected along these sequences, and using the learned model to keep track of their probability (for a detailed presentation, see Daw & Dayan 2014).

This approach is known as model-based learning due to its reliance on the internal model. Its main advantage is the simplicity of learning, but its main disadvantage is that this simplicity is offset by computational complexity at choice time because producing state-action values depends on extensive computation over many branching possible paths.

The second class of algorithms eschews learning a world model and instead learns a table of long-run state-action values Q (the left-hand side of Equation 1) directly from experience. The discovery of algorithms for accomplishing such model-free RL [in particular, the family of temporal-difference (TD) learning algorithms; Sutton 1988] was a major advance in machine learning that continues to provide the foundation for modern applications (e.g., Mnih et al. 2015).

Briefly, these algorithms use experienced states, actions, and rewards to approximate the right-hand side of Equation 1 and average these to update a table of long-run reward predictions.

More particularly, many algorithms are based on the temporal difference reward prediction error occasioned by comparing the value $Q(s_t, a_t)$ to a sample computed one time step later:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t). \tag{2}$$

When the value function is well estimated, this difference should on average be zero [because $Q(s_t, a_t)$ should in expectation equal $r_t + \gamma Q(s_{t+1}, a_{t+1})$, according to Equation 1]. When the error is nonzero, stored Qs can be updated to reduce it.

Choice is, accordingly, much simpler using model-free algorithms than using model-based algorithms because the long-run values are already computed and need only be compared to find the best action. However, these computational savings come at the cost of inflexibility and less-efficient learning.

Model-Free Learning in the Brain

The initial and still the most-celebrated success of RL theory in neuroscience was the observation that the firing of dopamine neurons in the midbrain of monkeys behaving for reward resembles the reward prediction error of Equation 2 (Houk et al. 1995, Montague et al. 1996, Schultz et al. 1997), suggesting that the brain may use this signal for RL. The trial–trial fluctuations in this signal track the model quite precisely (Bayer & Glimcher 2005) and can also be measured in rodents using both physiology and voltammetry (Cohen et al. 2012, Hart et al. 2014). A similar signal can also be measured in the ventral striatum (an important dopamine target) in humans using fMRI (e.g., Hare et al. 2008). Although fMRI measurements are not specific to the underlying neural causes, dopaminergic involvement in these prediction error correlates is suggested by the finding that these correlates are modulated by dopaminergic medication (Pessiglione et al. 2006) and by Parkinson's disease (Schonberg et al. 2010), which is marked by the relatively selective degeneration of dopaminergic nuclei.

Many researchers believe that dopamine drives learning about actions by modulating plasticity at its targets, notably medium spiny neurons in striatum (Frank et al. 2004). Via their projections to other basal ganglia nuclei (and ultimately to the motor cortex), these neurons drive elicitation and withholding of behavior (Alexander & Crutcher 1990). Accordingly, Parkinson's disease and dopamine replacement therapy in humans modulate learning in RL tasks (Frank et al. 2004, Shohamy et al. 2005). More temporally specific optogenetic elicitation and suppression of dopaminergic responses in rodents also drives learning in tasks specifically designed to isolate error-driven learning (Parker et al. 2016, Steinberg et al. 2013). These studies refine an earlier literature using less-selective electrical or pharmacological stimulation of the systems; notably, drugs of abuse invariably agonize dopamine as a common link of effect. This suggests that the reinforcing effects of these drugs are ultimately driven by the same RL mechanisms discussed in this review (Everitt & Robbins 2005, Redish 2004).

The behavioral experiments discussed above mainly involve nonsequential decision tasks such as one-step bandit tasks, in which a subject repeatedly chooses between a set of actions (e.g., different slot machines) and receives reward or punishment. Indeed, the trial-by-trial dependency of choices on rewards in such tasks is quantitatively consistent with the pattern predicted by error-driven learning in both monkeys (Lau & Glimcher 2005) and humans (Seymour et al. 2012). However, model-free learning according to Equation 2 makes more specific and characteristic predictions about the progression of learning across states in multistep, sequential tasks. The predicted patterns have been confirmed in humans (Daw et al. 2011, Fu & Anderson 2008), although not exclusively. Indeed, long before the advent of the neurophysiological models, behavioral psychologists had established that basic TD learning cannot by itself explain a number of learning effects, a point we examine in the next section.

Model-Based Learning in the Brain

Although model-free and model-based algorithms both ultimately converge to the optimal value predictions (under various technical assumptions and in the theoretical limit of infinite experience in a fixed MDP; e.g., Bertsekas & Tsitsiklis 1996), they differ in the trial-by-trial dynamics by which they approach the solution. One difference between model-free and model-based algorithms is the fact that, because the model-free algorithms learn long-run action values by sampling them directly along experienced trajectories, they can in some cases fail to integrate information encountered in different trajectories (e.g., separate trials or task stages).

This basic insight has been investigated using tasks involving staged sequences of experience that are ordered in such a way as to defeat a model-free learner. For instance, in latent learning (Gläscher et al. 2010, Tolman 1948) and a similar task called sensory preconditioning (Brogden 1939, Wimmer & Shohamy 2012), organisms are first preexposed to the state-action contingencies in an environment without any rewards (e.g., by exploring a maze), then subsequently learn that reward is available at a particular location.

For a model-based learner, this experience has the effect of teaching them first the transition function $P(s_{t+1} \mid s_t, a_t)$, i.e., the map of the maze, and then, separately, the reward function $P(r_t \mid s_t)$. Together, this information enables them in a subsequent probe phase to navigate to the reward from any location by evaluating Equation 1. However, for a model-free learner, the preexposure stage teaches them nothing useful for the probe (only that Q is everywhere zero); in particular, because they don't separately learn a representation of the map of the maze (the state transition distribution), they must learn the navigation task from scratch when reward is introduced.

Humans and even rodents can, at least under some circumstances, successfully integrate these experiences, demonstrated in this case by facilitated navigation learning in groups who received the preexposure (Gläscher et al. 2010, Tolman 1948). These results and logically similar ones involving studying whether animals require additional experience to adjust their decisions following changes in reward value (e.g., outcome devaluation) or task contingencies (e.g., introduction of blockades or shortcuts, contingency degradation) have been taken as a rejection of model-free RL as a complete account of behavior (Daw et al. 2005, Dickinson & Balleine 2002).

However, the same types of experiments actually do support the predictions of model-free learning mechanisms such as TD because, under other circumstances, organisms fail, as the theories predict, to integrate well- (but separately) learned information about contingencies and rewards. For instance, following overtraining on lever pressing for food, rodents will press the lever even after the outcome is devalued by satiety (Adams 1982), although less thoroughly trained animals can successfully adjust. In psychology, these two sorts of behaviors (incapable and capable of integration, respectively) are known as habitual and goal-directed. Lesion studies in rodents suggest that they are dependent on discrete networks in the brain, involving different parts of the frontal cortex and striatum (for a review, see Daw & O'Doherty 2013).

Altogether, the predictions of model-free learning and the prediction error theories of dopamine are well matched to habitual behavior but fail to account for the additional category of goal-directed behavior and the ability of organisms to integrate experiences. This deficiency led to the suggestion that the latter behavior might be understood in terms of model-based learning operating alongside the model-free system and competing to control behavioral output (Daw et al. 2005). This proposal put hitherto looser ideas about deliberative behavior and cognitive maps on more equal quantitative footing with the more specific neurocomputational theories of habitual learning, enabling further investigation of its properties.

For instance, with more specific characterizations of both sorts of learning, it is possible to dissociate trial-by-trial behavioral adjustments and neural correlates of decision variables like Q

associated with either model-based or model-free learning in multistep decision tasks (e.g., two-step, three-state MDPs; Daw et al. 2011). Experiments using this technique have verified that signatures of both types of learning coexist in humans. Their prevalence can be manipulated situationally (Otto et al. 2013a,b), varies across individuals (e.g., with symptoms of compulsive disorders such as drug abuse) (Gillan et al. 2016), and tracks prospective representation of future states measured in fMRI at the time of choice (consistent with choice-time evaluation via mental simulation) (Doll et al. 2015). Research using elaborate multistep decision tasks has also begun to shed light on the computational shortcuts by which the brain manages to compute the expected reward (Cushman & Morris 2015, Dezfouli & Balleine 2013, Diuk et al. 2013, Huys et al. 2015, Solway & Botvinick 2015).

Less is known about the neural circuits supporting putatively model-based behavior. Particularly in human neuroimaging, there appears to be more overlap between neural signals associated with model-based and model-free learning than might have been expected on the basis of lesion work. For instance, prediction error signals in human striatum (Daw et al. 2011) and rodent dopamine neurons (Sadacca et al. 2016) both reflect integrated, model-based valuations. (This is surprising because those signals provide the foundation for the standard model-free account.) Such results might suggest that the systems interact more cooperatively in the intact than the lesioned brain, that model-based computations are built in part by leveraging phylogenetically earlier model-based circuitry, that there is more of a continuum between them, or that the integration of value that is taken as a signature of model-based computation is actually heterogeneous and may occur via a number of different mechanisms at different times (Gershman et al. 2014, Shohamy & Daw 2015, Wimmer & Shohamy 2012).

Other data point to the hippocampus as an important player in model-based RL. The model-free versus model-based distinction appears to track a similar dichotomy in the study of multiple memory systems, which in broad terms distinguishes a rigid striatal procedural learning system from a more flexible declarative memory system associated with the hippocampus (Gabrieli 1998, Knowlton et al. 1996, Squire 1992). Many of the particular aspects of hippocampal function also suggest it as a candidate site for world models as envisioned in RL. For example, the hippocampus has been viewed as a seat of the cognitive maps useful for spatial navigation (O'Keefe & Nadel 1978). Perhaps the most directly suggestive data concerning a potential neural circuit for model-based evaluation also come from spatial navigation tasks, in which representations of place cells in the rodent hippocampus appear to run ahead of the animal during navigation and at choice points (Johnson & Redish 2007, Pfeiffer & Foster 2013). This prospective activity has been suggested to instantiate a search of future trajectories to support model-based evaluation (e.g., decision-time computation in Equation 1). However, this phenomenon has yet to be specifically linked to choice behaviors (such as latent learning or other integrative tasks) that demonstrate model-based evaluation.

In addition to spatial navigation, the hippocampus is also associated with more abstract relational information reminiscent of the state transition function (Eichenbaum & Cohen 2004, Shohamy & Wagner 2008). However, perhaps the most well-known function of the hippocampus is the formation of episodic memories, which are long-term, autobiographical snapshots of particular events. This function has also been linked to prospective construction of imagined future episodes for planning or other decisions (Schacter et al. 2012). It has not, however, received as much attention in RL. Below, we argue that it may underlie some decisions that appear to be model-based. The relationship between these seemingly disparate aspects of hippocampal memory function is a deep conceptual issue that has given rise to ongoing debate in the cognitive neuroscience of memory.

Computational Shortcomings of the Current Picture

The computational and neural mechanisms described in the two sections above appear to be reasonably well-supported, albeit with some uncertainty related to the neural implementation of world modeling and integrative evaluation. However, the ways in which these mechanisms could scale up to real-world tasks remain unclear. Not only are the tasks that have been studied in the laboratory small and artificial but, more importantly, the very assumptions that allow RL to work well in these sorts of tasks are inapplicable to many richer real-world settings.

Many of the problems with the current conception of RL arise from the definition of the state s_t. Laboratory experiments typically involve at most a handful of discrete states and actions, which are clearly signaled to the subjects and designed to satisfy the Markov conditional independence property. Real-world sensations rarely meet these conditions. The typical sensory experiences of an organism are both too vast and too impoverished to serve as s_t in algorithms based on Equation 1. They are too vast because they are continuous and high-dimensional, such that effective learning requires identifying the subset of relevant dimensions and generalizing appropriately across situations that will never exactly recur (Niv et al. 2015).

Real-world sensations are also too impoverished because, despite the extraneous detail in one's immediate sensory observations, they rarely satisfy the Markov property; other information observed in the past but not currently observable affects future state and reward expectancies. Immediate sensations routinely fail to satisfy the Markov property in real-world tasks; for example, the property is violated whenever two different locations look similar enough to be indistinguishable (state aliasing) during navigation or when there are long-run dependencies between day-to-day events, such as when someone tells you they'll be back tomorrow at noon for lunch. If the Markov property fails to hold for some putative state s, it is not possible to decompose the state-action value via the Bellman equation (Equation 2).

Of course, extensive machine learning work exists on ways to cope with some of these circumstances. Particularly relevant for neuroscience is the theory of partially observable Markov decision processes (POMDPs) (Kaelbling et al. 1998), which treats Markov violations as arising from latent states that would satisfy the Markov property but can be only indirectly (and perhaps ambiguously) observed. With training, one can learn to infer the identity of these states (which may indeed provide part of a theoretical basis for state representation for RL; Daw et al. 2006; Gershman et al. 2010, 2015; Rao 2010), but only after having done so is one in a firm position to learn action values. In the following sections, we consider mechanisms that might be applicable to earlier stages in the learning process and might also be flexible and able to adapt in the face of ongoing learning about how to define the state, which dimensions are relevant, and how to infer latent aspects.

EPISODIC MEMORY FOR NONPARAMETRIC VALUE FUNCTION APPROXIMATION

If the current computational conception of RL is incomplete, how can progress be made? One approach is to further examine what the brain's memory systems might suggest about RL.

Existing RL theories have recognized the links between RL and what is known in memory research as procedural memory (for model-free policies or action values), as well as semantic declarative memories (for world maps or models). Strikingly, these quantities, such as procedural knowledge of how to ride a bike or semantic knowledge of what a typical breakfast might contain, all represent statistical summaries extracted from a series of events. In contrast, a predominant focus of research in memory concerns memory for one-shot events, from word lists to autobiographical events such as your 30th birthday party or what you had for breakfast this morning. The remainder of this review considers how memories for individual events might serve RL and, in particular, why these memories might help theories of RL to escape some of their previous weaknesses and the restrictive assumptions under which they operate.

Though an interesting computational object in the abstract, one-shot memories are not unique to long-term episodic memory. For instance, working memory clearly plays a role in maintaining and manipulating information briefly, as is the case for phone numbers. This type of memory is also discussed in our review. However, we mainly consider long-term episodic memory, which, apart from having a number of appealing computational features for RL, is also associated with the hippocampus, which has other mnemonic roles already implicated in model-based RL. (Although we are not yet in a position to entirely reconcile these functions, it is nevertheless clear that episodic aspects are conspicuously lacking from the current conception of RL.)

Psychologically, episodic memory is associated with detailed autobiographical memories, such as what you had for breakfast this morning, that link many different sensory features of an experience at a particular time and place (Tulving 1972). Computationally, for the purpose of this review, we would stress the notion of a record of an individual event (like a trial in a task) and the connection between many aspects of that event, including multiple sensory dimensions and sensations experienced sequentially. Below, we reason about what sort of advantages episodic memories might confer on an organism's decision making and argue that these memories are well suited to the situations poorly handled by the mechanisms considered above and well linked to another class of estimation algorithm.

For this explanation, we build on an earlier proposal by Lengyel & Dayan (2007), who suggested that episodic memories could be used to record and later mimic previously rewarding sequences of states and actions, a process they dubbed episodic control. In this review, we suggest a somewhat different computational rationale for a similar idea, which we call episodic RL, in which episodic memories are used to construct estimates of the state or state-action value function (rather than for extracting policies, i.e., action sequences, directly). These evaluations can then be compared to derive choice policies in the usual way.

The previous section identified two difficulties with existing algorithms in real-world circumstances. First, the space of situations (states) is vast, and which features or dimensions of it are relevant to value prediction are not typically known in advance. Second, many RL systems harness the recursive structure of the Bellman equation, but the Markov assumptions that underpin this recursive structure are invalid in many real-world environments (e.g., when there are long-term dependencies). Memory for individual episodes can help ameliorate these problems by allowing the later construction of a nonparametric approximation of the value function that need not precommit at the time of encoding to averaging with respect to particular relevant sensory dimensions or to reliance on the Bellman equation for a particular choice of state.

To understand what this means, recall that the value of a state represents the cumulative future reward over a (possibly infinitely long) trajectory. Model-free algorithms store and update a running average of this value, whereas model-based algorithms compute the value on the fly using estimates of the reward and transition functions. These approaches are parametric in the sense that they estimate a set of parameters that specify the value function (cached values in the case of model-free control, model parameters in the case of model-based control). Once these parameters have been estimated, the raw data can be discarded.

Episodic RL keeps the raw data in memory and approximates state values by retrieving samples from memory. Intuitively, this works because the value of a state can be approximated simply by summing rewards collected along a remembered trajectory initiated in that state or averaging such sums across several such trajectories. Because these trajectories are individual and temporally extended, they capture arbitrary long-range, non-Markovian dependencies among events. Moreover, as discussed below, this procedure allows for flexible and adaptive generalization in terms of what counts as a similar state for the purpose of forecasting value in novel circumstances.

Episodic RL is nonparametric in the sense that it does not rely on a fixed, parameterized form of the value function. The effective complexity of the approximation (i.e., the number of episodes) grows as more data are observed. This approach is similar to that of a well-developed literature in statistics and machine learning on nonparametric estimation (for a textbook treatment, see Wasserman 2006) and a more specialized set of applications of these techniques to value estimation in the RL setting (e.g., Engel et al. 2005, Ormoneit & Sen 2002).

Formalization of Episodic Reinforcement Learning

The simplest implementation of episodic RL (**Figure 1**) is to store individual trajectories in memory and, when a familiar state is encountered, retrieve the set of trajectories that have followed each candidate action in that state, averaging the rewards subsequently obtained to estimate the value of each action. Formally,

$$Q_\pi(s_1, a) = E_\pi\left[\sum_{n=1}^{N} \gamma^{n-1} r_n \mid s_1, a\right] \approx \frac{1}{M}\sum_{m=1}^{M} R_m,$$

where M is the number of retrieved trajectories, R_m is the cumulative discounted return for each trajectory, and π is the prevailing policy. This approach works reasonably well when the state space is small and sequences are not deep. However, there are several problems with this implementation when applied to more general environments (e.g., with large state spaces and long planning horizons). First, because it seems likely that only relatively short trajectories can be stored in memory (much work in memory concerns the segmentation of events between episodes; e.g., Ezzyat & Davachi 2011), episodic RL may tend to be myopic, neglecting long-term future events due to truncation of the trace. Computationally, estimates of long-run reward based on sample trajectories also have large variance as the horizon grows longer because increasing numbers of random events intervene along the way (Kearns & Singh 2000). Second, in complex or continuous state spaces, states may be rarely, if ever, revisited; thus, the controller needs a mechanism for generalization to new states.

The first problem can be addressed by combining episodic RL with the Bellman equation. Consider an agent who retrieves a set of trajectories M starting with action a in state s_1 and ending N time steps later in some state s_{mN}, which may differ for each episode m. The value of this state can be expressed as follows:

$$Q_\pi(s_1, a) = \frac{1}{M}\sum_{m=1}^{M}\left[R_m + \gamma^N \sum_{s} P(s_{N+1} = s \mid s_{mN}, \pi(s_{mN}))\, Q_\pi(s_{N+1}, \pi(s_{N+1}))\right].$$

The first term in this equation represents the expected return from an episode of length N and the second term represents the expected return after that trace has terminated. The second term could be computed using model-based or model-free value estimates or by chaining together a sequence of episodes. Combining these terms allows episodic RL to correctly take into account the long-term consequences of a finite trajectory. Notably, the individual sequences capture arbitrary long-run dependencies among events (up to their length), and a Markovian assumption is invoked only to knit them together. It is also possible to knit together shorter sequences or, in the limit, individual state transitions themselves, each drawn from a set of sample episodes (Ormoneit & Sen 2002), to the extent the Markovian assumption can be relied upon. Unlike traditional

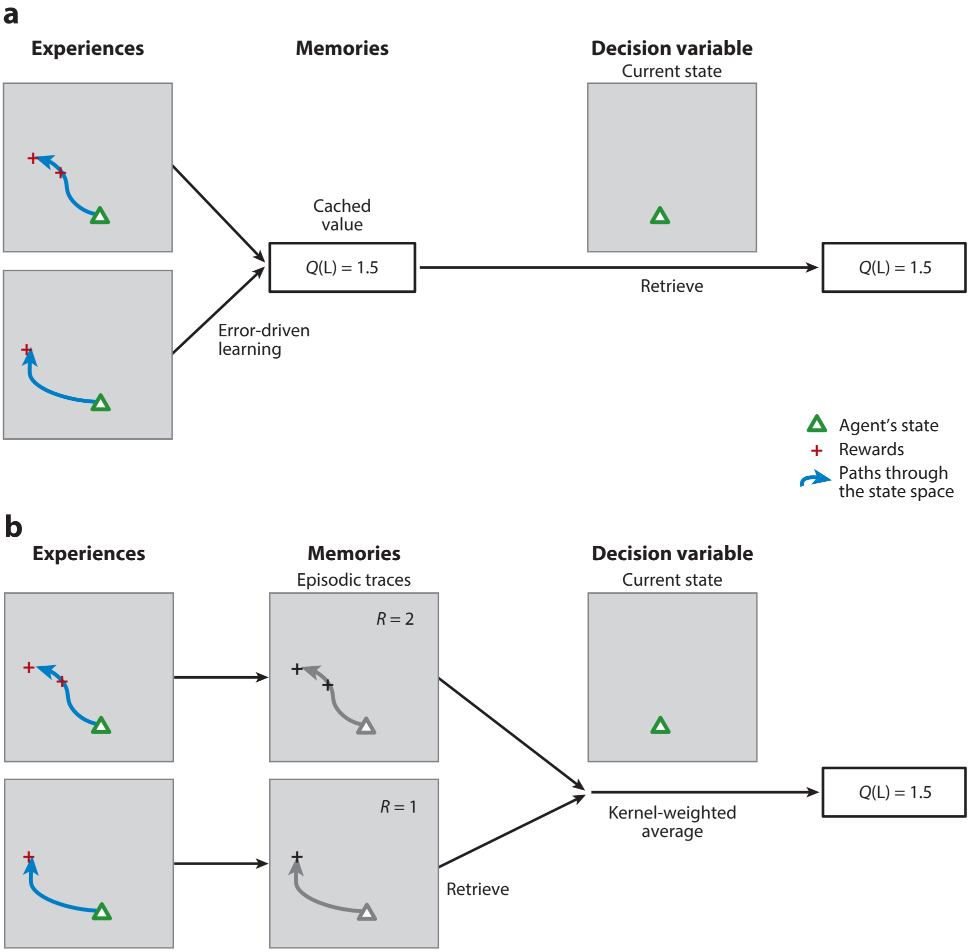

Figure 1

Schematic representation of different approaches to value computation. (*a*) In model-free reinforcement learning, individual experiences are integrated into a cached value, which is then used to compute action values in a new state. Only cached values are stored in memory; individual experiences are discarded. (*b*) In episodic reinforcement learning, individual experiences, along with their associated returns, are retained in memory and retrieved at choice time. Each episodic trace is weighted by its similarity to the current state according to a kernel function. This kernel-weighted average implements a nonparametric value estimate. *Q*(L) is the value of taking the left action.

model-free approaches (Sutton 1988), the decision of how heavily to rely on the Markovian assumption need not be made when experience is first acquired but instead can be made later, at choice time, when it is used to compute decision variables. Thus, this decision can be informed by additional experience in the interim.

The process of chaining episodes bears a striking resemblance to the use of options in hierarchical RL (Botvinick et al. 2009). Options are policies that have specific initiation and termination conditions; when one option terminates, another option is invoked. Just as options allow an agent to build reusable subroutines out of primitive actions, episodes allow an agent to reuse past experience. In fact, episodic retrieval may be one way in which options are created.

The second problem—generalization—can be addressed by allowing values to be smooth interpolations of episodes. Specifically, the expected return of a trajectory can be estimated by

$$E_\pi\left[\sum_{n=1}^{N}\gamma^{n-1}r_n \mid s_1, a\right] \approx \frac{\sum_{m=1}^{M} R_m K(s_1, s_{m1})}{\sum_{m=1}^{M} K(s_1, s_{m1})},$$

where M is the number of retrieved memory traces, s_{m1} is the initial state of the trajectory stored in memory trace m, and R_m is the return for the trajectory. The kernel function $K(s_1, s_{m1})$ measures the similarity between the current and retrieved states. The kernel function can also be defined over state-action or state-action-reward tuples. Such generalization is important for the purposes of choice because it allows an agent to estimate the value of taking a particular action in novel circumstances or in continuous state spaces. Again, an important feature of this model is that the kernel function K need not be fixed at the time of initial learning but can be shaped by subsequent experience before the episodes are used to guide choice. This contrasts with traditional generalization based on parametric function approximation schemes, such as neural networks, which amount to averaging values over some area of the state space at encoding time (e.g., Sutton & Barto 1998).

The appropriate kernel depends on the structure of the state space. For example, in a smooth, real-valued state space, a commonly used kernel is the Gaussian:

$$K(s, s') = \exp\left(-\frac{\|s - s'\|^2}{2\sigma^2}\right),$$

where the bandwidth parameter σ^2 governs the smoothness of the value function approximation; a smaller bandwidth induces sharper generalization gradients and, in the limit, produces no generalization (i.e., a pure episodic memory). The optimal bandwidth decreases with the number and increases with dispersion of episodes (Wasserman 2006). Intuitively, the bandwidth provides a form of regularization, preventing the kernel estimate from overgeneralizing. Kernels can also be defined over discrete state spaces, as well as structured objects like graphs, grammars, and trees (Gärtner et al. 2004), and an analogous parameterization of bandwidth can sometimes be specified.

Kernel-based approaches to RL fit well with similar approaches applied to other areas of cognition (Jäkel et al. 2009). Exemplar models of memory, categorization, object recognition, and function learning can be interpreted as forms of kernel density estimation. Of particular relevance is Gilboa & Schmeidler's (2001) case-based decision theory, which (as we discuss in the section Case-Based Decision Theory and Decision by Sampling) applies kernel density estimation to decision problems. Research on machine learning has demonstrated the efficacy of kernel-based approaches (Ormoneit & Sen 2002), although relatively little work has compared the computational and statistical trade-offs of these approaches with those of conventional model-based and model-free RL.

Reinforcement Learning and Memory for Individual Episodes

The framework outlined in the previous section and the predecessor proposal by Lengyel & Dayan (2007) suggest that RL behavior should, under some circumstances, be driven by memory for individual episodes distinct from the aggregate statistics of these episodes that would be employed by a model-based or model-free learner. The empirical literature directly supporting these predictions is, at present, fairly sparse, mostly because the sorts of behavioral tasks most commonly used in studies of RL do not easily lend themselves to addressing these issues. Two limitations of these tasks contribute to these deficiencies.

First, unlike studies of categorization—in which subjects render judgments about many unique stimuli and exemplar-based models reminiscent of our framework have long been successful (Nosofsky 1986)—most laboratory studies of RL consist of many repetitions of essentially identical trials. Thus, there has been little, experimentally or psychologically, to differentiate episodes and few objectively predictable features other than temporal recency to govern which episodes subjects might retrieve. Second, although some of the most interesting features of nonparametric episodic evaluation (like RL evaluation in general) occur during the evaluation of sequential decision tasks, existing work relevant to these ideas has mostly taken place in repeated choice-reward bandit tasks without sequential structure. However, some supporting evidence does exist.

Recently, Collins & Frank (2012) proposed a model, as well as an associated experimental task to test it, that argued that many trial-by-trial choices in RL tasks in humans were driven by a small set of memories of previous events held in working memory rather than incremental running averages of the sort associated with model-free (and model-based) RL. This idea bears some resemblance to the current episodic RL proposal (although focusing on a different memory system as the store). In support of this theory, the researchers found that increasing the number of stimuli (the set size) or time delays between state visits in a bandit-like task slowed learning, a finding inconsistent with standard RL models but well explained by a model that uses a limited memory buffer over stimulus history to determine action values. Individuals with a genetic polymorphism associated with higher levels of prefrontal dopamine exhibited greater retention of previous stimulus history in the action values. Further work using this task has shown that schizophrenic patients have a selective impairment in the working memory component of RL (Collins et al. 2014) consistent with the observation of reduced prefrontal dopamine levels in schizophrenia.

This mechanism does not fully coincide with episodic evaluation as we have described it. First, the task is deterministic and the state space discrete, so aspects of generalization and averaging over noisy outcomes are not exercised. Second, we [and other theorists, such as Zilli & Hasselmo (2008)] have assumed that, for an episode-based RL system to be useful over longer delays (including retaining learning from, for example, one day to the next) and larger state spaces, it likely must involve the episodic memory system of the hippocampus rather than short-term working memory.

Other work on bandit tasks, in this case with stochastic outcomes, has been carried out by Erev and colleagues (e.g., Erev et al. 2008). These investigators have argued that many aggregate features of subjects' choice preferences are best explained by a model that maintains individual trial outcomes rather than running averages. According to the model, which can be thought of as an instance of episodic RL, subjects evaluate bandits on the basis of a small sample (e.g., one or two) of particular rewards previously received from them, although not always (as would be predicted by running averages) the most recent ones. The statistics of decision variables implied by such sampling explain a variety of features of preferences in these tasks, such as sensitivity to risk and loss.

One issue standing in the way of examining this sort of model is the basic similarity of all trials in a bandit task to one another. Other research has integrated incidental trial-unique images with bandit tasks to begin to gain leverage over individual episodes. For example, Bornstein et al. (2015) found that using these images to remind subjects of previous trials influenced their subsequent action immediately after the reminder: If a past action resulted in a reward, then a reminder of that trial induced subjects to repeat it, whereas if the action resulted in a loss, then a reminder induced subjects to avoid it. This manipulation might be understood as influencing memory retrieval in episodic RL.

Wimmer et al. (2014) investigated a similar manipulation using fMRI. In this study, episodic memory for the trial-unique objects (tested after the experiment) covaried negatively with the influence of reward history on decisions at encoding time such that better (subsequently measured) episodic memory was associated with weaker feedback-driven learning. This negative effect of successful episodic encoding was also associated with an attenuated striatal prediction error signal and increased connectivity between the hippocampus and the striatum. One possible interpretation of this result in terms of episodic RL is that, because the trial-unique objects were entirely incidental to the task, episodic evaluation mechanisms (to the extent that they were engaged) effectively injected uncontrolled noise into the evaluation process, obscuring both reward-driven choice behavior and associated striatal signals.

Overcoming State Aliasing

One advantage of episodic RL is its robustness: State values can be validly estimated by remembered trajectories even when the Markov properties do not hold within the trajectory. That is, a set of returns following some current state s_1 validly estimate its long-term value even if there are arbitrary long-range dependencies across the events within the sample trajectories. However, this property only partly solves the problems of state representation. In particular, if the starting state s_1 does not itself satisfy the Markov property (that is, if outcomes following s_1 depend on events that happened prior to s_1 but aren't reflected in it), then the set of returns matching s_1 will not reflect this additional information. This will introduce additional noise in even episodic value estimates.

Violations of this assumption can occur when states are aliased: If multiple states are indistinguishable on the basis of the current observation, then the value is not conditionally independent of the agent's history given the current observation. Work on this problem again looks to memory (in this case, short-term working memory) to disambiguate the state by augmenting it with appropriate recent stimulus history. For example, if you received instructions to turn left after the second traffic light, the value of a left turn is not specified simply by whether you are at a traffic light but by the trajectory preceding it. This dependence is eliminated, though, if you can remember how many traffic lights you passed. In other words, the number of traffic lights is a sufficient statistic for your history, and storing it in memory allows you to incorporate it into the state representation and validly apply standard RL algorithms. The main problem here is how much and what sort of history to store.

This insight is the basis of several computational models of the ways in which working memory aids RL. Dopamine functions as a gating signal in the prefrontal cortex, allowing phasic bursts of dopamine to transiently increase the gain of prefrontal neurons, making them more responsive to afferent input (Cohen et al. 2002). Importantly, Braver & Cohen (2000) demonstrated that TD learning could be used to adaptively gate relevant information into working memory, excluding irrelevant distractors. In essence, this work treated the evaluation (via RL) and selection of cognitive actions (inserting and removing items from working memory) in the same way as the selection of motor actions, providing an integrative explanation of dopamine's role in both cognitive and motor control. O'Reilly & Frank (2006) extended this idea by showing how adaptive gating could be realized in a biologically detailed model of prefrontal–basal ganglia interactions. Further insight was provided by Todd et al. (2008), who articulated how adaptive gating could be understood as a normative computational solution to partial observability.

The challenge that all of these models seek to address is discovering which particular past events need to be retained in working memory and for how long. It is noteworthy that Todd et al.'s (2008) model discovers these long-run relationships by leveraging a form of TD value estimation known as TD(λ), in which all previously visited states are eligible for updating on every time step; this variant of the algorithm is statistically related to the evaluation of state values by episodic sample trajectories (Sutton & Barto 1998). This suggests that episodic memory might also be useful for the same purpose. From the perspective of an episodic RL model, learning of this sort, in effect, allows the organism to figure out under what circumstances to apply the Markov property. This understanding can then be applied, going forward, to computing values using the experience stored in episodic traces. In keeping with a recurring theme of this review, one advantage of this understanding, relative to state learning models such as Todd et al.'s (2008), is that action values need not be relearned from new experience, only recomputed as the understanding of the state space evolves.

Approximating Value Functions Over Complex State Spaces

As discussed above, raw memory traces are of limited use when making decisions in novel situations because they generalize poorly. To use a previous example, exactly counting the number of traffic lights will fail if one is forced to take a detour; in this case, it is necessary to use a value function approximation that degrades gracefully with deviations from the stored memory traces. This limitation has motivated the use of kernel methods that allow some degree of generalization.

In studies of RL, this problem is typically addressed as a question of value function approximation: How does an agent approximate the function $Q(s,a)$ over (potentially continuous and high-dimensional) states? Much work in computational neuroscience has been devoted to trying to understand how these issues play out in the brain. Proposed architectures typically implement linear or nonlinear parametric approximations, e.g., taking $Q(s,a)$ to be approximated by a weighted sum of a set of basis functions defined over the state space. However, it is unclear whether such parametric approximations can scale up to real-world problems, in which the appropriate feature space is elusive. One approach pursued in machine learning has been to develop complex architectures, such as deep neural networks, which can learn to discover good parametric representations from a large amount of training data (Mnih et al. 2015). However, this approach does not seem to provide a complete account of human performance, which can in certain cases be effective after observing a very small amount of data (e.g., Griffiths et al. 2010, Lee et al. 2014). This ability is partly attributable to strong inductive biases that guide learning (Griffiths et al. 2010). Another factor may be the brain's use of kernel methods that generalize from sparse training examples to new testing situations in a way that captures the underlying structure of the state space.

Intuitively, a good kernel assigns high similarity to states that have similar values, allowing the value function approximation to average across the rewards in these states while abstaining from averaging over states with different values. In the literature on biological reinforcement learning, these issues of generalization have mainly been discussed in terms of selecting an appropriate set of basis functions for parametric (linear) value function approximation (e.g., Foster et al. 2000, Ludvig et al. 2008), but exactly the same considerations apply to the choice of kernel for nonparametric generalization. A particular advantage of the latter is that the kernel is used at choice time rather than at encoding time, so it can be learned or adapted by subsequent experience, as in many of the schemes discussed below.

In spatial domains, appropriate generalization can be given a concrete, geometric interpretation. For example, a Gaussian kernel defined over Euclidean spatial coordinates would incorrectly predict that standing outside a bank vault is highly valuable. This mistake is the result of failing to encode the fact that getting inside the vault has low probability. Geometric boundaries induce discontinuities in an otherwise smooth value function, and such discontinuities can be encoded by representing similarity in terms of geodesic distance (the shortest path along the connectivity graph of the space). This principle extends beyond physical space to arbitrary feature spaces (Mahadevan 2007, Tenenbaum et al. 2000).

Gustafson & Daw (2011) suggested that place cells in the hippocampus (conceived by them as basis functions rather than approximation kernels) encode a geodesic spatial metric, as evidenced by

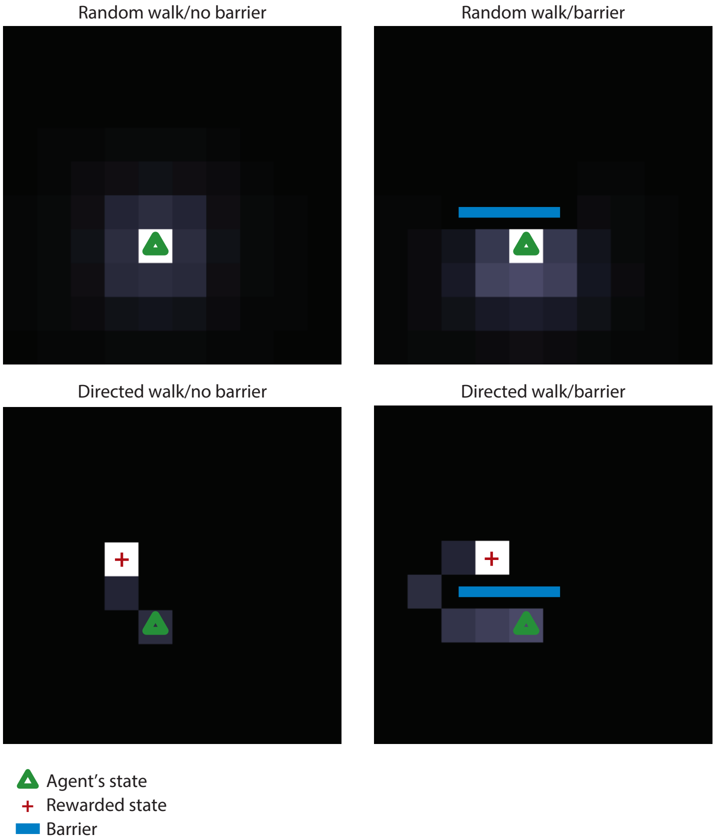

Figure 2

Comparison of the successor representation in different environments. Each image shows the successor representation for the state indicated by the green triangle. The rewarded state is indicated by a red cross. The left column shows an open field. The right column shows a field with a barrier, indicated by the blue line. The top row shows the successor representation for an undirected or random walk induced by a policy that moves through the state space randomly. The bottom row shows the results for a directed policy that moves deterministically along the shortest path to the reward.

systematic spatial distortions in geometrically irregular environments. This idea was extended by Stachenfeld et al. (2014), who argued that a geodesic spatial metric in the hippocampus might arise from a more general predictive representation known as the successor representation (SR) (Dayan 1993). In particular, each state (e.g., spatial location) can be represented in terms of the expected future occupancy of successor states (see **Figure 2** for an illustration). Distance measured in terms of the SR captures the key properties of geodesic distance because passing through boundaries is very unlikely and thus the expected future occupancy is low. The SR goes beyond geodesic distance by also incorporating spatial distortions induced by changes in behavioral policy. An important computational virtue of the SR is that it renders value computation trivial: The value of a state is simply the sum of expected future occupancies for each successor state weighted by the expected reward in that state.

Instead of a basis function for encoding a parametric value approximation, one can think of the SR (or the geodesic distance function) as a particular choice of kernel that encodes the underlying structure of the state space. The Bellman equation implies that states and their successors will tend to have similar values, and thus the SR is a good kernel precisely because it is predictive. The SR can be learned directly from state transitions using TD methods (Dayan 1993, Stachenfeld et al. 2014) and therefore provides a plausible mechanism for adapting the kernel function, with learning, to arbitrary state spaces.

Another aspect of kernel design pertains to multidimensional state spaces: In many real-world tasks, only some of the dimensions are relevant for task performance, necessitating some form of selective attention applied to the feature space. In the kernel view, selective attention would manifest as a distortion of the similarity structure between states depending on the task at hand. This idea has been embodied in several influential exemplar models of categorization, which posit that error-driven learning shapes the mapping from feature inputs to similarity (Kruschke 1992, Love et al. 2004). Related ideas have begun to be explored in RL tasks (Gershman et al. 2010, Niv et al. 2015, Vaidya & Fellows 2015). This research has shown that classical attention areas in the parietal and prefrontal cortices are involved in credit assignment to stimulus features on the basis of reward history. Although researchers have offered an account of this phenomenon in terms of model-free RL, it is possible that the same dimensional attention filter impinges on the kernel used by episodic RL. Again, this would be advantageous because applying the attentional filter at choice time rather than at encoding time reduces the need for relearning values once appropriate dimensional attention is discovered.

Learning with Sparse Data

Another advantage of various sorts of episodic estimation is that they can succeed (relatively speaking) in the extreme low-data limit when model-based and model-free learning fail, as demonstrated in simulations by Lengyel & Dayan (2007). This analysis is consistent with evidence for a shift in behavioral control from the hippocampus to the striatum over the course of training in a variety of tasks (Packard & McGaugh 1996, Poldrack et al. 2001), although these tasks do not specifically isolate an episodic RL strategy.

Some evidence suggests that the hippocampus plays a special role in one-shot learning in decision tasks. Lee et al. (2014) found that humans could learn a novel stimulus-reward outcome after a single observation, and this rapid learning selectively recruited the hippocampus. Rapid learning was also associated with increased coupling between the hippocampus and ventrolateral prefrontal cortex, which was interpreted as evidence supporting an earlier hypothesis that the ventrolateral prefrontal cortex acts as a metacontroller arbitrating between different RL systems (Lee et al. 2014).

The statistically minded reader may object that nonparametric approximations such as kernel density estimation are typically less data-efficient than parametric methods, which is paradoxical in light of our claim that such approximations may be utilized in the low-data limit. It is true that strong parametric assumptions (such as the Markovian assumption) can offer an inductive bias to guide and discipline inference, but this bias is only useful to the extent that the assumptions are correct. Given the twin problems of high dimensionality and state aliasing in the natural environment, it may well be that standard parametric assumptions can only be relied upon if they are validated and tuned by an initial learning phase that identifies relevant dimensions and stimulus history. Furthermore, although the convergence rate of nonparametric approximations is typically slower, such approximations achieve an asymptotically lower error because of their superior flexibility (Wasserman 2006). This is an example of a bias–variance trade-off (Geman et al. 1992): Nonparametric methods more closely approximate the value function (lower bias) at the expense of poorer generalization (higher variance). The purpose of kernel smoothing is precisely to reduce variance by introducing bias (i.e., regularization). If the value function is itself smooth and this smoothness is well matched to the kernel function, the added bias will be small; as discussed in the previous section, kernel smoothing should be strongest across states with similar expected values, a point that can be made precise using the theory of reproducing kernel Hilbert spaces (Schölkopf & Smola 2002). From this discussion, we can posit that episodic RL should perform relatively well in the low-data limit when the value function cannot be well approximated by a parametric family but the values are nonetheless smooth over the state space in a way that is captured by the kernel.

INTERACTIONS BETWEEN LEARNING SYSTEMS

A central theme in contemporary research on RL is the interplay between multiple learning and control systems (Daw et al. 2005, Dolan & Dayan 2013). Much of this research has focused on the principles guiding competitive interactions between model-free and model-based systems—for instance, under what circumstances is it worth engaging in model-based deliberation versus simply acting according to previously learned model-free preferences (Daw et al. 2005, Keramati et al. 2011)—but the full story is more complex and unsettled, particularly in light of the suggested involvement of episodic memory. First, the possibility of additional influences extends the arbitration questions: When should the brain consult episodes versus plan using a previously learned map or model, and which episodes should be consulted? Second, the influences may interact in ways other than simple competition. For instance, as discussed below in this section, in addition to being used to compute values at decision time, episodes may also be useful for offline training of model-free values, e.g., during sleep. Third, and relatedly, all of these considerations may complicate or confound the working of the model-free and model-based systems as they have previously been conceived. In particular, the cognitive and computational bases for putatively model-based choice are as yet underdetermined, and at least some of what has been taken as model-based behavior may arise from some of these episodic influences.

As we have made clear, episodic RL may well constitute yet another system alongside (or as part of) the model-based and model-free systems. Indeed, in other research, influences of individual episodes on choice may have been mistaken for either model-free or model-based learning, which are typically assumed to instead depend on statistical summaries learned over many episodes. For instance, in one-step bandit choice tasks, memory for individual recent episodes can support trial-by-trial choice adjustment that appears similar to model-free incremental learning of action values (Bornstein et al. 2015, Collins & Frank 2012, Erev et al. 2008).

Episodic influences may also have masqueraded as model-based in a number of ways. For example, in multistep sequential tasks, episodic snapshots of individual trajectories also contain information about the sequential state-state map of the task and may support behavior that has the signatures of map- or model-based choice (Daw et al. 2011, Tolman 1948) without actual use of a statistical world model (e.g., Gershman et al. 2014). Indeed, the idea that planning by mental simulation is supported by episodic rather than (or in addition to) semantic representations is a prominent proposal in the cognitive neuroscience of hippocampal function (Hassabis & Maguire 2009, Schacter et al. 2012).

As we have already described, episodic and model-free RL also appear to compete with each other, much like model-based and model-free RL are thought to. Such competition might be understood as a third system or an episodic aspect to the model-based system. Successful episodic memory on individual trials is negatively correlated with sensitivity to reward history and neural prediction error signals (Wimmer et al. 2014; for contrasting results, see Murty et al. 2016). Interfering effects of episodic memory on reward-guided choice can also be directly induced by adding incidental reminders of past actions (Bornstein et al. 2015). More generally, hippocampal involvement in behavioral control tends to predominate early in training, whereas striatal involvement predominates later in training (Packard & McGaugh 1996, Poldrack et al. 2001).

These competitive interactions fit with the picture of largely independent systems vying for behavioral control, with a metacontroller arbitrating between the three (or two) systems according to their relative efficacy at different points during training. In particular, episodic RL may be primarily useful early in training, when parametric value approximations break down due to the sparsity of data and complexity of the state space (Lengyel & Dayan 2007). In all these respects, episodic RL as we have envisioned it echoes features that have also been attributed to model-based RL. Although it seems unlikely that episodic RL alone can account for all of the manifestations of model-based RL, these two putative systems have not been clearly teased apart in the same way that they (collectively) have been dissociated from model-free learning. Doing so will require more precise identification of influences on behavior and brain activity that are verifiably tied to the retrieval of individual episodes versus statistical summaries of them as in a map or world model. If individual episodes are found to directly contribute to evaluation, then this will necessitate fleshing out the emerging theoretical and experimental account of competition between model-based and model-free influences—broadly speaking, thought to reflect a rational speed–accuracy trade-off induced by spending time recomputing better action values (Keramati et al. 2011)—to also weigh the relative costs and benefits of consulting raw episodes for these recomputations versus a summary model.

The influences of episodic memory may also crosscut the model-based and model-free distinction, complicating the picture still further. For instance, the striatum and hippocampus may interact cooperatively as well as competitively (for a review, see Pennartz et al. 2011). Evidence suggests that replay of memories (Lansink et al. 2009) and oscillatory dynamics (van der Meer & Redish 2011) in the two regions are coordinated. Human neuroimaging studies have demonstrated functional connectivity between the hippocampus and striatum during virtual navigation (Brown et al. 2012) and context-dependent decision making (Ross et al. 2011).

One functional explanation for some of these interactions is that they support synergistic influences of episodic memory on model-free values. Such interactions would further leverage episodic memory for choice (beyond the nonparametric value computation) and also produce choices that might, again, appear to mimic some of the behaviors of a model-based system. Model-free RL is, in its traditional conception, limited to learning from direct experience, which renders it inflexible. For example, separately experiencing different parts of an environment will result in a disjointed model-free value function, in which the consistency of values implied by the Bellman equation is violated at the part boundaries. One of the traditional signatures of a model-based system is the ability to stitch these parts together by using them to build a world model that can then be used to simulate sequences of state transitions and rewards that were never experienced together (Shohamy & Daw 2015). However, another way to achieve the same effect is to feed such ersatz experience to a model-free learner, which can then use it in the same way as it would use actual experience to update its stored values. This can be achieved without even building a world model by simply replaying snippets of experience from episodic memory, interleaved across the otherwise separate experiences. Such a replay mechanism is another way (in addition to nonparametric evaluation) in which episodic memories might influence choice, in this case by driving model-free value learning. This hybrid architecture was originally proposed in the machine learning literature by Sutton (1991), who referred to it as Dyna.

Gershman and colleagues (2014) reported behavioral evidence that valuation in humans is supported by Dyna-like offline replay. In these experiments, participants separately learned different parts of a single MDP and were then given a retrospective revaluation test to see if their decisions reflected an integrated value. The experiment indeed found evidence for revaluation, which has typically been taken as a signature of model-based value computation. However, the experiments showed that the extent of successful revaluation was sensitive to several manipulations designed to affect Dyna-style offline replay but irrelevant to model-based choice (in the sense of just-in-time computation of decision variables by mental simulation at choice time). In particular, revaluation was disrupted by placing people under cognitive load during the learning, rather than during the subsequent choice phase, using a secondary task. The deleterious effects of load could be mitigated simply by giving people a brief period of quiescence (listening to classical music) before the revaluation test, consistent with the operation of an offline simulation process.

Neuroimaging studies (Kurth-Nelson et al. 2015, Wimmer & Shohamy 2012) also demonstrate that successful revaluation in a similar integration task is supported by memories retrieved at learning (rather than choice) time. More generally (though without being linked to decisions or learning), replay of the neural responses to previous experiences has repeatedly been observed in neuronal recordings from the hippocampus during quiet rest, sleep, and even ongoing behavior (Carr et al. 2011, Skaggs & McNaughton 1996). These phenomena suggest that the hippocampus may be a candidate for the neural substrate of replay-based learning. However, in all these cases, including the human experiments, it is not yet wholly clear whether the memories being retrieved are episodic (e.g., in the sense of autobiographical snapshots of individualized events) or reflect more semantic knowledge derived from the statistics of multiple episodes (e.g., a statistical world model).

RELATIONSHIP TO OTHER FRAMEWORKS

Case-Based Decision Theory and Decision by Sampling

Research on behavioral economics has explored the role of memory in decision making, focusing on one-shot decision problems rather than the sequential problems on which we have focused. The starting point of this work is a critique of expected utility theory, the cornerstone of neoclassical economics, which assumes that a decision maker will consider all possible states of the world and all possible outcomes so as to average over these in computing expected value. As pointed out by Gilboa & Schmeidler (2001), many real-world situations fit the expected utility framework poorly: The sets of states and outcomes are not readily available to the decision maker. For example, the choice of a nanny would require the enumeration of all possible nanny profiles and all possible consequences of hiring a particular nanny. These sets are, for all practical purposes, infinite. To address this problem, Gilboa & Schmeidler (2001) developed a case-based decision theory (CBDT), drawing upon a venerable tradition in cognitive science (Riesbeck & Schank 1989).

A basic primitive of this theory is the case, consisting of a decision problem, an act, and an outcome. Previously observed cases constitute memories. The decision maker is endowed with a similarity function on decision problems and a utility function on outcomes and is assumed to rank acts for a new decision problem by comparing it to previous cases using the similarity function. This formulation does not require the exhaustive enumeration of states and outcomes, only the retrieval of a subset from memory. Interestingly, the ranking mechanism is a form of kernel-based value estimation, with the similarity function corresponding to a kernel and the cases corresponding to episodes.

The similarity function posited by CBDT effectively determines what memories are available. For simplicity, we can imagine that the similarity is zero for some memories, so that these memories are not retrieved into the available subset, and a constant value for all the retrieved memories. In the most basic form of CBDT, the utility assigned to an act is then proportional to the summed utilities of outcomes stored in the subset of retrieved memories for which the act was chosen. This model has interesting implications for the role played by memory in determining reference points because acts will only be judged with respect to available memories (Simonson & Tversky 1992). For example, Simonsohn & Loewenstein (2006) reported that a household moving to a new city will exhibit dramatically different spending on rent depending on the distribution of rents in their city of origin. In related theoretical work, P. Bordalo, N. Gennaioli, and A. Shleifer (unpublished manuscript) formulated a memory-based model of decision making that allows retrieved memories to influence the decision maker's reference points.

Stewart et al. (2006) took this logic a step further in their decision by sampling theory by arguing that all decision-theoretic quantities (utility, probability, temporal duration, etc.) are based on samples from memory. They demonstrated that the descriptive parameterization of these quantities in prospect theory (Kahneman & Tversky 1979) can be empirically derived from their ecological distribution (a proxy for their availability in memory). For example, Stewart et al. (2006) found that the distribution of credits to bank accounts (a measure of the ecological distribution of gains) is approximately power-law distributed, implying a power-law-revealed utility function under the assumption that the utility function reflects the relative rank of gains. This analysis reproduces the curvature of the utility function proposed by Kahneman & Tversky (1979) on purely descriptive grounds to explain risk aversion; analogous considerations about the relative distribution of debits explain loss aversion.

The idea that subjective utility is computed relative to a memory-based sample has profound implications for models of decision making. It suggests that there is no stable valuation mechanism that consistently obeys the axioms of rational choice. This idea is grounded in a set of psychological principles that extend far beyond economic decisions. Essentially all judgments, ranging from the psychophysics of magnitude, duration, and pain to causal reasoning and person perception, are relative: The same object can be perceived as dramatically different depending on contextual factors that determine a comparison set (Kahneman & Miller 1986, Stewart et al. 2005). This point has not been lost on marketing researchers, who have long recognized the importance of comparison (or consideration) set composition in consumer choice (Bettman 1979, Lynch & Srull 1982, Nedungadi 1990).

Contingent Sampling Models and Instance-Based Learning

Although most economic models have been developed to explain decisions from description (e.g., explicitly described lotteries), RL paradigms typically involve decisions from experience (in which the lottery structure must be learned). Behavioral economists have also studied experiential learning in bandit-like problems in a literature that is largely separate from the study of RL. The most important finding stressed here is that experiential learning often produces striking divergences from description-based decisions (Hertwig & Erev 2009). For example, the classic description-based experiments of Kahneman & Tversky (1979) demonstrated apparent overweighting of rare events, but experience-based experiments have found the opposite phenomenon: underweighting of rare events (e.g., Barron & Erev 2003, Hau et al. 2008). Biele and colleagues (2009) have argued that this underweighting is the result of contingent sampling from memory, in which samples are drawn based on similarity to the current situation. Because rare events are less likely to appear in the sampled set, these events will be relatively neglected. This model can also explain a number of other puzzling behaviors, such as overconfidence (due to a biased estimate of variance from small samples) and inertia (tendency to repeat previous choices; Biele et al. 2009). Gonzalez and colleagues (Gonzalez & Dutt 2011, Gonzalez et al. 2003) have developed closely related instance-based learning models. The important point for the present discussion is that the samples resemble episodes, and the sampling process itself effectively implements a form of kernel smoothing and thus fits into our general framework.

If decisions from experience depend on some form of contingent sampling, then we should expect that memory biases will influence decisions. Ludvig, Madan, and their colleagues (Ludvig et al. 2015, Madan et al. 2014) have shown that the bias to recall extremely positive or negative events is systematically related to risk preferences. In one set of experiments (Madan et al. 2014), individual differences in the tendency to recall extreme events was positively correlated with preference for risky gains and negatively correlated with preference for risky losses. Another experiment (Ludvig et al. 2015) manipulated memory using a priming cue and showed that priming past wins promotes risk seeking. On the theoretical side, Lieder et al. (2014) have shown how a sampling strategy that overweights extreme events is rational when the goal is to minimize the variance of the expected utility estimate from a limited number of samples.

CONCLUSIONS

We have reviewed the current cognitive neuroscience conception of RL, in which model-based and model-free systems compete (and sometimes cooperate) for control of behavior. This dual-system architecture is motivated computationally by the need to balance speed and flexibility, but we have argued that neither system (at least as traditionally conceived) is designed to perform well in high-dimensional, continuous, partially observable state spaces when data are sparse and observations have dependencies over long temporal distances. Unhappily, this situation may be characteristic of many real-world learning problems. A third system—episodic RL—may offer a partial solution to these problems by implementing a form of nonparametric value function approximation. As we have shown, this notion can tie together many disparate observations about the role of episodic memory in RL. Nonetheless, our theory is still largely speculative. We have framed it abstractly to highlight the generality of the ideas, but to make progress the theory must first be more precisely formalized so that it can make quantitative predictions. We expect this to be an exciting frontier for research, both theoretical and experimental, in the near future.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors are grateful to Daphna Shohamy for longstanding collaborative research underlying many of the ideas in this review and helpful comments on the manuscript. The authors also thank Rahul Bhui for comments on an earlier draft of the manuscript. The authors' research contributing to this review is funded by the National Institute on Drug Abuse, grant number R01DA038891 (N.D.D.); Google DeepMind (N.D.D.); and the Center for Brains, Minds and Machines (CBMM) via National Science Foundation Science and Technology Centers award CCF-1231216 (S.J.G.).

LITERATURE CITED

Adams CD. 1982. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol.* 34:77–98

Alexander GE, Crutcher MD. 1990. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.* 13:266–71

Barron G, Erev I. 2003. Small feedback-based decisions and their limited correspondence to description-based decisions. *J. Behav. Decis. Making* 16:215–33

Bayer HM, Glimcher PW. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129–41

Bellman R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press

Bertsekas DP, Tsitsiklis JN. 1996. *Neuro-Dynamic Programming*. Nashua, NH: Athena Sci.

Bettman JR. 1979. *Information Processing Theory of Consumer Choice*. Boston: Addison-Wesley

Biele G, Erev I, Ert E. 2009. Learning, risk attitude and hot stoves in restless bandit problems. *J. Math. Psychol.* 53(3):155–67

Bornstein AM, Khaw MW, Shohamy D, Daw ND. 2015. What's past is present: Reminders of past choices bias decisions for reward in humans. *bioRxiv* 033910. doi: 10.1101/033910

Botvinick MM, Niv Y, Barto AC. 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113:262–80

Braver TS, Cohen JD. 2000. On the control of control: the role of dopamine in regulating prefrontal function and working memory. In *Control of Cognitive Processes: Attention and Performance XVIII*, ed. S Monsell, J Driver, pp. 713–37. Cambridge, MA: MIT Press

Brogden W. 1939. Sensory pre-conditioning. *J. Exp. Psychol.* 25:323–32

Brown TI, Ross RS, Tobyne SM, Stern CE. 2012. Cooperative interactions between hippocampal and striatal systems support flexible navigation. *NeuroImage* 60:1316–30

Carr MF, Jadhav SP, Frank LM. 2011. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* 14:147–53

Cohen JD, Braver TS, Brown JW. 2002. Computational perspectives on dopamine function in prefrontal cortex. *Curr. Opin. Neurobiol.* 12:223–29

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N. 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482:85–88

Collins AG, Brown JK, Gold JM, Waltz JA, Frank MJ. 2014. Working memory contributions to reinforcement learning impairments in schizophrenia. *J. Neurosci.* 34:13747–56

Collins AG, Frank MJ. 2012. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* 35:1024–35

Cushman F, Morris A. 2015. Habitual control of goal selection in humans. *PNAS* 112:13817–22

Daw ND. 2013. Advanced reinforcement learning. See Glimcher & Fehr 2013, pp. 299–320

Daw ND, Courville AC, Touretzky DS. 2006. Representation and timing in theories of the dopamine system. *Neural Comput*. 18:1637–77

Daw ND, Dayan P. 2014. The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130478

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69:1204–15

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8:1704–11

Daw ND, O'Doherty JP. 2013. Multiple systems for value learning. See Glimcher & Fehr 2013, pp. 393–410

Daw ND, Tobler PN. 2013. Value learning through reinforcement: the basics of dopamine and reinforcement learning. See Glimcher & Fehr 2013, pp. 283–98

Dayan P. 1993. Improving generalization for temporal difference learning: the successor representation. *Neural Comput*. 5:613–24

Dezfouli A, Balleine BW. 2013. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLOS Comput. Biol.* 9:e1003364

Dickinson A, Balleine BW. 2002. The role of learning in the operation of motivational systems. In *Steven's Handbook of Experimental Psychology*, Volume 3: *Learning, Motivation and Emotion*, ed. CR Gallistel, pp. 497–534. New York: John Wiley & Sons. 3rd ed.

Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y. 2013. Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *J. Neurosci.* 33:5797–805

Dolan RJ, Dayan P. 2013. Goals and habits in the brain. *Neuron* 80:312–25

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. 2015. Model-based choices involve prospective neural activity. *Nat. Neurosci.* 18:767–72

Eichenbaum H, Cohen NJ. 2004. *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford, UK: Oxford Univ. Press

Engel Y, Mannor S, Meir R. 2005. Reinforcement learning with Gaussian processes. *Proc. Int. Conf. Mach. Learn., 22nd, Bonn, Ger.*, pp. 201–8. New York: Assoc. Comput. Mach.

Erev I, Ert E, Yechiam E. 2008. Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *J. Behav. Decis. Making* 21:575–97

Everitt BJ, Robbins TW. 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* 8:1481–89

Ezzyat Y, Davachi L. 2011. What constitutes an episode in episodic memory? *Psychol. Sci.* 22(2):243–52

Foster DJ, Morris RGM, Dayan P. 2000. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* 10:1–16

Frank MJ, Seeberger LC, O'Reilly RC. 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306:1940–43

Fu W-T, Anderson JR. 2008. Solving the credit assignment problem: explicit and implicit learning of action sequences with probabilistic outcomes. *Psychol. Res.* 72:321–30

Gabrieli JD. 1998. Cognitive neuroscience of human memory. *Annu. Rev. Psychol.* 49:87–115

Gärtner T, Lloyd JW, Flach PA. 2004. Kernels and distances for structured data. *Mach. Learn.* 57:205–32

Geman S, Bienenstock E, Doursat R. 1992. Neural networks and the bias/variance dilemma. *Neural Comput.* 4:1–58

Gershman SJ, Blei DM, Niv Y. 2010. Context, learning, and extinction. *Psychol. Rev.* 117:197–209

Gershman SJ, Markman AB, Otto AR. 2014. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* 143:182–94

Gershman SJ, Norman KA, Niv Y. 2015. Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* 5:43–50

Gilboa I, Schmeidler D. 2001. *A Theory of Case-Based Decisions*. Cambridge, UK: Cambridge Univ. Press

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. 2016. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* 5:e11305

Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66:585–95

Glimcher PW, Fehr E. 2013. *Neuroeconomics: Decision-Making and the Brain*. Cambridge, MA: Academic

Gonzalez C, Dutt V. 2011. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychol. Rev.* 118:523–51

Gonzalez C, Lerch JF, Lebiere C. 2003. Instance-based learning in dynamic decision making. *Cogn. Sci.* 27:591–635

Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB. 2010. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14:357–64

Gustafson NJ, Daw ND. 2011. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLOS Comput. Biol.* 7:e1002235

Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A. 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* 28:5623–30

Hart AS, Rutledge RB, Glimcher PW, Phillips PE. 2014. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J. Neurosci.* 34:698–704

Hassabis D, Maguire EA. 2009. The construction system of the brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364(1521):1263–71

Hau R, Pleskac TJ, Kiefer J, Hertwig R. 2008. The description-experience gap in risky choice: the role of sample size and experienced probabilities. *J. Behav. Decis. Making* 21:493–518

Hertwig R, Erev I. 2009. The description-experience gap in risky choice. *Trends Cogn. Sci.* 13:517–23

Houk JC, Adams JL, Barto AG. 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia*, ed. JC Houk, JL Davis, DG Beiser, pp. 249–70. Cambridge, MA: MIT Press

Huys QJ, Lally N, Faulkner P, Eshel N, Seifritz E, et al. 2015. Interplay of approximate planning strategies. *PNAS* 112:3098–103

Jäkel F, Schölkopf B, Wichmann FA. 2009. Does cognitive science need kernels? *Trends Cogn. Sci.* 13:381–88

Johnson A, Redish AD. 2007. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* 27:12176–89

Kaelbling LP, Littman ML, Cassandra AR. 1998. Planning and acting in partially observable stochastic domains. *Artif. Intell.* 101:99–134

Kahneman D, Miller DT. 1986. Norm theory: comparing reality to its alternatives. *Psychol. Rev.* 93:136–53

Kahneman D, Tversky A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47:263–91

Kearns MJ, Singh SP. 2000. "Bias-variance" error bounds for temporal difference updates. *Proc. Annu. Conf. Comput. Learn. Theory, 13th, Stanford, CA*, pp. 142–47. New York: Assoc. Comput. Mach.

Keramati M, Dezfouli A, Piray P. 2011. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Comput. Biol.* 7:e1002055

Knowlton BJ, Mangels JA, Squire LR. 1996. A neostriatal habit learning system in humans. *Science* 273:1399–402

Kruschke JK. 1992. ALCOVE: an exemplar-based connectionist model of category learning. *Psychol. Rev.* 99:22–44

Kurth-Nelson Z, Barnes G, Sejdinovic D, Dolan R, Dayan P. 2015. Temporal structure in associative retrieval. *eLife* 4:e04919

Lansink CS, Goltstein PM, Lankelma JV, McNaughton BL, Pennartz CM. 2009. Hippocampus leads ventral striatum in replay of place-reward information. *PLOS Biol.* 7:e1000173

Lau B, Glimcher PW. 2005. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* 84:555–79

Lee SW, Shimojo S, O'Doherty JP. 2014. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81:687–99

Lengyel M, Dayan P. 2007. Hippocampal contributions to control: the third way. *Adv. Neural Inf. Process. Syst.*, 20:889–96

Lieder F, Hsu M, Griffiths TL. 2014. The high availability of extreme events serves resource-rational decision-making. *Proc. Ann. Conf. Cogn. Sci. Soc., 36th, Quebec City, Can.*, pp. 2567–72. Wheat Ridge, CO: Cogn. Sci. Soc.

Love BC, Medin DL, Gureckis TM. 2004. SUSTAIN: a network model of category learning. *Psychol. Rev.* 111:309–32

Ludvig EA, Madan CR, Spetch ML. 2015. Priming memories of past wins induces risk seeking. *J. Exp. Psychol. Gen.* 144:24–29

Ludvig EA, Sutton RS, Kehoe EJ. 2008. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* 20:3034–54

Lynch JG Jr., Srull TK. 1982. Memory and attentional factors in consumer choice: concepts and research methods. *J. Consum. Res.* 9:18–37

Madan CR, Ludvig EA, Spetch ML. 2014. Remembering the best and worst of times: memories for extreme outcomes bias risky decisions. *Psychon. Bull. Rev.* 21:629–36

Mahadevan S, Maggioni M. 2007. Proto-value functions: a Laplacian framework for learning representation and control in Markov decision processes. *J. Mach. Learn. Res.* 8:2169–231

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529–33

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16:1936–47

Murty VP, Feldman Hall O, Hunter LE, Phelps EA, Davachi L. 2016. Episodic memories predict adaptive value-based decision-making. *J. Exp. Psychol. Gen.* 145:548–58

Nedungadi P. 1990. Recall and consumer consideration sets: influencing choice without altering brand evaluations. *J. Consum. Res.* 17:263–76

Niv Y. 2009. Reinforcement learning in the brain. *J. Math. Psychol.* 53:139–54

Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, et al. 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* 35:8145–57

Nosofsky RM. 1986. Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol. Gen.* 115:39–57

O'Keefe J, Nadel L. 1978. *The Hippocampus as a Cognitive Map*. Oxford, UK: Clarendon Press

O'Reilly RC, Frank MJ. 2006. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput*. 18:283–328

Ormoneit D, Sen Ś. 2002. Kernel-based reinforcement learning. *Mach. Learn.* 49:161–78

Otto AR, Gershman SJ, Markman AB, Daw ND. 2013a. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* 24:751–61

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. 2013b. Working-memory capacity protects model-based learning from stress. *PNAS* 110:20941–46

Packard MG, McGaugh JL. 1996. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.* 65:65–72

Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, et al. 2016. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat. Neurosci.* 19:845–54

Pennartz CMA, Ito R, Verschure PFMJ, Battaglia FP, Robbins TW. 2011. The hippocampal-striatal axis in learning, prediction and goal-directed behavior. *Trends Neurosci*. 34:548–59

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442:1042–45

Pfeiffer BE, Foster DJ. 2013. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497:74–79

Poldrack RA, Clark J, Pare-Blagoev E, Shohamy D, Moyano JC, et al. 2001. Interactive memory systems in the human brain. *Nature* 414:546–50

Rao RP. 2010. Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* 4:146

Redish AD. 2004. Addiction as a computational process gone awry. *Science* 306:1944–47

Riesbeck CK, Schank RC. 1989. *Inside Case-based Reasoning*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Ross RS, Sherrill KR, Stern CE. 2011. The hippocampus is functionally connected to the striatum and orbitofrontal cortex during context dependent decision making. *Brain Res*. 1423:53–66

Sadacca BF, Jones JL, Schoenbaum G. 2016. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife* 5:e13665

Schacter DL, Addis DR, Hassabis D, Martin VC, Spreng RN, Szpunar KK. 2012. The future of memory: remembering, imagining, and the brain. *Neuron* 76:677–94

Schölkopf B, Smola AJ. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press

Schonberg T, O'Doherty JP, Joel D, Inzelberg R, Segev Y, Daw ND. 2010. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. *NeuroImage* 49:772–81

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science* 275:1593–99

Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R. 2012. Serotonin selectively modulates reward value in human decision-making. *J. Neurosci.* 32:5833–42

Shohamy D, Daw ND. 2015. Integrating memories to guide decisions. *Curr. Opin. Behav. Sci.* 5:85–90

Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA. 2005. The role of dopamine in cognitive sequence learning: evidence from Parkinson's disease. *Behav. Brain Res.* 156:191–99

Shohamy D, Wagner AD. 2008. Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* 60:378–89

Simonson I, Tversky A. 1992. Choice in context: tradeoff contrast and extremeness aversion. *J. Mark. Res.* 29:281–95

Simonsohn U, Loewenstein G. 2006. Mistake #37: the effect of previously encountered prices on current housing demand. *Econ. J.* 116:175–99

Skaggs WE, McNaughton BL. 1996. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* 271:1870–73

Solway A, Botvinick MM. 2015. Evidence integration in model-based tree search. *PNAS* 112:11708–13

Squire LR. 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* 99:195–231

Stachenfeld KL, Botvinick M, Gershman SJ. 2014. Design principles of the hippocampal cognitive map. *Adv. Neural Inf. Process. Sys.* 27:2528–36

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. 2013. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16:966–73

Stewart N, Brown GD, Chater N. 2005. Absolute identification by relative judgment. *Psychol. Rev.* 112:881–911

Stewart N, Chater N, Brown GD. 2006. Decision by sampling. *Cogn. Psychol.* 53:1–26

Sutton RS. 1988. Learning to predict by the methods of temporal differences. *Mach. Learn.* 3:9–44

Sutton RS. 1991. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull.* 2:160–63

Sutton RS, Barto AG. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press

Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–23

Todd MT, Niv Y, Cohen JD. 2008. Learning to use working memory in partially observable environments through dopaminergic reinforcement. *Adv. Neural Inf. Process. Sys.* 21:1689–96

Tolman EC. 1948. Cognitive maps in rats and men. *Psychol. Rev.* 55:189–208

Tulving E. 1972. Episodic and semantic memory 1. In *Organization and Memory*, ed. E Tulving, W Donaldson, pp. 381–402. New York: Academic

Vaidya AR, Fellows LK. 2015. Ventromedial frontal cortex is critical for guiding attention to reward-predictive visual features in humans. *J. Neurosci.* 35:12813–23

van der Meer MA, Redish AD. 2011. Theta phase precession in rat ventral striatum links place and reward information. *J. Neurosci.* 31:2843–54

Wasserman L. 2006. *All of Nonparametric Statistics*. Berlin: Springer Science & Business Media

Wimmer GE, Braun EK, Daw ND, Shohamy D. 2014. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J. Neurosci.* 34:14901–12

Wimmer GE, Shohamy D. 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* 338:270–73

Zilli EA, Hasselmo ME. 2008. Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus* 18:193–209

Annual Review of Psychology

Volume 68, 2017

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Psychology* articles may be found at http://www.annualreviews.org/errata/psych