Published in final edited form as:

Psychol Bull. 2014 March ; 140(2): 466–486. doi:10.1037/a0033455.

Navigating complex decision spaces: Problems and paradigms in sequential choice

Matthew M. Walsh^{a,1} and John R. Anderson^b

^aAir Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433

^bCarnegie Mellon University, Department of Psychology, Pittsburgh, PA 15213

Abstract

To behave adaptively, we must learn from the consequences of our actions. Doing so is difficult when the consequences of an action follow a delay. This introduces the problem of temporal credit assignment. When feedback follows a sequence of decisions, how should the individual assign credit to the intermediate actions that comprise the sequence? Research in reinforcement learning provides two general solutions to this problem: model-free reinforcement learning and model-based reinforcement learning. In this review, we examine connections between stimulus-response and cognitive learning theories, habitual and goal-directed control, and model-free and model-based reinforcement learning. We then consider a range of problems related to temporal credit assignment. These include second-order conditioning and secondary reinforcers, latent learning and detour behavior, partially observable Markov decision processes, actions with distributed outcomes, and hierarchical learning. We ask whether humans and animals, when faced with these problems, behave in a manner consistent with reinforcement learning techniques. Throughout, we seek to identify neural substrates of model-free and model-based reinforcement learning. The former class of techniques is understood in terms of the neurotransmitter dopamine and its effects in the basal ganglia. The latter is understood in terms of a distributed network of regions including the prefrontal cortex, medial temporal lobes cerebellum, and basal ganglia. Not only do reinforcement learning techniques have a natural interpretation in terms of human and animal behavior, but they also provide a useful framework for understanding neural reward valuation and action selection.

Keywords

Reinforcement learning; sequential choice; temporal credit assignment

Introduction

To behave adaptively, we must learn from the consequences of our actions. These consequences sometimes follow a single decision, and they sometimes follow a sequence of decisions. Although single-step choices are interesting in their own right (for a review, see

¹to whom correspondence may be addressed, **Address:** Air Force Research Laboratory, 711 HPW/RHAC – Cognitive Models and Agents Branch, 2620 Q Street, Building 852, Wright-Patterson AFB, OH 45433, **Phone:** (937) 938-4057, **Fax:** (937) 904-8810, matthew.walsh@wpafb.af.mil.

Fu & Anderson, 2006), we focus here on the multi-step case. Sequential choice is significant for two reasons. First, sequential choice introduces the problem of temporal credit assignment (Minsky, 1963). When feedback follows a sequence of decisions, how should one assign credit to the intermediate actions that comprise the sequence? Second, sequential choice makes contact with everyday experience. Successful resolution of the challenges imposed by sequential choice permits fluency in domains where achievement hinges on a multitude of actions. Unsuccessful resolution leads to suboptimal performance at best (Fu & Gray, 2004; Yechiam et al., 2003) and pathological behavior at worst (Herrnstein & Prelec, 1991; Rachlin, 1995).

Parallel Learning Processes

Contemporary accounts of sequential choice build on classic theories of behavioral control. We consider these theories in brief to prepare for our discussion of sequential choice.

In describing how humans and animals select actions, psychologists have long distinguished between habitual and goal-directed behavior (James, 1950/1890). This distinction was made rigorous in stimulus-response and cognitive theories of learning during the behaviorist era. Stimulus-response theories portrayed action as arising directly from associations between stimuli and responses (Hull, 1943). These theories emphasized the role of reinforcement in augmenting habit strength. Cognitive theories, on the other hand, portrayed action as arising from prospective inference over internal models, or maps, of the environment (Tolman, 1932). These theories stressed the interplay between planning, anticipation, and outcome evaluation in goal-directed behavior.

Research in psychology and neuroscience has provided new insight into the distinction between stimulus-response and cognitive learning theories. The emerging view is that two forms of control, habitual and goal-directed, coexist as complementary mechanisms for action selection (Balleine & O'Doherty, 2010; Daw, Niv, & Dayan, 2005; Doya, 1999; Rangel, Camerer, & Montague, 2008). Although both forms of control allow the individual to attain desirable outcomes, behavior is only considered goal-directed if (1) the individual has reason to believe that an action will result in a particular outcome, and (2) the individual has reason to pursue that outcome. Evidence for this distinction comes from animal conditioning studies, which show that different factors engender habitual and goal-directed control (Balleine & O'Doherty, 2010). Further evidence comes from physiological studies, which show that different neural structures are necessary for the expression of habitual and goal-directed behavior (Balleine & O'Doherty, 2010).

The same distinction has arisen in the computational field of reinforcement learning (RL). RL addresses the question of how one should act to maximize reward. The key feature of the reinforcement learning problem is that the individual does not receive instruction, but must learn from the consequences of their actions. Two general solutions to this problem exist: model-free RL and model-based RL. Model-free techniques use stored action values to evaluate candidate behaviors, whereas model-based techniques use an internal model of the environment to prospectively calculate the values of candidate behaviors (Sutton & Barto, 1998).

Deep similarities exist between model-free RL, habitual control, and stimulus-response learning theories. All treat stimulus-response associations as the basic unit of knowledge, and all use experience to adjust the strength of these associations. Deep similarities also exist between model-based RL, goal-directed control, and cognitive learning theories. All treat action-outcome contingencies as the basic unit of knowledge, and all use this knowledge to simulate prospective outcomes. The dichotomy between model-free and model-based learning, which has been applied to decisions that require a single action, extends to decisions that involve a sequence of actions.

Scope of Review

The goal of this review is to synthesize research from the fields of cognitive psychology, neuroscience, and artificial intelligence. We focus on model-free and model-based learning. Existing reviews concentrate on the computational or neural properties of these techniques, but no review has systematically evaluated the results of studies that involve multi-step decision making. A spate of recent experiments has shed light on the behavioral and neural basis of sequential choice, however. To that end, we examine the correspondence between predictions of reinforcement learning models and the results of these experiments. The first and second sections of this review introduce the computational and neural underpinnings of model-free RL and model-based RL. The third section examines a range of problems related to temporal credit assignment that humans and animals face. The final section identifies outstanding questions and directions for future research.

Model-Free Reinforcement Learning

Reinforcement learning is not a monolithic technique, but rather a class of techniques designed for a common problem: learning through trial-and-error to act so as to maximize reward. The individual is not told what to do, but instead must select an action, observe the result of performing the action, and learn from the outcome. Because feedback pertains only to the selected action, the individual must sample alternative actions to learn about their values. Also, because the consequences of an action may be delayed, the individual must learn to select actions that maximize immediate *and* future reward.

Computational Instantiation of Model-Free Reinforcement Learning

Prediction—In model-free RL, differences between actual and expected outcomes, or *reward prediction errors*, serve as teaching signals. After the individual receives an outcome, a prediction error is computed,

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

The value r_{t+1} denotes immediate reward, $V(s_{t+1})$ denotes the estimated value of the new world state (i.e., future reward), and $V(s_t)$ denotes the estimated value of the previous state. The temporal discount rate (γ) controls the weighting of future reward. Discounting ensures that when state values are equal, the individual will favor states that are immediately rewarding.

The prediction error equals the difference between the value of the outcome, $[r_{t+1} + \gamma V(s_{t+1})]$, and the value of the previous state, $V(s_t)$. The prediction error is used to update the estimated value of the previous state,

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta_t. \quad (2)$$

The learning rate (α) scales the size of updates. When expectations are revised in this way, the individual can learn to predict the sum of immediate and future rewards. This is called temporal difference (TD) learning.

TD learning relates to the integrator model (Bush & Mosteller, 1955) and the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972), two prominent accounts of animal conditioning and human learning. Like these models, TD learning explains many conditioning phenomena in terms of the discrepancy between actual and expected rewards (e.g., blocking, overshadowing, and conditioned inhibition; for a review, see Sutton & Barto, 1990). TD learning differs, however, in that it is sensitive to immediate and future reward, whereas the integrator model and the Rescorla-Wagner learning rule are only sensitive to immediate reward. Thus, while all three models account for first-order conditioning, TD learning alone accounts for second-order conditioning. These strengths notwithstanding, TD learning fails to account for some of the same conditioning phenomena that challenge the integrator model and the Rescorla-Wagner learning rule (e.g., latent inhibition, sensory preconditioning, facilitated reacquisition after inhibition, and the partial reinforcement extinction effect), a point that we return to throughout the review and in the discussion.

Control—Prediction is only useful insofar as it facilitates selection. The actor/critic model (Sutton & Barto, 1998) advances a two-process account of how humans and animals deal with this control problem (Figure 1). The critic computes and uses prediction errors to learn state values (Eqs. 1 and 2). Positive prediction errors indicate things have gone better than expected, and negative prediction errors indicate things have gone worse than expected. The actor uses prediction errors to adjust preferences, $p(s, a)$,

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \alpha \cdot \delta_t. \quad (3)$$

This learning rule effectively states that the individual should repeat actions that result in greater rewards than usual, and that the individual should avoid actions that result in smaller rewards than usual. As the individual increasingly favors actions that maximize reward, the discrepancy between actual and expected outcomes decreases and learning ceases. Other model-free techniques like Q-learning and SARSA bypass state values entirely and learn action values directly (Sutton & Barto, 1998).

Preferences must still be converted to decisions. This can be done with the softmax selection rule,

$$\pi(s_t, a_t) = \frac{\exp\left(\frac{p(s_t, a_t)}{\tau}\right)}{\sum_{a'} \exp\left(\frac{p(s_t, a')}{\tau}\right)}. \quad (4)$$

The temperature parameter (τ) controls the degree of stochasticity in behavior. Selections become more random as τ increases, and selections become more deterministic as τ decreases. The softmax selection rule allows the individual to exploit knowledge of the best action while exploring alternatives in proportion to their utility. Aside from its computational appeal, the softmax selection rule resembles Luce's choice axiom and Thurstonian theory, two attempts to map preference strengths to response probabilities (Luce, 1977). The softmax selection rule also approximates greedy selection among actions whose utility estimates are subject to continuously varying noise (Fu & Anderson, 2006).

The chief advantage of model-free techniques is that they learn state and action values without a model of the environment (i.e., state transitions and reward probabilities).¹ Another advantage of model-free RL is that action selection is computationally simple. The individual evaluates actions based on stored preferences or utility values. This simplicity comes at the cost of inflexibility, however. Because state and action values are divorced from outcomes, the individual must experience outcomes to update these values. For example, if a rat unexpectedly discovers reward in a maze blind, the value of the blind will increase immediately. The value of the corridor leading to the blind will only increase, however, after the rat has subsequently passed through the corridor and arrived at the revalued location.

Eligibility traces—Eligibility traces mitigate this problem (Singh & Sutton, 1996). When a state is visited, an eligibility trace is initiated. The trace marks the state as eligible for update and fades according to the decay parameter (λ),

$$e_t(s) = \begin{cases} \lambda \cdot e_{t-1}(s) + 1 & \text{if } s = s_t \\ \lambda \cdot e_{t-1}(s) & \text{if } s \neq s_t \end{cases}. \quad (5)$$

Prediction error is calculated in the conventional manner (Eq. 1), but the signal is used to update all states according to their eligibility,

$$V(s) \leftarrow V(s) + \alpha \cdot \delta_t \cdot e_t(s). \quad (6)$$

Separate traces are assigned to state-action pairs, and state-action pairs are also updated according to their eligibility. Eligibility traces allow prediction errors to pass beyond immediate states and actions, and to reach other recent states and actions.

¹Model-free RL, in turn, has been linked with the sensorimotor loop. Lesions of the rodent dorsolateral striatum, a part of the sensorimotor loop, restore sensitivity to outcome devaluation and contingency degradation, indicating that this region supports habitual control (Yin, Knowlton, & Balleine, 2005). This division suggests a refinement of the actor/critic model where the dorsolateral striatum functions as the actor element.

Neural Instantiation of Model-Free Reinforcement Learning

Dopamine—Researchers have extensively studied the neural basis of model-free RL. Much of this work focuses on dopamine, a neurotransmitter that plays a role in appetitive approach behavior (Berridge, 2007), and is a key component in pathologies of behavioral control such as addiction, Parkinson's disease, and Huntington's disease (Hyman & Malenka, 2001; Montague, Hyman, & Cohen, 2004; Schultz, 1998).² The majority of dopamine neurons are located in two midbrain structures, the substantia nigra pars compacta (SNc) and the medially adjoining ventral tegmental area (VTA). The SNc and VTA receive highly convergent inputs and project to virtually the entire brain. The striatum and prefrontal cortex receive the greatest concentration of dopamine afferents, with the SNc selectively targeting the dorsal striatum and the VTA targeting the ventral striatum and prefrontal cortex.

In a series of studies, Schultz and colleagues demonstrated that the phasic responses of dopamine neurons mirrored reward prediction errors (Schultz, 1998). When a reward was unexpectedly presented, neurons showed enhanced activity at the time of reward delivery. When a conditioned stimulus preceded reward, however, neurons no longer responded to reward delivery. Rather, the dopamine response transferred to the earlier stimulus. Finally, when a reward was unexpectedly omitted following a conditioned stimulus, neurons showed depressed activity at the expected time of reward delivery. These observations motivated the idea that the phasic response of dopamine neurons codes for reward prediction errors (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997).

Basal ganglia—The basal ganglia are a collection of linked subcortical structures that mediate learning and cognitive functions (Packard & Knowlton, 2002). The main input structure of the basal ganglia, the striatum, receives excitatory connections from the frontal cortex. Striatal neurons modulate activity in the thalamus via a direct pathway that passes through the internal segment of the globus pallidus (GPi), and an indirect pathway that passes through the external segment of the globus pallidus (GPe) and to the GPi (Joel, Niv, & Ruppín, 2002). At an abstract level, frontal collaterals convey information about the state of the world to the striatum. Activation from the direct and indirect pathways converges on the thalamus, resulting in facilitation or suppression of action representations.

Dopamine mediates plasticity at corticostriatal synapses (Wickens, Begg, & Arbuthnott, 1996; Reynolds, Hylan, & Wickens, 2001). This has led to the proposal that three factors govern changes in the strength of corticostriatal synapses in a multiplicative fashion: presynaptic depolarization, postsynaptic depolarization, and dopamine concentration (Reynolds & Wickens, 2002). By adjusting the strength of corticostriatal synapses according to a reward prediction error, the basal ganglia comes to facilitate actions that yield positive outcomes, and to suppress actions that do not.

Actor/critic—The actor and critic elements have been associated with the dorsal and ventral subdivisions of the striatum (Joel, Niv, & Ruppín, 2002; O'Doherty et al., 2004).

²These techniques do require an accurate representation of the state space. How humans and animals form such a representation remains an active research topic (Gershman, Blei, & Niv, 2010; Kemp & Tenenbaum, 2008).

Structural connectivity studies support this proposal. The dorsal striatum shares reciprocal connections with motor cortices, whereas the ventral striatum receives information about context, stimuli, and rewards through its connections with limbic and associative structures. Additionally, the ventral striatum, through its connections with the VTA and SNc, can influence activity in midbrain nuclei that project to itself and the dorsal striatum (Haber et al., 1990; 2000; Joel, Niv, & Ruppin, 2002). In a similar way, the critic affects the computation of prediction error signals that reach itself and the actor.

Physiological and lesion studies also implicate the ventral striatum in the acquisition of state values, and the dorsal striatum in the acquisition of action values (Balleine & O'Doherty, 2010; Cardinal et al., 2002; Packard & Knowlton, 2002). Human neuroimaging results further support this proposal. Instrumental conditioning tasks, which require behavioral responses, engage the dorsal and ventral striatum. Classical conditioning tasks, which do not require behavioral responses, mainly engage the ventral striatum (Elliott et al., 2004; O'Doherty et al., 2004; Tricomi, Delgado, & Fiez, 2004).

Model-Based Reinforcement Learning

Model-free techniques assign utility values directly to states and actions. To select among candidate actions, the individual compares the utility of each. Model-based techniques adopt a fundamentally different approach. From experience, the individual learns the reward function—that is, the rewards contained in each state. The individual also learns the state transition function—that is, the mapping between the current state, actions, and resulting states. Using this world model (i.e., the reward function and the transition function), the individual prospectively calculates the utility of candidate actions in order to select among them.

Computational Instantiation of Model-Based Reinforcement Learning

Learning a world model—Model-based approaches require a model of the environment. How can the individual learn about the reward function, $R(s_{t+1})$, and the state transition function, $T(s_t, a_t, s_{t+1})$, from experience? One solution is to compute something akin to the prediction errors used in model-free RL. After the individual enters a state and receives reward r_{t+1} , a reward prediction error is calculated,

$$\delta_{RPE} = r_{t+1} - R(s_{t+1}). \quad (7)$$

This differs from the prediction error in model-free RL because it does not include a term for future reward. The reward prediction error is used to update the value of $R(s_{t+1})$,

$$R(s_{t+1}) \leftarrow R(s_{t+1}) + \alpha_{RPE} \cdot \delta_{RPE}. \quad (8)$$

After the individual arrives at state s_{t+1} , a state prediction error is also calculated,

$$\delta_{SPE} = 1 - T(s_t, a_t, s_{t+1}). \quad (9)$$

The state prediction error is used to update the value of $T(s_t, a_t, s_{t+1})$,

$$T(s_t, a_t, s_{t+1}) \leftarrow T(s_t, a_t, s_{t+1}) + a_{SPE} \cdot \delta_{SPE}. \quad (10)$$

The likelihoods of all states not arrived at are then normalized to ensure that transition probabilities sum to one. The use of error-driven learning to acquire a causal relationship model constitutes a variation of the Rescorla-Wagner learning rule (1972).

Policy search—In model-based RL, information about the reward and transition functions is used to calculate state values,

$$V(s_t) = \sum_{a'} \pi(s_t, a') \sum_{s'} T(s_t, a', s') \cdot \{R(s') + \gamma \cdot V(s')\}. \quad (11)$$

This quantity equals the value of possible outcomes, $\{R(s') + \gamma V(s')\}$, weighted according to their probability given the individual's selection policy, $\pi(s_t, a')$, and the state transition function, $T(s_t, a', s')$. After state values are calculated, actions are evaluated in terms of their immediate and future rewards,

$$Q_{FWD}(s_t, a_t) = \sum_{s'} T(s_t, a_t, s') \cdot \{R(s') + \gamma \cdot V(s')\}. \quad (12)$$

Finally, the selection policy is updated by setting $\pi(s_t, a_t) = \max_a Q_{FWD}(s_t, a)$. In other words, the individual selects the most valuable action in each state.³ By iteratively applying these evaluations over all states and actions, the individual will arrive at an optimal selection policy. This is called policy iteration (Sutton & Barto, 1998).

Policy iteration is but one example of how the individual can use a world model to select actions. The individual can also use transition and reward functions to calculate expected rewards over the next n time steps for all sequences of actions to identify the one that maximizes return (i.e., forward or depth-first search; Daw, Niv, & Dayan, 2005; Johnson & Redish, 2007; Simon & Daw, 2011; Smith, Becker, & Kapur, 2006). Alternatively, working from the desired final state, the individual can reason back to the current state (i.e., backward induction; Busemeyer & Pleskac, 2009). Finally, using the transition and reward functions as a generative world model, the individual can apply Bayesian techniques to infer which policy is most likely to maximize return (Solway & Botvinick, 2012).

The chief advantage of model-based RL is that it efficiently propagates experience to antecedent states and actions. For example, if a rat unexpectedly discovers reward in a maze blind, policy iteration allows the rat to immediately revalue states and actions leading to that blind. Model-based RL has two main drawbacks, however. First, model-based RL requires a complete model of the environment. Second, in environments with many states, the costs of querying the world model become prohibitive in terms of time and computation. The finite

³Two-factor theories also posit a role for dopamine in active avoidance (Dayan, 2012; Maia, 2010). Besides signaling reward, dopamine responses signal the potential for, and success in, avoiding punishment.

capacity of human working memory magnifies this concern: people cannot hold an endlessly branching decision tree in memory.

Neural Instantiation of Model-Based Reinforcement Learning

Reward function—Neurons in the orbitofrontal cortex (OFC) encode stimulus incentive values. OFC responses decrease when participants are satiated on the evoking unconditioned stimulus (Rolls, Kringelbach, & de Araujo, 2003; Valentin, Dickinson, & O'Doherty, 2007), and OFC responses correlate with people's willingness to pay for appetitive stimuli (Plassmann, O'Doherty, & Rangel, 2007). Additionally, experienced and imagined rewards activate the OFC (Bray, Shimojo, & O'Doherty, 2010). These functions have been linked with goal-directed behavior, a position further supported by the observations that OFC lesions abolish devaluation sensitivity (Izquierdo, Suda, & Murray, 2004), and impair animals' ability to associate distinct stimuli with different types of food rewards (McDannald et al., 2011).

Transition function—Several regions provide candidate transition functions. Chief among these is the hippocampus. Rodent navigation studies show that receptive fields of hippocampal pyramidal cells form a cognitive map (O'Keefe & Nadel, 1978). Neurons preferentially fire as the animal traverses specific locations in the environment. Ensemble activity sometimes correlates with positions other than the rodent's current location, however. For instance, when rodents rest after traversing a maze, the sequence of active hippocampal fields 'replays' navigation paths (Foster & Wilson, 2006). Additionally, when rodents arrive at choice points in a maze, they pause and engage in vicarious trial-and-error behavior such as looking down alternate paths (Tolman, 1932). The sequence of active hippocampal fields simultaneously 'preplays' alternate paths at such points, hinting at the involvement of the hippocampus in forward search (Johnson & Redish, 2007).

The acquisition and storage of relational knowledge among non-spatial stimuli also depends on the hippocampus and the surrounding medial temporal lobes (MTL; Bunsey & Eichenbaum, 1996). MTL-mediated learning occurs rapidly, MTL-based memories contain information about associations among stimuli, and MTL-based memories are accessible in and transferrable to novel contexts (Cohen & Eichenbaum, 1993). These features of the MTL coincide with core properties of model-based RL. Interestingly, the hippocampus becomes active when people remember the past and imagine the future, paralleling reports of hippocampal replay and preplay events in rodents (Schacter, Addis, & Buckner, 2007). Thus, among the many types of knowledge it stores, the hippocampus may contain state transition functions that permit forward search in humans as well.

The cerebellum contains a different type of transition function in the form of internal models of the sensorimotor apparatus (Doya, 1999; Ito, 2008). These models allow the sensorimotor system to identify motor commands that will produce target outputs. Movement is not a prerequisite for cerebellar activation, however. Cognitive tasks also engage the cerebellum (Stoodley, 2012; Strick, Dum, & Fiez, 2009). This suggests that the cerebellum contains internal models that contribute to non-motoric planning as well (Imamizu & Kawato, 2009; Ito, 2008).

Lastly, specific basal ganglia structures encode information about relationships between actions and outcomes. Electrophysiological recordings and lesion studies have identified three anatomically and functionally distinct cortico-basal-ganglia loops: a sensorimotor loop that mediates habitual control, an associative loop that mediates goal-directed control, and a limbic loop that mediates the impact of primary reward values on habitual and goal-directed control (Balleine, 2005; Balleine & O'Doherty, 2010). Model-based RL has been linked with the associative loop (Daw, Niv, & Dayan, 2005; Dayan & Niv, 2008; Niv, 2009). Lesions of the rodent prelimbic cortex and dorsomedial striatum, parts of the associative loop, abolish sensitivity to outcome devaluation and contingency degradation, two assays used to establish that behavior is goal-directed (Balleine & O'Doherty, 2010; Ostlund & Balleine, 2005; Yin, Knowlton, & Balleine, 2005).⁴

Researchers have identified homologous areas in the primate brain (Wunderlich, Dayan, & Dolan, 2012). Neuroimaging studies have found that the ventromedial prefrontal cortex (vmPFC) encodes expected reward attributable to chosen actions (Daw et al., 2006; Gläscher, Hampton, & O'Doherty, 2009; Hampton, Bossaerts, & O'Doherty, 2006). Further, activity in the vmPFC is modulated by outcome devaluation (Valentin, Dickinson, & O'Doherty, 2007), and the vmPFC and dorsomedial striatum (the anterior caudate) detect the strength of the contingency between actions and rewards (Liljeholm et al., 2011; Tanaka, Balleine, & O'Doherty, 2008). These findings demonstrate that the vmPFC and dorsomedial striatum are sensitive to the defining properties of goal-directed control.

Decision policy—The prefrontal cortex supports more abstract forms of responses. For example, certain neurons in the lateral prefrontal cortex (latPFC) appear to encode task sets, or “rules” that establish context-dependent mappings between stimuli and responses (Asaad, Rainer, & Miller, 2000; Mansouri, Matsumoto, & Tanaka, 2006; Muhammad, Wallis, & Miller, 2006; White & Wise, 1999). Human neuroimaging experiments underscore the involvement of the latPFC in rule retrieval and use (Bunge, 2004). The task sets evoked in these studies are akin to a literal decision policy.

The PFC also supports sequential choice behavior. For example, in a maze navigation task, neurons in the latPFC exhibited sensitivity to initial, intermediate, and final goal positions prior to movement onset (Mushiaki et al., 2006; Saito et al., 2005). Additionally, human neuroimaging experiments show that the PFC is active in tasks that require planning, and that PFC activation increases with planning difficulty (Anderson, Albert, & Fincham, 2005; Owen, 1997). Moreover, latPFC damage impairs planning and rule-guided behavior while leaving other types of responses intact (Bussey, Wise, & Murray, 2001; Fuster, 1997; Hoshi, Shima, & Tanji, 2000; Owen, 1997; Shallice, 1982). Collectively, these results highlight the involvement of the PFC in facets of model-based RL. That is not to say that the PFC *only* performs model-based RL. Rather, model-based RL draws on a multitude of functions performed by the PFC and other regions as described above.

⁴This formalism does not permit exploration. In model-based RL, exploration is only useful for resolving uncertainty in the world model. Techniques exist for incorporating exploration into model-based RL (Thrun, 1992).

Hybrid Models

The computational literature contains proposals for pairing model-free and model-based approaches. There is evidence that the brain also combines model-free and model-based RL. For example, some computational algorithms augment model-free learning by replaying or simulating experiences offline (e.g., the Dyna-Q algorithm; Lin, 1992; Sutton, 1990). The model-free system treats simulated episodes as real experience to accelerate TD learning. In a similar way, hippocampal replay may facilitate model-free RL by allowing the individual to update cached values offline using simulating experiences (Gershman, Markman, & Otto, *in press*; Johnson & Redish, 2005). Thus, the model-based system may train a model-free controller.

Other computational algorithms use cached values to limit the depth of search in model-based RL (Samuel, 1959). This is especially pertinent in light of the limited capacity of human short-term memory. For example, although a chess player has complete knowledge of the environment, and can enumerate the full state-space to discover the optimal move in principle, chess masters consider a far smaller subspace before acting (de Groot, 1946/1978). This amounts to pruning the branches of the decision tree (Huys et al., 2012). Rather than exhaustively calculating future reward, the individual estimates future reward using heuristics (Newell & Simon, 1972) or cached values from the model-free system (Daw, Niv, & Dayan, 2005). Although neural evidence for such pruning is sparse, one study demonstrated that ventral-striatal neurons responded when rats received reward and when they engaged in vicarious trial-and-error behavior at choice points in a T-maze (van der Meer & Redish, 2009). Thus, the model-free system may contribute information about branch values to a model-based controller.

Problems and Paradigms in Sequential Choice

Model-free and model-based RL provide normative solutions to the problem of temporal credit assignment, and neuroscientific investigations have begun to map components of these frameworks onto distinct neural substrates (Table 1). We now turn to empirical work that presents variants of the temporal credit assignment problem. We ask whether humans and animals can cope with these challenges. If so, how, and if not, why not?

Second-Order Conditioning and Secondary Reinforcers

In the archetypal classical conditioning experiment, a conditioned stimulus (CS) precedes an unconditioned stimulus (US). For example, a dog views a light (the CS) before receiving food (the US). Initially, the US evokes an unconditioned response, such as salivation, but the CS does not. When the CS and US are repeatedly paired, however, the CS comes to evoke a conditioned response, salivation, as well (Pavlov, 1927). Holland and Rescorla (1975) asked whether second-order conditioning was possible – that is, can a CS be used to condition a neutral stimulus? They found that when a neutral stimulus was paired with a CS, the neutral stimulus came to evoke a conditioned response as well.

The spread of reinforcement is also seen in tasks that require behavioral responses. In the standard instrumental conditioning paradigm, an animal performs a response and receives reward. For example, a pigeon presses a lever and is given a food pellet. Skinner (1938)

asked whether a neutral stimulus, once conditioned, could act as a secondary reinforcer – that is, can a CS shape instrumental responses? Indeed, when an auditory click was first associated with food pellets, pigeons learned to press a lever that simply produced the auditory click (Skinner, 1938).

In TD learning, a model-free technique, the CS inherits the value of the US that follows it. Consequently, the CS can condition neutral stimuli (i.e., second-order conditioning), and the CS can support the acquisition of instrumental responses (i.e., secondary reinforcement). By this view, the CS mediates behavior directly through its reinforcing potential. Model-based accounts can also accommodate these results. The individual may learn that a neutral stimulus or action leads to the CS, and that the CS leads to the US. By this view, the CS mediates behavior indirectly through its link with the US.

The question of model-free or model-based RL maps onto the classic question of stimulus-response or stimulus-stimulus association. Some results are consistent with the model-free/stimulus-response position. For example, in higher-order conditioning experiments, animals continue to exhibit a conditioned response to the second-order CS even after the first-order CS is extinguished (Rizley & Rescorla, 1972). This shows that once conditioning is complete, the second-order conditioned response no longer depends on the first-order CS. Other results are consistent with the model-based/stimulus-stimulus position. For example, in sensory preconditioning, the second stimulus is paired with the first stimulus, and the first stimulus is only then paired with the US (Brogden, 1939; Rizley & Rescorla, 1972). Once conditioning is complete, the second-order CS evokes a conditioned response even though it was never paired with the revalued, first-order CS. This shows that conditioning can occur even if the training schedule does not permit the backward propagation of reward to the second-order CS.

Physiological studies—Some of the strongest evidence for TD learning comes from studies of the phasic responses of dopamine neurons to rewards and reward-predicting stimuli. The dopamine response conforms to basic properties of the reward prediction error signal in classical and instrumental tasks (Pan et al., 2005; Schultz, 1998), and during self-initiated movement sequences (Wassum et al., 2012). More nuanced tests substantiate the dopamine prediction error hypothesis. First, dopamine neurons respond to the earliest predictors of reward in classical and instrumental conditioning tasks (Schultz, Apicella, & Ljungberg, 1993). Second, dopamine neurons respond more strongly to stimuli that predict probable rewards (Fiorillo, Tobler, & Schultz, 2003) and rewards with large magnitudes (Morris et al., 2006; Satoh et al., 2003; Tobler, Fiorillo, & Schultz, 2005). Additionally, when probability and magnitude are crossed, dopamine neurons respond to expected value rather than to the constituent parts (Tobler, Fiorillo, & Schultz, 2005). Third, in blocking paradigms, animals fail to learn associations between blocked stimuli and rewards. Accordingly, dopamine neurons respond more weakly to blocked stimuli than to unblocked stimuli (Waelti, Dickinson, & Schultz, 2001). Fourth and finally, animals and humans discount delayed rewards (Frederick, Loewenstein, & O'Donoghue, 2002). Dopamine neurons also respond more weakly to stimuli that predict delayed rewards (Kobayashi & Schultz, 2008; Roesch, Calu, & Schoenbaum, 2007).

fMRI studies—Researchers have attempted to identify prediction error signals in humans using fMRI. Many experiments have examined BOLD responses to parametric manipulations expected to produce prediction errors. Others have used hidden-variable analyses in conjunction with TD models to identify regions where activation correlates with a prediction error signal. Both approaches consistently show that prediction errors modulate activity throughout the striatum, a region densely innervated by dopamine neurons (Berns et al., 2001; Delgado et al., 2003; McClure, Berns, & Montague, 2003; O'Doherty et al., 2004; O'Doherty, Hampton, & Kim, 2007; Pagnoni et al., 2002; Rutledge et al., 2010).

A characteristic feature of the TD learning signal and of dopamine responses is that they propagate back to the earliest outcome predictor. This is also true of BOLD responses. In one illustrative study, participants underwent appetitive conditioning (O'Doherty et al., 2003). Activity in the ventral striatum at the time of reward delivery conformed to a prediction error signal. The BOLD response was maximal when a pleasant liquid was unexpectedly delivered, and the response was minimal when a pleasant liquid was unexpectedly withheld. Activity at the time of stimulus presentation also conformed to a prediction error signal. The BOLD response was greatest when the stimulus predicted delivery of the pleasant liquid.

Other fMRI studies have replicated the finding that conditioned stimuli evoke neural prediction errors during classical and instrumental conditioning (Abler et al., 2006; O'Doherty et al., 2004; Tobler et al., 2006). In one such study, cues predicted rewards with probabilities ranging from 0% to 100% (Abler et al., 2006). Following cue presentation, activity in the nucleus accumbens (NAc) increased as a linear function of reward probability, and following reward delivery, activity increased as a linear function of the unexpectedness of reward. NAc activity also increases with anticipated reward magnitude (Knutson et al., 2005).

In these examples, conditioning was successful. In contrast, when a blocking procedure is used, participants fail to associate blocked stimuli with rewards (Tobler et al., 2006). Paralleling this behavioral result, the ventral putamen showed weaker responses to blocked stimuli than to unblocked stimuli. Additionally, and as anticipated by TD learning, the ventral putamen showed greater responses to rewards that followed blocked stimuli than to rewards that followed unblocked stimuli.

These studies support the notion that the BOLD signal in the striatum conveys a model-free report. One recent study challenges that notion, however (Daw et al. 2011). In that study, participants made two selections before receiving feedback. The first selection led to one of two intermediate states with fixed probabilities (Figure 2). Participants memorized these probabilities in advance. The second selection, made from the intermediate state, was rewarded probabilistically. Participants learned these probabilities during the experiment. Model-based and model-free RL make opposing predictions for how outcomes will influence first-stage selections. Model-free RL credits first- and second-stage selections for outcomes, and so predicts that participants will repeat first-stage selections whenever they receive reward. Model-based RL only credits second-stage selections for outcomes, and so predicts that participants will favor the first-stage selection that leads to the most rewarding

intermediate state. Consequently, model-based RL predicts that the individual will repeat first-stage selections when the expected transition occurs and the trial is rewarded, and when the *unexpected* transition occurs and the trial is *not rewarded*.

Participants' behavior reflected a blend of these predictions: they were more likely to repeat the initial selection when the trial was rewarded, and they did so most often when the initial selection led to the expected intermediate state. To account for these results, Daw et al. (2011) proposed a hybrid model that combined estimated action values from model-free and model-based controllers. Daw et al. generated prediction errors using TD learning. They also generated prediction errors based on the difference between the value of the actual outcome and the value predicted by the model-based controller. BOLD responses in the ventral striatum correlated with the difference between outcomes and model-free predictions, supporting the idea that the ventral striatum is involved in TD learning. BOLD responses further correlated with the difference between outcomes and model-based predictions, however. Daw et al. (2011) concluded that the ventral striatum did not literally implement model-based RL, as this result might suggest. Rather, other regions that implement model-based RL influenced the utility values represented in the ventral striatum.

Electrophysiological studies—Researchers have also attempted to identify neural prediction errors in humans using scalp-recorded event-related potentials (ERPs). Early studies revealed a frontocentral error-related negativity (ERN) that appeared 50 to 100 ms after error commission (Falkenstein et al., 1991; Gehring et al., 1993). Subsequent studies have revealed a frontocentral negativity that appears 200 to 300 ms after the display of aversive feedback (Miltner, Braun, & Coles, 1997). Many features of this feedback-related negativity (FRN) indicate that it relates to reward prediction error. First, the FRN amplitude depends on the difference between actual and expected reward (Holroyd & Coles, 2002; Walsh & Anderson, 2011a, 2011b). Second, the FRN amplitude correlates with post-error adjustment (Cohen & Ranganath, 2007). Third, converging methodological approaches indicate that the FRN originates from the anterior cingulate (Holroyd & Coles, 2002), a region implicated in cognitive control and behavioral selection (Kennerley et al., 2006). These ideas have been synthesized in the reinforcement learning theory of the error-related negativity (RL-ERN), which proposes that midbrain dopamine neurons transmit a prediction error signal to the anterior cingulate, and that this signal reinforces or punishes actions that preceded outcomes (Holroyd & Coles, 2002; Walsh & Anderson, 2012).

According to RL-ERN, outcomes and stimuli that predict outcomes should be able to evoke an FRN. To test this hypothesis, researchers have examined whether predictive stimuli produce an FRN. In some studies, cues perfectly predicted outcomes. In those studies, ERPs were more negative after cues that predicted losses than after cues that predicted wins (Baker & Holroyd, 2009; Dunning & Hajcak, 2007). In other studies, cues provided probabilistic information about outcomes. There too, ERPs were more negative after cues that predicted probable losses than after cues that predicted probable wins (Holroyd, Krigolson, & Lee, 2011; Liao et al., 2011; Walsh & Anderson, 2011a). In each case, the latency and topography of the cue-locked FRN coincided with the feedback-locked FRN.

RL-ERN is but one account of the FRN (Holroyd & Coles, 2002). According to another proposal, the anterior cingulate monitors response conflict (Botvinick et al., 2001; Yeung et al., 2004). Upon detecting coactive, incompatible responses, the anterior cingulate signals the need to increase control in order to resolve the conflict. Consistent with this view, ERP studies have revealed a frontocentral negativity called the N2 that appears when participants must inhibit a response (Pritchard et al., 1991). Source localization studies indicate that the N2, like the FRN, arises from the anterior cingulate (van Veen & Carter, 2002; Yeung et al., 2004). Further, fMRI studies have reported enhanced activity in the anterior cingulate following incorrect responses as well as correct responses under conditions of high response conflict (Carter et al., 1998; Kerns et al., 2004).

RL-ERN focuses on the ERN and the FRN, whereas the conflict monitoring hypothesis focuses on the N2 and the ERN. RL-ERN can be augmented to account for the N2, however, by assuming that conflict resolution incurs cognitive costs, penalizing high conflict states (Botvinick, 2007). Alternatively, high conflict states may have lower expected value because they engender greater error likelihoods (Brown & Braver, 2005).

Goal Gradient

Hull noted that reactions to stimuli followed immediately by rewards were conditioned more strongly than reactions to stimuli followed by rewards after a delay (1932). Based on this observation, he proposed that the power of primary reinforcement transferred to earlier stimuli, producing a spatially extended goal gradient that decreased in strength with distance from reward. The concept of the goal-gradient has inspired much research on maze learning. This theory predicts that errors will occur most frequently at early choice points in mazes. Because future reward is discounted, the values of states and actions most remote from reward are near zero. As the difference between the values of actions decreases, the probability that the individual will select the correct response with a softmax selection rule (Eq. 4) decreases as well.

Consistent with these predictions, rats make the most errors at choice points furthest from reward (Spence, 1932; Tolman & Honzik, 1930). Researchers have since identified other factors that affect maze navigation. For example, subjects enact incorrect responses that anticipate future choice points (Spragg, 1934), they more frequently enter blinds oriented toward the goal (Spence, 1932), and they commit a disproportionate number of errors in the direction of more probable turns (Buel, 1935). Although a multitude of factors affect maze learning, the backward elimination of blinds operates independently of these factors.

Fu and Anderson (2006) asked whether humans also show a goal gradient. In their task, participants navigated through rooms in a virtual maze (Figure 3). Each room contained one object that marked its identity, and two cues that participants chose between. After choosing a cue, participants transitioned to a room that contained a new object and cues. If participants selected the correct cues in Rooms 1, 2, and 3, they arrived at the exit. If they made an incorrect selection in any room, they ultimately arrived at a dead end. Participants only received feedback upon reaching the exit or a dead end, and upon reaching a dead end, they were returned to the last correct room. The goal-gradient hypothesis predicts that errors

will be most frequent in Room 1, followed by Room 2, and then by Room 3. The results of the experiment confirmed this prediction.

Fu and Anderson fit a SARSA model to their data. Like participants, the model produced a negatively accelerated goal gradient. This is a natural consequence of discounting future reward (i.e., application of γ in Eq. 1). The maximum reward that can reach a room, R_D , decreases as a function of its distance (D) from the exit,

$$R_D = r \cdot \gamma^D. \quad (13)$$

The discount term (γ) controls the steepness of the gradient. This function is equivalent to one derived by Spence to describe the goal gradient in rat maze learning (1932).

This interpretation of the goal gradient is in terms of model-free RL. Might a model-based controller also exhibit a goal gradient? Yes. Future reward is discounted in some instantiations of forward search (Eqs. 11 and 12). Consequently, differences among state and action values decrease with their distance from reward. In other versions of model-based RL, future reward is not discounted but error is accrued with each subsequent step in forward search (Daw, Niv, & Dayan, 2005). The compounding of error with increasing search depth would yield a goal gradient as well.

Eligibility traces—Fu and Anderson's model predicted slower learning in Room 1 than was observed. This relates to a weakness of TD learning: when reward follows a delay, credit slowly propagates to distant states and actions. Walsh and Anderson (2011a) directly compared the behavior of TD models with and without eligibility traces in a sequential choice task. Like Fu and Anderson (2006), they found that models without eligibility traces learned the initial choice in a sequence more slowly than participants did. The addition of eligibility traces resolved this discrepancy.

In these examples, deviations between model predictions and behavior were slight. Even moderately complex problems exacerbate the challenge of assigning credit to distant states and actions, however (Janssen & Gray, 2012). For instance, Gray et al. (2006) found that Q-learning required 100,000 trials to match the proficiency of human participants after 50 trials. Eligibility traces greatly accelerate learning in such cases where rewards are delayed by multiple states.

Latent Learning and Detour Behavior

Latent learning—Work aimed at distinguishing between stimulus-response and cognitive learning theories provided early support for model-based RL. Two classic examples are latent learning and detour behavior. Latent learning experiments examine whether individuals can learn about the structure of the environment in the absence of reward. In one such experiment, rats navigated through a compound T-maze until they reached an end box (Blodgett, 1929). Following several unrewarded sessions, food was placed at the end box. After discovering the food, rats committed far fewer errors as they navigated to the end box

in the next trial. This indicates that they acquired information about the structure of the maze during training and in the absence of reward.⁵

In a recent study of latent learning in humans, participants navigated through two intermediate states before arriving at a terminal state (Gläscher et al., 2010). Before performing the task, they learned the transition probabilities leading to terminal states, and they then learned the reward values assigned to terminal states. Gläscher et al. compared the behavior of three models to participants' performance at test. The SARSA model used reward prediction errors to learn action values from experience during the test phase. The FORWARD model used knowledge of the transition and reward probabilities to calculate action values prospectively with policy iteration. Lastly, the HYBRID model averaged action values from the separate SARSA and FORWARD models. Gläscher et al. found that the HYBRID model best matched participants' behavior. The FORWARD component accounted for the fact that participants immediately exercised knowledge of the transition and reward functions, and the SARSA component accounted for the fact that they continued to learn from experience during the test.

By collecting neuroimaging data as participants performed the task, Gläscher et al. could search for neural correlates of model prediction errors. Reward prediction errors (δ_t) generated by the SARSA model correlated with activity in the ventral striatum, corroborating other reports of this region's involvement in model-free RL. State prediction errors (δ_{SPE} ; Eq. 9) generated by the FORWARD model correlated with activity in the intraparietal sulcus and the latPFC, indicating that these regions contribute to the acquisition or storage of a state transition function. The finding that the latPFC represents transitions is consistent with the role of this region in planning (Fuster, 1997; Mushiaki et al., 2006; Owen, 1997; Saito et al., 2005; Shallice, 1982).

Detour behavior—Detour behavior experiments examine how individuals adjust their behavior upon detecting environmental change. In one such study (Tolman & Honzik, 1930), rats selected from three paths, two of which shared a segment leading to the end box (Figure 4). When Path 1 was blocked at point A, rats immediately switched to Path 2, and when Path 1 was blocked at point B, rats immediately switched to Path 3. This indicates that they revised their internal model upon encountering detours, and that they used this revised model to identify the shortest remaining path.

Simon and Daw asked whether humans were as sensitive to detours (2011). In their task, participants navigated through a grid of rooms, some of which contained known rewards. Doors connecting the rooms changed between trials, eliminating old paths and creating new paths. Simon and Daw fit model-free and model-based controllers to each participant's behavior, and found that the model-based controller best accounted for nearly all participants' choices. This result indicates that humans, like rats, update their model of the environment upon encountering detours and that they use this model to infer the shortest path to the goal.

⁵Simply removing the rat from the terminal section of the maze may be rewarding. Control experiments indicate that this is unlikely to account completely for latent learning effects, however (Thistlethwaite, 1951).

Simon and Daw used fMRI to identify neural correlates of model-free and model-based prediction errors. Activity in the dorsolateral PFC and OFC correlated with model-based values. Additionally, activity in the ventral striatum, though correlated with model-free values, was more strongly correlated with model-based values. Simon and Daw also identified regions that responded according to the expected value of the reward in the next room. This analysis revealed significant clusters of activation in the superior frontal cortex and the parahippocampal cortex.

The discovery that model-based value signals correlated with activity in the dorsolateral PFC and OFC is consistent with the purported role of these regions in planning. Additionally, the finding that future reward correlated with activity in the PFC and MTL is consistent with the idea that these regions contribute to goal-directed behavior. The discovery that activity in the ventral striatum correlated more strongly with model-based than with model-free value signals is surprising, however, because this region is typically associated with TD learning. As in Daw et al. (2011), this result need not imply that the striatum itself implements model-based RL. Rather, other regions that implement model-based RL may bias striatal representations of action values.⁶

Partially Observable Markov Decision Processes (POMDP)

Most studies of choice involve Markov decision processes (MDP). The key feature of MDPs is that state transitions and rewards depend on the current state and action, but not on earlier states or actions. An important challenge is to extend learning theories to partially observable Markov decision processes (POMDPs). In POMDPs, the system's dynamics follow the Markov property, but the individual cannot directly observe the system's underlying state. POMDPs can be dealt with in three ways. First, the individual can simply ignore hidden states (Loch & Singh, 1998). Second, the individual can maintain a record of past states and actions to disambiguate the current state (McCallum, 1995). Third, the individual can maintain a belief state vector that contains the relative likelihood of each state (Kaelbling, Littman, & Cassandra, 1998). We consider each of these alternatives in turn.

Eligibility traces—In one task that violated the Markov property, participants selected from two images (Tanaka et al., 2009). For some image pairs, monetary reward or punishment was delivered immediately. For other image pairs, reward or punishment was delivered after three trials (Figure 5). In trials with delayed outcomes, the hidden state is the correctness of the response, which affects the value of the score displayed three trials later. Tanaka et al. asked whether participants could learn correct responses for the outcome-delayed image pairs. Although one response resulted in a future loss, and the other in a future gain, the states that immediately followed both responses mapped onto the same visual percept. Because TD learning calculates future reward based on the value of the state that immediately follows an action (Eq. 1), this creates a credit assignment bottleneck. Yet participants learned the correct responses for the outcome-delayed image pairs.

⁶The basal ganglia are involved in functions beyond reward evaluation. One recent model proposes that the basal ganglia play a general role in the conditional routing of information between cortical areas (Stocco, Lebiere, & Anderson, 2010). By this proposal, striatal activation might reflect the passage of model-based predictions through the basal ganglia, rather than the impact of model-based evaluations on striatal learning.

Tanaka et al. evaluated two computational models. The first used internal memory elements to store the past three decisions. The representation of states in this model included the last three decisions and the current image pair. This representation restores the Markov property, but expands the size of the state space. The second model applied TD learning with eligibility traces to observable states, and did not store past decisions. The eligibility trace model better accounted for participants' choices, while the model with internal memory elements heavily fractionated the state space resulting in slow learning. In a related study, participants learned the initial and final selections in a sequential choice task even though the final selection violated the Markov property (Fu & Anderson, 2008a). Although Fu and Anderson interpreted their results in terms of TD learning, eligibility traces would be needed to acquire the initial response in their task as well.

How do eligibility traces allow TD learning to overcome these credit assignment bottlenecks? Because both actions lead to the same intermediate state, the intermediate outcome does not adjudicate between actions. When the final outcome is delivered, wins produce positive prediction errors and losses produce negative prediction errors. Because the initial action remains eligible for update, the TD model can assign these later, discriminative prediction errors to earlier actions.

We have considered violations of the Markov property in sequential choice tasks. Embodied agents face a more pervasive type of credit bottleneck when perceptual and motor behaviors separate choices from rewards (Anderson, 2007; Walsh & Anderson, 2009). For example, in representative instrumental conditioning studies, a rat chooses between levers, but only receives reward upon entering the magazine (Balleine et al., 1995; Killcross & Coutureau, 2003). In a strict sense, the act of entering the magazine would block assignment of credit to the lever press. Eligibility traces are applicable to these credit assignment bottlenecks as well.

Memory—The second technique for dealing with POMDPs is to remember past states and actions to disambiguate the current state (McCallum, 1995). Tanaka et al.'s (2009) memory model, described in the previous section, did just that. Cognitive architectures that include a short-term memory component also accomplish tasks in this way (Anderson, 2007; Frank & Claus, 2006; O'Reilly & Frank, 2006; Wang & Laird, 2007). For example, in the 12-AX task, the individual views a continuous stream of letters and numbers (O'Reilly & Frank, 2006; Todd, Niv, & Cohen, 2009). The correct response mappings for the current item depend on the identity of earlier items. As such, the individual must maintain and update information about context. According to the gating hypothesis, the prefrontal cortex maintains such contextual information (O'Reilly & Frank, 2006). The same mechanism that supports the acquisition of stimulus-response mappings, the dopamine predication error signal, teaches the basal ganglia when to update the contents of working memory. Memory can also be used to disambiguate states in sequential choice tasks. For example, in the absence of salient cues, a rat must remember prior moves to determine its current location in a maze. Wang and Laird (2007) showed that a cognitive model that used past actions to disambiguate the current state better accounted for the idiosyncratic navigation errors of rats than did a model without memory.

Belief states—The final technique for dealing with POMDPs is to maintain a belief state vector that contains the relative likelihood of each state (Kaelbling, Littman, & Cassandra, 1998). Calculating these likelihoods is non-trivial when the underlying state of the system is not observable; for example, when different events or actions lead probabilistically to different states with identical appearances. Upon acting and receiving a new observation, the individual updates the belief state vector. Updated beliefs depend on states' prior probabilities, and states' conditional probabilities given the new observation. When paired with model-based techniques, belief states allow the individual to calculate the values of actions, weighted according to the likelihood that the individual is in each state.

Belief states have been used to capture sequential sampling results (Dayan & Daw, 2008; Rao, 2010). For example, in the dot-motion detection task, the individual reports the direction in which an array of dots is moving. Some dots move coherently and some move randomly. In the dot-motion detection task, states are the possible directions of motion, observations are the array of coherent and incoherent dots at each moment, and the belief state vector contains the subject's expectations and confidence about the direction of motion. Belief states have also been used to model navigational uncertainty (Stankiewicz et al., 2006; Yoshida & Ishii, 2006). The challenge in these navigation tasks is to move toward an occluded location while gathering information to disambiguate one's current location. Interestingly, in sequential sampling and navigation studies, humans perform worse than ideal observer models (Doshi-Velez & Ghahramani, 2011; Stankiewicz et al., 2006). These shortcomings have been attributed to errors in updating belief states.

Distributed Outcomes

The problem of temporal credit assignment arises when feedback follows a sequence of decisions. A related problem arises when the consequences of a decision are distributed over a sequence of outcomes. This is the case in the Harvard Game, a task where participants select between an action that increases immediate reward and an action that increases future reward (Herrnstein et al., 1993). Like Aesop's fabled grasshopper, humans and animals struggle to forgo immediate gratification to secure future rewards when confronted with such a scenario.

In one Harvard Game experiment (Tunney & Shanks, 2002), participants chose between two actions, *left* and *right*. The immediate (or local) reward associated with *left* exceeded the reward associated with *right* by a fixed amount (Figure 6), but the future (or global) reward associated with both actions grew in proportion to the percentage of responses allocated to *right* during the previous ten trials. Consequently, selection of *left* (melioration) increased immediate reward, but selection of *right* (maximization) increased total reward. In this and other Harvard Game experiments, responses change the probability or magnitude of reward without altering the observable state of the system. Consequently, the basic TD model can learn only to meliorate (Montague & Berns, 2002), as do participants in most Harvard Game experiments.

Yet participants do not always meliorate (Tunney & Shanks, 2002). Four factors increase the percentage of maximizing responses. First, displaying cues that mark the underlying state of the system increases maximization (Gureckis & Love, 2009; Herrnstein et al., 1993).

Second, shortening the inter-trial interval increases maximization (Bogacz et al., 2007). Third, maximizing is negatively associated with the length of the averaging window (i.e., the number of trials over which response allocation is computed; Herrnstein et al., 1993; Yarkoni et al., 2005). Fourth, maximizing is negatively associated with the immediate difference between the payoff functions for the two responses (Heyman & Dunn, 2002; Warry, Remington, & Sonuga-Barke, 1999).

The Harvard Game constitutes a POMDP; the probability or magnitude of reward depends on the current choice *and* the history of actions. One could restore the Markov property by providing participants with information about their action history. Indeed, displaying cues that mark the underlying state of the system increases maximization (Gureckis & Love, 2009; Herrnstein et al., 1993). By restoring the Markov property, cues may allow participants to credit actions for future reward based on observable states. Do POMDP techniques permit maximization when state cues are absent? Yes. A model with memory elements can record the history of actions. Such an internal record can disambiguate states in the same manner as external cues. Likewise, a model with belief states can learn the rewards associated with each of the system's underlying states. The model can then simulate the long-term utility associated with performing sequences of minimizing and maximizing responses to decide which yields greater reward.

Eligibility traces provide an especially elegant account of various Harvard Game manipulations. Bogacz et al. (2007) modeled the effect of inter-trial interval duration using real-time decaying eligibility traces. Over long inter-trial intervals, traces decayed to zero, reducing their ability to support learning. Like participants, the model exhibited maximization when inter-trial intervals were short, and melioration when inter-trial intervals were long. Additionally, as the length of the averaging window increases, and as the difference between the payoff functions increases, the individual must integrate outcomes over a larger number of trials before the value of maximizing exceeds the value of meliorating. Melioration may ensue when the decay of eligibility traces fails to permit integration of outcomes over a sufficient number of trials. In sum, a TD model with eligibility traces can account for the results of several Harvard Game manipulations.

Hierarchical Reinforcement Learning

As the field of reinforcement learning has matured, focus has shifted to factors that limit its applicability. Foremost among these is the scaling problem: the performance of standard reinforcement learning techniques declines as the number of states and actions increase. Hierarchical reinforcement learning (HRL) is one solution to the scaling problem (Barto & Mahadevan, 2003; Botvinick, Niv, & Barto, 2009; Dietterich, 2000). The HRL framework is expanded to include temporally abstract options, representations comprised of primitive, interrelated actions. For instance, the actions involved in adding sugar to coffee (grasp spoon, scoop sugar, lift spoon to cup, deposit sugar) are represented by a single high-level option ('add sugar'). Still more complex skills ('make coffee') are assembled from high-level options (Botvinick, Niv, & Barto, 2009).

In HRL, each option has a designated subgoal. Pseudo-reward is issued upon subgoal completion, and is used to reinforce the actions selected as the individual enacted an option.

External reward, in turn, is issued upon task completion and is used to reinforce the options selected as the individual performed the task. This segmentation of the learning episode enhances scalability in two ways. First, because pseudo-rewards are issued following subgoal completion, the individual need not wait until the end of the task to receive feedback. Thus, reward is not discounted as substantially, and the actions that comprise an option are insulated against errors that occur as the individual pursues later subgoals. Second, HRL allows the individual to learn more efficiently from experience. When options can be applied to new tasks (e.g., adding sugar to coffee, and then adding sugar to tea), the individual can recombine options rather than relearn the larger number of primitive actions they entail.

One psychological prediction of HRL is that credit assignment will occur when participants make progress toward subgoals, even when subgoals do not directly relate to primary reinforcement. To test this hypothesis, Ribas-Fernandes et al. (2011) conducted an experiment where participants navigated a cursor to an intermediate target and then to a final target. The location of the intermediate target sometimes changed unexpectedly. Location shifts that decreased the distance to the intermediate target without decreasing the distance to the final target triggered activation in the anterior cingulate cortex, a region implicated in signaling primary rewards. The shift would not be expected to produce a standard reward prediction error because it did not reduce the total distance to the final goal. The shift would be expected to produce a pseudo-reward prediction error, however, because it did reduce the distance to the intermediate subgoal. Thus, the ACC may also signal pseudo-rewards, as in the HRL framework.

Disorders and Addiction

Reinforcement learning has been applied to the study of psychological disorders (Maia & Frank, 2011; Redish, Jensen, & Johnson, 2008; Redgrave, 2010). For example, the initial stage of Parkinson's disease is characterized by the loss of dopaminergic inputs from the SNc to the striatum (Maia & Frank, 2011; Redish, Jensen, & Johnson, 2008). Given that these regions support model-free control, it is not surprising that Parkinson patients display an impaired ability to acquire and express habitual responses. Reinforcement learning has also been applied to the study of addiction. Redish et al. (2008) describe ten vulnerabilities in the decision-making system. These constitute points through which drugs of addiction can produce maladaptive responses. Some vulnerabilities relate to the goal-direct system; for example, drugs that alter memory storage and access could affect model-based, forward search. Other vulnerabilities relate to the habit system; for example, drugs that artificially increase dopamine could inflate model-free, action value estimates.

The reinforcement learning framework offers insight into deficits in sequential choice as well. For instance, 'impulsivity' is defined as choosing a smaller immediate reward over a larger delayed reward. This is the case in addiction: the individual selects behaviors that are immediately rewarding but ultimately harmful. Addicts discount future reward more steeply than do non-addicts. This was demonstrated in an experiment where participants chose between a large delayed reward and smaller immediate rewards (Madden et al., 1997).

Opioid-dependent participants accepted smaller immediate rewards than did controls, a result that holds across other types of addiction (Reynolds, 2006).

Impulsivity has been associated with changes in the availability of the neurotransmitter serotonin (Mobini et al., 2000). Diminished serotonin levels caused by lesions and pharmacological manipulations increase impulsivity. The impact of serotonin on choice behavior has been simulated in the reinforcement learning framework by decreasing the eligibility trace decay term or by lowering the temporal discounting rate (Doya, 2000; Schweighofer et al., 2008; Tanaka et al., 2009). Decreasing the eligibility trace decay term (λ) reduces the efficiency of propagating credit from delayed outcomes to earlier states and actions. Lowering the discounting rate (γ), in turn, minimizes the contribution of future outcomes to the calculation of utility values. In both cases, the net result is impulsive behavior.

Discussion

Psychologists have long distinguished between habitual and goal-directed behavior. This distinction was made rigorous in stimulus-response and cognitive theories of learning during the behaviorist era. Although this distinction has carried forward in studies of human and animal choice, the apparent dichotomy between these perspectives has given way to a unified view of habitual and goal-directed control as complementary mechanisms for action selection. Similar ideas have emerged in the field of artificial intelligence. Model-free RL resonates with stimulus-response theories and the notion of habitual control, whereas model-based RL resonates with cognitive theories and the notion of goal-directed control.

The pursuit of such compatible ideas in the fields of psychology and artificial intelligence is not coincidental. Humans and animals regularly face the very problems that reinforcement learning methods are designed to overcome. One such problem is temporal credit assignment. When feedback follows a sequence of decisions, how should credit be assigned to the intermediate actions that comprise the sequence? Model-free RL solves this problem by learning internal value functions that store the sum of immediate and future rewards expected from each state and action. Model-based RL solves this problem by learning state transition and reward functions, and by using this internal model to identify actions that will result in goal attainment. Not only do these techniques have a natural interpretation in terms of human and animal behavior, but they also provide a useful framework for understanding neural reward valuation and action selection. Reciprocally, the manner in which humans and animals cope with temporal credit assignment provides a model for designing artificial learning systems that can solve the very same problem (Sutton & Barto, 1998).

Throughout this review, we have emphasized the utility of RL as a model of reward valuation and action selection. We now explore three remaining questions that are central to a unified theory of habitual and goal-directed control. The answers to these questions have ramifications for theories of sequential choice, and decision making more generally.

What Factors Promote Model-Based and Model-Free Control?

Under what circumstances is behavior model-based, and under what circumstances is it model-free? Animal conditioning studies are informative with respect to this question. Animals are sensitive to outcome devaluation and contingency degradation early in conditioning but not after extended training (Adams et al., 1981; Dickinson et al., 1998), a result that has been replicated with humans (Tricomi, Balleine, & O'Doherty, 2009). This suggests that with overtraining, goal-directed (i.e., model-based) control gives way to habitual (i.e., model-free) control. Extending this result to sequential choice, Gläscher et al. (2010) found that participants' behavior was initially predicted by a model-based controller, and later by a model-free controller.

Secondary tasks that consume attention and working memory also shift the balance from model-based to model-free control. In two experiments, Fu and Anderson asked participants to make a pair of decisions before receiving feedback (2008a, 2008b). Participants in the dual-task condition performed a memory-intensive *n*-back task as they made selections. Although participants ultimately learned both choices in the single- and dual-task conditions, the order in which they learned the choices differed. Participants in the dual-task condition learned the second choice before the first. This is consistent with a TD model in which reward propagates from later states to earlier states. Conversely, participants in the single-task condition learned the first choice before the second. This is consistent with a model in which each of the choices and the outcome are encoded in memory, but the first choice is encoded more strongly owing to a primacy advantage (Drewnowski & Murdock, 1980). Likewise, Otto et al. (*in press*) found that having participants perform a demanding secondary task engendered reliance upon a model-free control strategy in the primary, sequential choice task. In the absence of the secondary task, participants reverted to a model-based control strategy.

Finally, time constraints evoke model-free control. In one study that examined this issue, participants learned the transition function for a grid navigation task (Fermin et al., 2010). At test, participants navigated to a novel location or to a well-rehearsed location. Fermin et al. manipulated start time by presenting the go signal immediately after the goal location appeared, or after a six-second delay. Prestart delay facilitated performance when participants navigated to a novel location but not when they navigated to a well-rehearsed location. This suggests that participants planned the sequence of moves to the novel location in the six-second start time condition, and that they used an automatized sequence of moves to navigate to well-rehearsed locations in both start time conditions. The finding that prestart delay facilitated performance when participants navigated to novel locations highlights the time costs associated with forward search. This echoes other findings that planning times increase with search depth and complexity (Hayes, 1965; Owen et al., 1995; Simon & Daw, 2011).

How Does the Cognitive System Arbitrate Between Model-Based and Model-Free Control?

According to several proposals, arbitration between model-free and model-based control is guided by two conflicting constraints (Daw, Niv, & Dayan, 2005; Fermin et al., 2010; Keramati, Dezfouli, & Piray, 2011). First, the computational simplicity of model-free RL

permits rapid response selection whereas model-based RL requires a time consuming search. Second, model-free RL is inflexible in adapting to changing conditions whereas model-based RL can quickly adapt with experience. Consequently, model-based RL is initially favored for its greater accuracy, and model-free RL is ultimately favored for its greater efficiency. When approximation techniques such as pruning are employed, model-based estimates remain uncertain despite extended training, further favoring the transition to model-free control (Daw, Niv, & Dayan, 2005).

The tradeoff between speed and accuracy emerges naturally in the integrated cognitive architecture, Adaptive Control of Thought – Rational (ACT-R; Anderson, 2007). ACT-R contains two types of knowledge; production rules that specify how to act when a set of conditions is met, and declarative knowledge that consists of propositional facts stored in long term memory. Procedural learning, which is driven by external reward, is analogous to model-free RL, whereas the coordinated retrieval of information from declarative memory fulfills a function akin to model-based RL. Because retrievals are instantiated within production rules, this path to action selection is also sensitive to external reward.

Through the process of production compilation, actions based on declarative knowledge are compiled into more specific rules that map stimuli directly to responses, bypassing time-consuming retrievals (Taatgen et al., 2008). Ultimately, procedural and declarative knowledge favor the same responses, but decisions based on procedural knowledge can be enacted more rapidly. Consequently, ACT-R comes to favor specialized productions over flexible, but slow, retrievals. One advantage to implementing choice models within this production system framework is that it permits arbitration among controllers without evoking a homunculus. The same machinery that allows ACT-R to choose between productions, reinforcement signals arising from dopamine neurons, allows ACT-R to choose between model-free and model-based control.

What Alternate Frameworks Exist for Studying Reward Learning?

Although RL explains a wide range of results in reward learning, other proposals have been advanced. For example, a selectionist approach to reinforcement has been influential in the field of behavioral analysis (Donahoe, Burgos, & Palmer, 1993; Thorndike, 1905). By this view, adaptive behavior arises as a byproduct of the same forces that shape Darwinian evolution: *variability*, *selection*, and *retention*. Variability describes the class of potential behaviors, selection describes the potentiating effects of reinforcement on behavior, and retention describes the physiological changes that permit maintenance of adaptive responses. McDowell (2004) developed a model of instrumental conditioning based on such evolutionary principles. He represented actions as ‘populations’ of behavior. Using an evolutionary reinforcement learning algorithm, McDowell demonstrated that actions similar to those rewarded became more prevalent in the population.

Selectionism provides an interesting counterpoint to model-free RL. Models such as McDowell’s (2004) account for operant conditioning phenomena neglected in work on reinforcement learning; for example, response rates under different schedules of reinforcement. Yet there are striking similarities between selectionism and model-free RL (Donahoe, Burgos, & Palmer, 1993). In both, learning only occurs contemporaneously with

discrepancies, classical and operant responses are acquired in a unified manner, and rewards influence behavior by strengthening associations between sensory inputs and motor outputs. Further, in their seminal physiological account of selectionism, Donahoe et al. (1993) attributed the potentiating effects of reinforcement to the release of dopamine. Even the differences between selectionism and model-free RL may be more apparent than real. For example, although most RL models treat responses discretely rather than as a population of behaviors, techniques exist for applying RL to continuous, graded response categories (Sutton & Barto, 1998). More work is needed to determine whether and how the predictions of selectionism differ from those of model-free RL.

Bayesian techniques have also been applied to reward learning. Theories of this form infer a model of the latent structure of the environment, and use this model to predict reward probability (Courville, Daw, & Touretzky, 2006). In the latent cause model, an exemplary Bayesian theory, stimulus and outcome are jointly attributed to a hidden variable (Courville, Daw, & Touretzky, 2006; Gershman & Niv, 2012). Upon viewing a stimulus, the individual infers the latent cause in order to predict the likely outcome.

The latent cause model accounts for the acquisition and extinction of conditioned responses in a manner distinct from associative theories. The individual attributes the conditioned stimulus and outcome to one latent cause during acquisition, and the individual attributes the conditioned stimulus and the *absence* of the outcome to a second latent cause during extinction (Gershman, Blei, & Niv, 2010). Renewal, the finding that the conditioned response is recovered rapidly when the individual is returned to the training context, is an emergent property of the latent cause model: the context change supports the inference that the initial latent cause is again active.

The latent cause model accounts for other conditioning phenomena that challenge associative theories. For example, the partial-reinforcement extinction effect refers to the fact that extinction is slower following training in which the conditioned stimulus is partially reinforced (Capaldi, 1957). Associative theories incorrectly predict that extinction will occur rapidly in such cases because the strength of association between the conditioned stimulus and outcome is weak to begin with. The latent cause model, in contrast, correctly predicts that extinction will occur gradually. When the change in reinforcement rate is large, the model is more likely to assign a new latent cause to the extinction context, but when the change is small, the model is less likely to assign a new latent cause (Gershman, Blei, & Niv, 2010). Latent inhibition, the finding that conditioning is slower when the animal is first preexposed to the stimulus in the absence of the outcome (Lubow, 1989), also challenges associative models.⁷ The latent cause theory explains this effect in terms of the animal's inference that the same cause is active during the preexposure and conditioning phases. Because no outcome followed the stimulus during preexposure, the animal does not expect for an outcome to follow the stimulus during conditioning.

⁷Associative theories can account for latent inhibition, however, by assuming that associability, a parameter akin to learning rate, is dynamic. For example, Maia (2009) demonstrates how a Kalman filter can be used to adjust stimulus associability in a statistically-optimal manner.

The latent cause model accounts for an impressive scope of conditioning phenomena, many of which fall beyond associative theories like TD learning. As with all theories, however, the latent cause model does not account for all conditioning results (Gershman & Niv, 2012). In addition to exploring the many yet unexplained results, it would be interesting to generalize the latent cause model to the case of instrumental conditioning. Belief states constitute a Bayesian solution to the problem of partial observability (Kaelbling, Littman, & Cassandra, 1998). Likewise, the latent cause model provides a way to infer the underlying world state, which can serve as a basis for action selection. It would also be interesting to examine how and when the organism infers the existence of a new latent cause. Redish et al. (2007) propose a solution in the form of a ‘state splitting’ mechanism that is activated by tonically low dopaminergic signals, which indicate that the environment has changed. Bayesian techniques for structural discovery, which may be viewed as reflecting a more abstract level of analysis than RL, exist as well (Courville, Daw, & Touretzky, 2006; Kemp & Tenenbaum, 2008; Sims et al., 2012).

Conclusion—In this review, we have examined connections between stimulus-response and cognitive theories, habitual and goal-directed control, and model-free and model-based RL. A chief strength of reinforcement learning techniques is their ability to overcome the problem of temporal credit assignment. Behavioral studies of sequential choice suggest that humans and animals solve this problem in a similar way. Moreover, neuroscientific investigations have begun to reveal how model-free and model-based RL are instantiated neurally.

Acknowledgments

This work was supported by National Institute of Mental Health training grant T32MH019983 to Matthew Walsh and National Institute of Mental Health grant MH068243 to John Anderson.

References

- Abler B, Walter H, Erk S, Kammerer H, Spitzer M. Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage*. 2006; 31:790–795. [PubMed: 16487726]
- Adams CD, Dickinson A. Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*. 1981; 33B:109–121.
- Anderson, JR. How can the human mind occur in the physical universe?. New York, NY: Oxford University Press; 2007.
- Anderson JR, Albert MV, Fincham JM. Tracing problem solving in real time: fMRI analysis of the subject-paced tower of Hanoi. *Journal of Cognitive Neuroscience*. 2005; 17:1261–1274. [PubMed: 16197682]
- Asaad WF, Rainer G, Miller EK. Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*. 2000; 84:451–459.10.1016/j.neuroimage.2010.12.036 [PubMed: 10899218]
- Baker TE, Holroyd CB. Which way do I go? Neural activation in response to feedback and spatial processing in a virtual T-Maze. *Cerebral Cortex*. 2009; 19:1708–1722.10.1093/cercor/bhn22 [PubMed: 19073622]
- Balleine BW. Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*. 1992; 18:236–250. [PubMed: 1619392]
- Balleine BW. Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology and Behavior*. 2005; 86:717–730. [PubMed: 16257019]

- Balleine BW, Garner C, Gonzalez F, Dickinson A. Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*. 1995; 21:203–217.
- Balleine BW, O'Doherty JP. Human and rodent homologues in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*. 2010; 35:48–69.10.1038/npp.2009.131 [PubMed: 19776734]
- Barto AG, Mahadevan S. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*. 2003; 13:41–77.10.1016/j.pneurobio.2010.06.007
- Berns GS, McClure SM, Pagnoni G, Montague PR. Predictability modulates human brain response to reward. *Journal of Neuroscience*. 2001; 21:2793–2798. [PubMed: 11306631]
- Berridge KC. The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology*. 2007; 191:391–431. [PubMed: 17072591]
- Blodgett HC. The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*. 1929; 4:113–134.
- Bogacz R, McClure SM, Li J, Cohen JD, Montague PR. Short-term memory traces for action bias in human reinforcement learning. *Brain Research*. 2007; 1153:111–121. [PubMed: 17459346]
- Botvinick M. Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, and Behavioral Neuroscience*. 2007; 7:356–366.
- Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. Conflict monitoring and cognitive control. *Psychological Review*. 2001; 108:624–652. [PubMed: 11488380]
- Botvinick MM, Niv Y, Barto AC. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*. 2009; 113:262–280.10.1016/j.cognition.2008.08.011 [PubMed: 18926527]
- Bray S, Shimojo S, O'Doherty JP. Human medial orbitofrontal cortex is recruited during experience of imagined and real rewards. *Journal of Neurophysiology*. 2010; 103:2506–2512.10.1152/jn.01030.2009 [PubMed: 20200121]
- Brogden WJ. Sensory pre-conditioning. *Journal of Experimental Psychology*. 1939; 25:223–232.
- Brown JW, Braver TS. Learned predictions of error likelihood in the anterior cingulate cortex. *Science*. 2005; 18:1118–1121. [PubMed: 15718473]
- Buel J. Differential errors in animal mazes. *Psychological Bulletin*. 1935; 32:67–99.
- Bunge SA. How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective, and Behavioral Neuroscience*. 2004; 4:564–579.
- Bunsey M, Eichenbaum H. Conservation of hippocampal memory function in rats and humans. *Nature*. 1996; 379:255–257. [PubMed: 8538790]
- Busmeyer JR, Pleskac TJ. Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology*. 2009; 53:126–138.
- Bush, RR.; Mosteller, F. *Stochastic models for learning*. Oxford, England: John Wiley & Sons, Inc; 1955.
- Bussey TJ, Wise SP, Murray EA. The role of ventral and orbital prefrontal cortex in conditional visuomotor learning and strategy use in Rhesus monkeys (*Macaca mulatta*). *Behavioral Neuroscience*. 2001; 115:971–982. [PubMed: 11584930]
- Capaldi EJ. The effect of different amounts of alternating partial reinforcement on resistance to extinction. *The American Journal of Psychology*. 1957; 70:451–452. [PubMed: 13458520]
- Cardinal RN, Parkinson JA, Hall J, Everitt BJ. Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*. 2002; 26:321–352. [PubMed: 12034134]
- Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*. 1998; 280:747–749. [PubMed: 9563953]
- Cohen MX, Ranganath C. Reinforcement learning signals predict future decisions. *Journal of Neuroscience*. 2007; 27:371–378. [PubMed: 17215398]
- Cohen, NJ.; Eichenbaum, H. *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press; 1993.

- Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*. 2006; 10:294–300. [PubMed: 16793323]
- Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron*. 2011; 69:1204–1215.10.1016/j.neuron.2011.02.027 [PubMed: 21435563]
- Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*. 2005; 8:1704–1711.
- Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. *Nature*. 2006; 441:876–879. [PubMed: 16778890]
- Dayan P. Instrumental vigour in punishment and reward. *European Journal of Neuroscience*. 2012; 35:1152–1168.10.1111/j.1460-9568.2012.08026.x [PubMed: 22487044]
- Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*. 2008; 8:429–453.10.3758/CABN.8.4.429
- Dayan P, Niv Y. Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*. 2008; 18:185–196.10.1016/j.conb.2008.08.003 [PubMed: 18708140]
- De Groot, AD. Thought and choice in chess. 2. The Hague: Mouton; 1978. (Original work published 1946)
- Delgado MR, Locke HM, Stenger VA, Fiez JA. Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, and Behavioral Neuroscience*. 2003; 3:27–38.
- Dickinson A, Squire S, Varga Z, Smith JW. Omission learning after instrumental pretraining. *Quarterly Journal of Experimental Psychology*. 1998; 51B:271–286.
- Dietterich TG. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*. 2000; 13:227–303.
- Donahoe JW, Burgos JE, Palmer DC. A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior*. 1993; 60:17–40. [PubMed: 8354965]
- Doshi-Velez, F.; Ghahramani, Z. A comparison of human and agent reinforcement learning. In: Carlson, L.; Hoelscher, C.; Shipley, TF., editors. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Sciences Society; 2011. p. 2703-2708.
- Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*. 1999; 12:961–974. [PubMed: 12662639]
- Doya, K. Metalearning, neuromodulation, and emotion. In: Hatano, G.; Okada, N.; Tanabe, H., editors. *Affective minds*. Amsterdam: Elsevier Science; 2000. p. 101-104.
- Drewnowski A, Murdock BB. The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*. 1980; 6:319–332. [PubMed: 7373250]
- Dunning JP, Hajcak G. Error-related negativities elicited by monetary loss and cues that predict loss. *Neuroreport*. 2007; 18:1875–1878. [PubMed: 18090330]
- Elliott R, Newman JL, Longe OA, Deakin JFW. Instrumental responding for rewards is associated with enhanced neuronal response in subcortical reward systems. *Neuroimage*. 2004; 21:984–990. [PubMed: 15006665]
- Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*. 1991; 78:447–455. [PubMed: 1712280]
- Fermin A, Yoshida T, Ito M, Yoshimoto J, Doya K. Evidence for model-based action planning in a sequential finger movement task. *Journal of Motor Behavior*. 2010; 42:371–379.10.1080/00222895.2010.526467 [PubMed: 21184355]
- Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*. 2003; 299:1898–1902. [PubMed: 12649484]
- Foster DJ, Wilson MA. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*. 2006; 440:680–683. [PubMed: 16474382]
- Frank MJ, Claus ED. Anatomy of a decision: Striato-Orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*. 2006; 113:300–326. [PubMed: 16637763]

- Frederick S, Loewenstein G, O'Donoghue T. Time discounting and time preference: A critical review. *Journal of Economic Literature*. 2002; 40:351–401.
- Fu WT, Anderson JR. From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*. 2006; 135:184–206. [PubMed: 16719650]
- Fu WT, Anderson JR. Solving the credit assignment problem: Explicit and implicit learning of action sequences with probabilistic outcomes. *Psychological Research*. 2008a; 72:321–330. [PubMed: 17447083]
- Fu WT, Anderson JR. Dual learning processes in interactive skill acquisition. *Journal of Experimental Psychology: Applied*. 2008b; 14:179–191.10.1037/1076-898X.14.2.179 [PubMed: 18590373]
- Fu WT, Gray WD. Resolving the paradox of the active user: stable suboptimal performance in interactive tasks. *Cognitive Science*. 2004; 28:901–935.
- Fuster, JM. The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe. Philadelphia, PA: Lippincott-Raven; 1997.
- Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E. A neural system for error detection and compensation. *Psychological Science*. 1993; 4:385–390.
- Gershman SJ, Blei D, Niv Y. Context, learning, and extinction. *Psychological Review*. 2010; 117:197–209.10.1037/a0017808 [PubMed: 20063968]
- Gershman SJ, Markman AB, Otto AR. Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*. (in press).
- Gershman SJ, Niv Y. Exploring a latent cause theory of classical conditioning. *Learning and Behavior*. 2012; 40:255–268.10.3758/s13420-012-0080-8 [PubMed: 22927000]
- Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*. 2010; 66:585–595. [PubMed: 20510862]
- Gläscher J, Hampton AN, O'Doherty JP. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*. 2009; 19:483–495.10.1016/j.neuron.2010.04.016 [PubMed: 18550593]
- Gray WD, Sims CR, Fu WT, Schoelles MJ. The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*. 2006; 113:461–482. [PubMed: 16802878]
- Gureckis TM, Love BC. Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*. 2009; 113:293–313.10.1016/j.cognition.2009.03.013 [PubMed: 19427635]
- Haber SN, Fudge JL, McFarland NR. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*. 2000; 20:2369–2382. [PubMed: 10704511]
- Haber SN, Lynd-Balta E, Klein C, Groenewegen HJ. Topographic organization of the ventral striatal efferent projections in the rhesus monkey: An anterograde tracing study. *Journal of Comparative Neurology*. 1990; 293:282–298. [PubMed: 19189717]
- Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*. 2006; 26:8360–8367. [PubMed: 16899731]
- Hayes JR. Problem topology and the solution process. *Journal of Verbal Learning and Verbal Behavior*. 1965; 4:371–379.
- Herrnstein RJ, Loewenstein GF, Prelec D, Vaughan W. Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making*. 1993; 6:149–185.
- Herrnstein RJ, Prelec D. Melioration: A theory of distributed choice. *Journal of Economic Perspectives*. 1991; 5:137–156.
- Heyman GM, Dunn B. Decision biases and persistent illicit drug use: An experimental study of distributed choice and addiction. *Drug and Alcohol Dependence*. 2002; 67:193–203. [PubMed: 12095669]
- Holland PC, Rescorla R. The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*. 1975; 1:355–363. [PubMed: 1202141]

- Holroyd CB, Coles MGH. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*. 2002; 109:679–709. [PubMed: 12374324]
- Holroyd CB, Krigolson OE, Lee S. Reward positivity elicited by predictive cues. *Neuroreport*. 2011; 22:249–252.10.1097/WNR.0b013e328345441d [PubMed: 21386699]
- Hoshi E, Shima K, Tanji J. Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*. 2000; 83:2355–2373. [PubMed: 10758139]
- Hull CL. The goal gradient hypothesis and maze learning. *Psychological Review*. 1932; 39:25–43.
- Hull, CL. *Principles of behavior: An introduction to behavioral theory*. New York, NY: Appleton-Century-Crofts; 1943.
- Huys QJM, Eshel N, O’Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*. 2012; 8:1–13.10.1371/journal.pcbi.1002410 [PubMed: 22629235]
- Hyman SE, Malenka RC. Addiction and the brain: The neurobiology of compulsion and its persistence. *Nature Reviews Neuroscience*. 2001; 2:695–703.
- Imamizu H, Kawato M. Brain mechanisms for predictive control by switching internal models: Implications for higher-order cognitive functions. *Psychological Research*. 2009; 73:527–544.10.1007/s00426-009-0235-1 [PubMed: 19347360]
- Ito M. Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*. 2008; 9:304–313.10.1038/nrn2332
- Izquierdo AD, Suda RK, Murray EA. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience*. 2004; 24:7540–7548. [PubMed: 15329401]
- James, W. *The principles of psychology*. New York, NY: Dover: 1950. (Original work published 1890)
- Janssen CP, Gray WD. When, what, and how much to reward in reinforcement learning-based models of cognition. *Cognitive Science*. 2012; 36:333–358.10.1111/j.1551-6709.2011.01222.x [PubMed: 22257174]
- Joel D, Niv Y, Ruppert E. Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*. 2002; 15:535–547. [PubMed: 12371510]
- Johnson A, Redish AD. Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*. 2005; 18:1163–1171. [PubMed: 16198539]
- Johnson A, Redish AD. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*. 2007; 27:12176–12189. [PubMed: 17989284]
- Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*. 1998; 101:99–134.
- Kemp C, Tenenbaum JB. The discovery of structural form. *Proceedings of the National Academy of Sciences*. 2008; 105:10687–10692.10.1073/pnas.0802631105
- Kennerley SW, Walton ME, Behrens TEJ, Buckley MJ, Rushworth MFS. Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*. 2006; 9:940–947.
- Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology*. 2011; 7:1–21.10.1371/journal.pcbi.1002055
- Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. Anterior cingulate conflict monitoring and adjustments in control. *Science*. 2004; 13:1023–1026. [PubMed: 14963333]
- Killcross S, Coutureau E. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*. 2003; 13:400–408. [PubMed: 12631569]
- Knutson B, Taylor J, Kaufman M, Peterson R, Glover G. Distributed neural representation of expected value. *Journal of Neuroscience*. 2005; 25:4806–4812. [PubMed: 15888656]
- Kobayashi S, Schultz W. Influences of reward delays on responses of dopamine neurons. *Journal of Neuroscience*. 2008; 28:7837–7846.10.1523/JNEUROSCI.1600-08.2008 [PubMed: 18667616]

- Liao Y, Gramann K, Feng W, Deák GO, Li H. This ought to be good: Brain activity accompanying positive and negative expectations and outcomes. *Psychophysiology*. 2011; 48:1412–1419.10.1111/j.1469-8986.2011.01205.x [PubMed: 21517899]
- Liljeholm M, Tricomi E, O'Doherty JP, Balleine BW. Neural correlates of instrumental contingency learning: Differential effects of action-reward conjunction and disjunction. *Journal of Neuroscience*. 2011; 31:2474–2480.10.1523/JNEUROSCI.3354-10.2011 [PubMed: 21325514]
- Lin LJ. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*. 1992; 8:293–321.
- Loch, J.; Singh, S. *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann; 1998. Using eligibility traces to find the best memoryless policy in partially observable Markov decision processes; p. 323-331.
- Lubow, R. *Latent inhibition and conditioned attention theory*. Cambridge, U.K.: Cambridge University Press; 1989.
- Luce RD. The choice axiom after twenty years. *Journal of Mathematical Psychology*. 1977; 15:215–233.
- Madden GJ, Petry NM, Badger GJ, Bickel WK. Impulsive and self-control choices in opioid-dependent patients and non-drug-using control participants: Drug and monetary rewards. *Experimental and Clinical Psychopharmacology*. 1997; 5:256–262. [PubMed: 9260073]
- Maia TV. Fear conditioning and social groups: Statistics, not genetics. *Cognitive Science*. 2009; 33:1232–1251. [PubMed: 21585503]
- Maia TV. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*. 2010; 38:50–67.10.3758/LB.38.1.50 [PubMed: 20065349]
- Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*. 2011; 14:154–162.10.1038/nn.2723
- Mansouri FA, Matsumoto K, Tanaka K. Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting test analog. *Journal of Neuroscience*. 2006; 26:2745–2756. [PubMed: 16525054]
- McCallum, RA. *The Proceedings of the Twelfth International Machine Learning Conference*. San Francisco, CA: Morgan Kaufmann; 1995. Instance-based utile distinctions for reinforcement learning with hidden states; p. 387-395.
- McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. *Neuron*. 2003; 38:339–346. [PubMed: 12718866]
- McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*. 2011; 31:2700–2705.10.1523/JNEUROSCI.5499-10.2011 [PubMed: 21325538]
- McDowell JJ. A computational model of selection by consequences. *Journal of the Experimental Analysis of Behavior*. 2004; 81:297–317. [PubMed: 15357512]
- Miltner WHR, Braun CH, Coles MGH. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a “generic” neural system for error detection. *Journal of Cognitive Neuroscience*. 1997; 9:788–798. [PubMed: 23964600]
- Minsky, M. Steps toward artificial intelligence. In: Feigenbaum, EA.; Feldman, J., editors. *Computers and thought*. New York, NY: McGraw-Hill; 1963. p. 406-450.
- Mobini S, Chiang TJ, Ho MY, Bradshaw CM, Szabadi E. Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology*. 2000; 152:390–397. [PubMed: 11140331]
- Montague PR, Berns GS. Neural economics and the biological substrates of valuation. *Neuron*. 2002; 36:265–284. [PubMed: 12383781]
- Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*. 1996; 16:1936–1947. [PubMed: 8774460]
- Montague PR, Hyman SE, Cohen JD. Computational roles for dopamine in behavioural control. *Nature*. 2004; 431:760–767. [PubMed: 15483596]
- Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*. 2006; 9:1057–1063.

- Muhammad R, Wallis JD, Miller EK. A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*. 2006; 18:974–989. [PubMed: 16839304]
- Mushiake H, Saito N, Sakamoto K, Itoyama Y, Tanji J. Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*. 2006; 50:631–641. [PubMed: 16701212]
- Newell, A.; Simon, H. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall; 1972.
- Niv Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology*. 2009; 53:139–154.
- O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. *Neuron*. 2003; 38:329–337. [PubMed: 12718865]
- O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*. 2004; 304:452–454. [PubMed: 15087550]
- O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Science*. 2007; 1104:35–53.
- O'Keefe, J.; Nadel, L. The hippocampus as a cognitive map. Oxford, England: Clarendon; 1978.
- O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in prefrontal cortex and basal ganglia. *Neural Computation*. 2006; 18:283–328. [PubMed: 16378516]
- Ostlund SB, Balleine BW. Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *Journal of Neuroscience*. 2005; 25:7763–7770. [PubMed: 16120777]
- Otto AR, Gershman SJ, Markman AB, Daw ND. The curse of planning: Dissecting multiple reinforcement learning systems by taxing the central executive. *Psychological Science*. (in press).
- Owen AM. Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*. 1997; 53:431–450. [PubMed: 9421831]
- Owen AM, Sahakian BJ, Hodges JR, Summers BA, Polkey CE, Robbins TW. Dopamine-dependent fronto-striatal planning deficits in early Parkinson's disease. *Neuropsychology*. 1995; 9:126–140.
- Packard MG, Knowlton BJ. Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*. 2002; 25:563–593.
- Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*. 2002; 5:97–98.
- Pan WX, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward learning network. *Journal of Neuroscience*. 2005; 25:6235–6242. [PubMed: 15987953]
- Pavlov, IP. Conditioned reflexes. Anrep, GV., translator. London, UK: Oxford University Press; 1927.
- Plassmann H, O'Doherty J, Rangel A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*. 2007; 27:9984–9988. [PubMed: 17855612]
- Pritchard, WS.; Shappell, SA.; Brandt, ME. Psychophysiology of N200/N400: A review and classification scheme. In: Jennings, JR.; Ackles, PK.; Coles, MGH., editors. *Advances in psychophysiology*. Vol. 4. Jessica Kingsley; London: 1991. p. 43-106.
- Rachlin H. Self-control: Beyond commitment. *Behavioral and Brain Sciences*. 1995; 18:109–159.
- Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*. 2008; 9:545–556.10.1038/nrn2357
- Rao RPN. Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*. 2010; 24:1–18.10.3389/fncom.2010.00146
- Redish AD, Jensen S, Johnson A. A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences*. 2008; 31:415–487.10.1017/S0140525X0800472X [PubMed: 18662461]
- Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*. 2007; 114:784–805. [PubMed: 17638506]

- Redgrave P, Rodriguez M, Smith Y, Rodriguez-Oroz MC, Lehericy S, Bergman H, Agid Y, DeLong MR, Obeso JA. Goal-directed and habitual control in the basal ganglia: Implications for Parkinson's disease. *Nature Reviews Neuroscience*. 2010; 11:760–772.10.1038/nrn2915
- Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. *Classical conditioning II: Current research and theory*. New York, NY: Appleton-Century-Crofts; 1972. p. 64-99.
- Reynolds B. A review of delay-discounting research with humans: Relations to drug use and gambling. *Behavioural Pharmacology*. 2006; 17:651–667. [PubMed: 17110792]
- Reynolds JNJ, Wickens JR. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*. 2002; 15:507–521. [PubMed: 12371508]
- Reynolds JNJ, Hyland BI, Wickens JR. A cellular mechanism of reward-related learning. *Nature*. 2001; 413:67–70. [PubMed: 11544526]
- Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM. A neural signature of hierarchical reinforcement learning. *Neuron*. 2012; 71:370–379.10.1016/j.neuron.2011.05.042 [PubMed: 21791294]
- Rizley RC, Rescorla RA. Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*. 1972; 81:1–11. [PubMed: 4672573]
- Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*. 2007; 10:1615–1624.
- Rolls ET, Kringelbach ML, de Araujo IET. Different representations of pleasant and unpleasant odours in the human brain. *European Journal of Neuroscience*. 2003; 18:695–703. [PubMed: 12911766]
- Rutledge RB, Dean M, Caplin A, Glimcher PW. Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*. 2010; 30:13525–13536.10.1523/JNEUROSCI.1747-10.2010 [PubMed: 20926678]
- Saito N, Mushiaki H, Sakamoto K, Itoyama Y, Tanji J. Representation of immediate and final behavioral goals in the monkey prefrontal cortex during an instructed delay period. *Cerebral Cortex*. 2005; 15:1535–1546. [PubMed: 15703260]
- Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*. 1959; 3:211–229. Reprinted in E. A. Feigenbaum & J. Feldman, (Eds.), *Computers and thought*. New York, NY: McGraw-Hill.
- Satoh T, Nakai S, Sato T, Kimura M. Correlated coding of motivation and outcome of decision by dopamine neurons. *Journal of Neuroscience*. 2003; 23:9913–9923. [PubMed: 14586021]
- Schacter DL, Addis DR, Buckner RL. Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*. 2007; 8:657–661.
- Schultz W. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*. 1998; 80:1–27. [PubMed: 9658025]
- Schultz W, Apicella P, Ljungberg T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*. 1993; 13:900–913. [PubMed: 8441015]
- Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science*. 1997; 275:1593–1599. [PubMed: 9054347]
- Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, Yamawaki S, Doya K. Low-serotonin levels increase delayed reward discounting in humans. *Journal of Neuroscience*. 2008; 28:4528–4532.10.1523/JNEUROSCI.4982-07.2008 [PubMed: 18434531]
- Shallice T. Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. 1982; 298:199–209.
- Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*. 2011; 31:5526–5539.10.1523/JNEUROSCI.4647-10.2011 [PubMed: 21471389]
- Sims CR, Neth H, Jacobs RA, Gray WD. Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review*. 2012; 120:139–154.10.1037/a0030850 [PubMed: 23230889]

- Singh SP, Sutton RS. Reinforcement learning with replacing eligibility traces. *Machine Learning*. 1996; 22:123–158.
- Skinner, BF. *The behavior of organisms: An experimental analysis*. Oxford, England: Appleton-Century; 1938.
- Smith A, Li M, Becker S, Kapur S. Dopamine, prediction error and associative learning: A model-based account. *Network: Computation in Neural Systems*. 2006; 17:61–84.
- Solway A, Botvinick MM. Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*. 2012; 119:120–154.10.1037/a0026435 [PubMed: 22229491]
- Spence KW. The order of eliminating blinds in maze learning by the rat. *Journal of Comparative Psychology*. 1932; 14:9–27.
- Spragg SDS. Anticipatory responses in the maze. *Journal of Comparative Psychology*. 1934; 18:51–73.
- Stankiewicz BJ, Legge GE, Mansfield JS, Schlicht EJ. Lost is virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance*. 2006; 32:688–704. [PubMed: 16822132]
- Stocco A, Lebiere C, Anderson JR. Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychological Review*. 2010; 117:541–574.10.1037/a0019077 [PubMed: 20438237]
- Stoodley CJ. The cerebellum and cognition: Evidence from functional imaging studies. *Cerebellum*. 2012; 11:352–365.10.1007/s12311-011-0260-7 [PubMed: 21373864]
- Strick PL, Dum RP, Fiez JA. Cerebellum and nonmotor function. *Annual Review of Neuroscience*. 2009; 32:413–434.10.1146/annurev.neuro.31.060407.125606
- Sutton, RS. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Porter, BW.; Mooney, RJ., editors. *Proceedings of the Seventh International Conference on Machine Learning*. San Francisco: Morgan Kaufmann; 1990. p. 216-224.
- Sutton, RS.; Barto, AG. Time-derivative models of Pavlovian reinforcement. In: Gabriel, M.; Moore, J., editors. *Learning and computational neuroscience: Foundations of adaptive networks*. Cambridge, MA: MIT Press; 1990. p. 497-537.
- Sutton, RS.; Barto, AG. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press; 1998.
- Taatgen NA, Huss D, Dickison D, Anderson JR. The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*. 2008; 137:548–565.10.1037/0096-3445.137.3.548 [PubMed: 18729715]
- Tanaka SC, Balleine BW, O'Doherty JP. Calculating consequences: Brain systems that encode the causal effects of actions. *Journal of Neuroscience*. 2008; 28:6750–6755.10.1523/JNEUROSCI.1808-08.2008 [PubMed: 18579749]
- Tanaka SC, Shishida K, Schweighofer N, Okamoto Y, Yamawaki S, Doya K. Serotonin affects association of aversive outcomes to past actions. *Journal of Neuroscience*. 2009; 29:15669–15674.10.1523/JNEUROSCI.2799-09.2009 [PubMed: 20016081]
- Thistlethwaite D. A critical review of latent learning and related experiments. *Psychological Bulletin*. 1951; 48:97–129. [PubMed: 14834282]
- Thorndike, EL. *Elements of psychology*. New York, NY: A. G. Seiler; 1905.
- Thrun, SB. The role of exploration in learning control. In: White, DA.; Sofge, DA., editors. *Handbook of intelligent control: Neural, fuzzy and adaptive approaches*. Florence, KY: Van Nostrand Reinhold; 1992.
- Tobler PN, Fiorillo CD, Schultz W. Adaptive coding of reward value by dopamine neurons. *Science*. 2005; 307:1642–1645. [PubMed: 15761155]
- Tobler PN, O'Doherty JP, Dolan RJ, Schultz W. Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*. 2006; 95:301–310. [PubMed: 16192329]
- Todd, MT.; Niv, Y.; Cohen, JD. Learning to use working memory in partially observable environments through dopaminergic reinforcement. In: Koller, D.; Schuurmans, D.; Bengio, Y.;

- Bottou, L., editors. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press; 2009. p. 1689-1696.
- Tolman, EC. *Purposive behavior in animals and men*. New York, NY: Century; 1932.
- Tolman EC, Honzik CH. Degrees of hunger, reward and non-reward, and maze learning in rats. *University of California Publications in Psychology*. 1930; 4:241-256.
- Tricomi E, Balleine BW, O'Doherty JP. A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*. 2009; 29:2225-2232.10.1111/j.1460-9568.2009.06796.x [PubMed: 19490086]
- Tricomi EM, Delgado MR, Fiez JA. Modulation of caudate activity by action contingency. *Neuron*. 2004; 41:281-292. [PubMed: 14741108]
- Tunney RJ, Shanks DR. A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making*. 2002; 15:291-311.
- Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*. 2007; 27:4019-4026. [PubMed: 17428979]
- van der Meer MAA, Redish AD. Covert expectation-of-reward in rat ventral striatum at decision points. *Frontiers in Integrative Neuroscience*. 2009; 3:1-15.10.3389/neuro.07.001.2009 [PubMed: 19225578]
- van Veen V, Carter CS. The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*. 2002; 14:593-602. [PubMed: 12126500]
- Waelti P, Dickinson A, Schultz W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*. 2001; 412:43-48. [PubMed: 11452299]
- Walsh MM, Anderson JR. The strategic nature of changing your mind. *Cognitive Psychology*. 2009; 58:416-440.10.1016/j.cogpsych.2008.09.003 [PubMed: 19013562]
- Walsh MM, Anderson JR. Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*. 2012; 36:1870-1884.10.1016/j.neubiorev.2012.05.008 [PubMed: 22683741]
- Walsh MM, Anderson JR. Learning from delayed feedback: Neural responses in temporal credit assignment. *Cognitive, Affective, and Behavioral Neuroscience*. 2011a; 11:131-143.10.3758/s13415-011-0027-0
- Walsh MM, Anderson JR. Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Science*. 2011b; 108:19048-19053.10.1073/pnas.1117189108
- Wang, Y.; Laird, JE. The importance of action history in decision making and reinforcement learning. *Proceedings of the Eighth International Conference on Cognitive Modeling*; Ann Arbor, MI. 2007.
- Warry CJ, Remington B, Sonuga-Barke EJS. When more means less: Factors affecting human self-control in a local versus global choice paradigm. *Learning and Motivation*. 1999; 30:53-73.
- Wassum KM, Ostlund SB, Maidment NT. Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. *Biological Psychiatry*. 2012; 71:846-854.10.1016/j.biopsych.2011.12.019 [PubMed: 22305286]
- White IM, Wise SP. Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*. 1999; 126:315-335. [PubMed: 10382618]
- Wickens JR, Begg AJ, Arbuthnott GW. Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience*. 1996; 70:1-5. [PubMed: 8848115]
- Wunderlich K, Dayan P, Dolan RJ. Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*. 2012; 15:786-793.10.1038/nn.3068
- Yarkoni T, Braver TS, Gray JR, Green L. Prefrontal brain activity predicts temporally extended decision-making behavior. *Journal of the Experimental Analysis of Behavior*. 2005; 84:537-554. [PubMed: 16596979]
- Yechiam E, Erev I, Yehene V, Gopher D. Melioration and the transition from touch-typing training to everyday use. *Human Factors*. 2003; 45:671-684. [PubMed: 15055463]

- Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*. 2004; 111:931–959. [PubMed: 15482068]
- Yin HH, Knowlton BJ, Balleine BW. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*. 2004; 19:181–189. [PubMed: 14750976]
- Yoshida W, Ishii S. Resolution of uncertainty in prefrontal cortex. *Neuron*. 2006; 50:781–789. [PubMed: 16731515]

Figure 1.

Actor/critic architecture. The actor records preferences for actions in each state. The critic combines information about immediate reward and the expected value of the subsequent state to compute reward prediction errors (δ). The actor uses reward prediction errors to update action preferences, $p(s, a)$, and the critic uses reward prediction errors to update state values, $V(s)$.

Figure 2.

Transition structure in sequential choice task (Daw et al., 2011). The first selection lead to one of two intermediate states with fixed probabilities, and the second selection was rewarded probabilistically.

Figure 3.

Experiment interface (left) and maze structure with correct path in gray (right)(Fu & Anderson, 2006). To exit the maze, participants needed to select the correct cues in Rooms 1, 2, and 3.

Figure 4.

Maze used to assess detour behavior in rats (Tolman & Honzik, 1930). In different trials, detours were placed at points A and B.

Figure 5.

Delayed reward task (Tanaka et al., 2009). Some rewards were delivered immediately (trial $t + 1$), and some rewards were delivered after a delay (trial $t + 3$).

Figure 6. Harvard Game payoff functions (Tunney & Shanks, 2002). Payoff for meliorating (choose *left*) and maximizing (choose *right*) as a function of the percentage of maximizing responses during the previous ten trials.

Table 1

Candidate Structures Implementing Model-Free and Model-Based RL

Component	Structure
Model-free RL	
Prediction error	Substantia nigra pars compacta
	Ventral tegmental area
Actor	Ventral striatum
Critic	Dorsolateral striatum
Model-based RL	
Reward function	Orbitofrontal cortex
Transition function	Hippocampus
	Cerebellum
	Dorsomedial striatum
	Ventromedial prefrontal cortex
Decision policy	Lateral prefrontal cortex