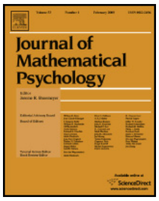
Journal of Mathematical Psychology 53 (2009) 139–154

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp



Reinforcement learning in the brain

Yael Niv[*]

Psychology Department & Princeton Neuroscience Institute, Princeton University, United States

ARTICLE INFO

Article history:
Received 9 June 2008
Received in revised form
3 December 2008
Available online 10 February 2009

ABSTRACT

A wealth of research focuses on the decision-making processes that animals and humans employ when selecting actions in the face of reward and punishment. Initially such work stemmed from psychological investigations of conditioned behavior, and explanations of these in terms of computational models. Increasingly, analysis at the computational level has drawn on ideas from *reinforcement learning*, which provide a *normative framework* within which decision-making can be analyzed. More recently, the fruits of these extensive lines of research have made contact with investigations into the neural basis of decision making. Converging evidence now links reinforcement learning to specific neural substrates, assigning them precise computational roles. Specifically, electrophysiological recordings in behaving animals and functional imaging of human decision-making have revealed in the brain the existence of a key reinforcement learning signal, the temporal difference reward prediction error. Here, we first introduce the formal reinforcement learning framework. We then review the multiple lines of evidence linking reinforcement learning to the function of dopaminergic neurons in the mammalian midbrain and to more recent data from human imaging experiments. We further extend the discussion to aspects of learning not associated with phasic dopamine signals, such as learning of goal-directed responding that may not be dopamine-dependent, and learning about the vigor (or rate) with which actions should be performed that has been linked to tonic aspects of dopaminergic signaling. We end with a brief discussion of some of the limitations of the reinforcement learning framework, highlighting questions for future research.

A fundamental question in behavioral neuroscience concerns the decision-making processes by which animals and humans select actions in the face of reward and punishment, and their neural realization. In behavioral psychology, this question has been investigated in detail through the paradigms of Pavlovian (classical) and instrumental (operant) conditioning, and much evidence has accumulated regarding the associations that control different aspects of learned behavior. The computational field of reinforcement learning (Sutton & Barto, 1998) has provided a normative framework within which such conditioned behavior can be understood. In this, optimal action selection is based on predictions of long-run future consequences, such that decision making is aimed at maximizing rewards and minimizing punishment. Neuroscientific evidence from lesion studies, pharmacological manipulations and electrophysiological recordings in behaving animals have further provided tentative links to neural structures underlying key computational constructs in these models. Most notably, much evidence suggests that the neuromodulator dopamine provides basal ganglia target structures with phasic signals that convey a reward prediction error that can influence learning and action selection, particularly in stimulus-driven habitual instrumental behavior (Barto, 1995; Schultz, Dayan, & Montague, 1997; Wickens & Kötter, 1995).

From a computational perspective, Pavlovian conditioning (Yerkes & Morgulis, 1909) is considered as a prototypical instance of *prediction learning* — learning the predictive relationships between events in the environment such as the fact that the scent of home-cooking usually predicts a tasty meal (e.g. Sutton and Barto (1990)). Instrumental conditioning, on the other hand, involves learning to select actions that will increase the probability of rewarding events and decrease the probability of aversive events (Skinner, 1935; Thorndike, 1911). Computationally, such decision making is treated as attempting to optimize the consequences of actions in terms of some long-term measure of total obtained rewards (and/or avoided punishments) (e.g. Barto (1994)). Thus, the study of instrumental conditioning is an inquiry into perhaps the most fundamental form of rational decision-making. This capacity to select actions that influence the environment to one's subjective benefit is the mark of intelligent organisms, and although animals such as pigeons and rats are capable of modifying their behaviors in response to the contingencies provided by the environment, choosing those behaviors that will maximize rewards and minimize punishments in an uncertain,

[*] Corresponding address: Psychology Department and Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, United States.
E-mail address: yael@princeton.edu.

0022-2496/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2008.12.005

often changing, and computationally complex world is by no means a trivial task.

In recent years computational accounts of these two classes of conditioned behavior have drawn heavily from the framework of *reinforcement learning* (RL) in which models all share in common the use of a scalar reinforcement signal to direct learning. Importantly, RL provides a *normative framework* within which to analyze and interpret animal conditioning. That is, RL models (1) generate predictions regarding the molar and molecular forms of optimal behavior, (2) suggests a means by which optimal prediction and action selection can be achieved, and (3) expose explicitly the computations that must be realized in the service of these. Different from (and complementary to) descriptive models that describe behavior *as it is*, normative models study behavior from the point of view of its hypothesized *function*, that is, they study behavior *as it should be* if it were to accomplish specific goals in an optimal way. The appeal of normative models derives from two primary sources. First, because throughout evolution animal behavior has been shaped and constrained by its influence on fitness, it is not unreasonable to view particular behaviors as optimal or near-optimal adaptations to some set of problems (Kacelnik, 1997). This allows for the generation of computationally explicit and directly testable hypotheses about the characteristics of those behaviors. Second, discrepancies between observed behavior and the predictions of normative models are often illuminating as they can shed light on the neural and/or informational constraints under which animals make decisions, or suggest that animals are, in fact, optimizing something other than what the model has assumed.

Adopting Marr's (1982) famous terminology, normative computational models span both the computational level in which the problem is defined (as they stem from an objective, such as maximizing future reward) and the algorithmic level of its principled solution. The relevance of RL models to human and animal learning and decision-making has recently been strengthened by research linking directly the computational and algorithmic levels to the implementation level. Specifically, extracellular recordings in behaving animals and functional imaging of human decision-making have revealed in the brain the existence of a key RL signal, the *temporal difference reward prediction error*. In this review we will focus on these links between the theory of reinforcement learning and its implementation in animal and human neural processing.

The link to the level of a neural implementation requires a (perhaps not obviously motivated) leap beyond the computer-science realm of RL, into an inquiry of how the brains of animals and humans bring about complex behavior. We believe that this connection between neuroscience and reinforcement learning stands to benefit both lines of research, making (at least) two important contributions. First, although behavioral predictions are extremely useful for the purpose of testing the relevance of RL to animal and human decision-making, neural data provide an important source of support and constraints, grounding the theory in another level of empirical support. This is especially true for a theory that makes clear predictions about learning – a fundamentally *unobservable* process, and its underlying hidden variables (such as prediction errors). Because different learning processes can lead to similar choice behavior, neural evidence is key to selecting one model of learning over another. Prime examples of this are the arbitration between different variants of RL based on dopaminergic firing patterns (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Roesch, Calu, & Schoenbaum, 2007), or the separation versus combination of model-based and model-free approaches to RL based on lesion studies (Daw, Niv, & Dayan, 2005), which we will discussed below. The fact that animals and humans clearly solve the RL problem successfully despite severe constraints on real-time neural computation suggests that the neural mechanisms can also provide a source for new theoretical developments such as approximations due to computational limitations and mechanisms for dealing with continuous and noisy sensory experience. A second contribution that a wedding of the computational and algorithmic levels to the neural implementation level allows, which is of even greater importance, is to our understanding of the neural processes underlying decision-making in the normal and abnormal brain. The potential advantages of understanding learning and action selection at the level of dopamine-dependent function of the basal ganglia cannot be exaggerated: dopamine is implicated in a huge variety of disorders ranging from Parkinson's disease, through schizophrenia, major depression, attentional deficit hyperactive disorder etc, and ending in decision-making aberrations such as substance abuse and addiction. Understanding the computational and algorithmic role of dopamine in learning and action selection is a first step to reversing or treating such unfortunate conditions.

In the following, we first introduce the formal RL framework (for a comprehensive textbook account of RL methods, see Sutton and Barto (1998)). We then review (in Section 2) the multiple lines of evidence linking RL to the function of dopaminergic neurons in the mammalian midbrain. These data demonstrate the strength of the computational model and normative framework for interpreting and predicting a wide range of (otherwise confusing) neural activity patterns. Section 3 extends these results to more recent data from human imaging experiments. In these experiments, the combination of RL models of choice behavior and online imaging of whole-brain neural activity has allowed the detection of specific 'hidden variables' controlling behavior (such as the subjective value of different options) in the human brain. In Section 4, we discuss aspects of learning not associated with phasic dopamine signals, such as goal directed learning (which may be relatively dopamine-independent) and learning about the vigor (or rate) with which actions should be performed (whose neural underpinning has been suggested to be tonic levels of dopamine in the striatum). We conclude with a discussion of some of the limitations of the RL framework of learning, and highlight several open questions.

1. Reinforcement learning: Theoretical background

The modern form of RL arose historically from two separate and parallel lines of research. The first axis is mainly associated with Richard Sutton, formerly an undergraduate psychology major, and his doctoral thesis advisor, Andrew Barto, a computer scientist. Interested in artificial intelligence and agent-based learning and inspired by the psychological literature on Pavlovian and instrumental conditioning, Sutton and Barto developed what is today the core algorithms and concepts of RL (Barto, Sutton, & Anderson, 1983; Sutton, 1978; Sutton & Barto, 1990, 1998). In the second axis, stemming from a different background of operations research and optimal control, electrical engineers such as Dimitri Bertsekas and John Tsitsiklis developed stochastic approximations to dynamic programming methods (which they termed 'neuro-dynamic programming'), which led to similar reinforcement learning rules (e.g. Bertsekas and Tsitsiklis (1996)). The fusion of these two lines of research couched the behaviorally-inspired heuristic reinforcement learning algorithms in more formal terms of optimality, and provided tools for analyzing their convergence properties in different situations.

1.1. The Rescorla–Wagner model

The early impetus for the artificial intelligence trajectory can be traced to the early days of the field of 'mathematical psychology' in the 1950's, within which statistical models of learning were considered for the first time. In a seminal paper Bush and Mosteller (1951) developed one of the first detailed formal accounts of learning. Together with Kamin's (1969) insight that learning should occur only when outcomes are 'surprising', the Bush and Mosteller 'linear operator' model found its most popular expression in the now-classic Rescorla–Wagner model of Pavlovian conditioning (Rescorla & Wagner, 1972). The Rescorla–Wagner model, arguably the most influential model of animal learning to date, explained puzzling behavioral phenomena such as blocking, overshadowing and conditioned inhibition (see below) by postulating that learning occurs *only when events violate expectations*. For instance, in a conditioning trial in which two conditional stimuli CS_1 and CS_2 (say, a light and a tone) are presented, as well as an affective stimulus such as food or a tail-pinch (the unconditional stimulus; US), Rescorla and Wagner postulated that the associative strength of each of the conditional stimuli $V(CS_i)$ will change according to

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta \left[\lambda_{US} - \sum_i V_{old}(CS_i) \right]. \quad (1)$$

In this *error correcting* learning rule, learning is driven by the discrepancy between what was predicted ($\sum_i V(CS_i)$ where i indexes all the CSs present in the trial) and what actually happened (λ_{US}, whose magnitude is related to the worth of the unconditional stimulus, and which quantifies the maximal associative strength that the unconditional stimulus can support). η is a learning rate that can depend on the salience properties of both the unconditional and the conditional stimuli being associated.

At the basis of the Rescorla–Wagner model are two important (and innovative) assumptions or hypotheses: (1) learning happens only when events are not predicted, and (2) predictions due to different stimuli are summed to form the total prediction in a trial. Due to these assumptions, the model could explain parsimoniously several anomalous features of animal learning such as why an already predicted unconditional stimulus will not support conditioning of an additional conditional stimulus (as in blocking; Kamin, 1969); why differently salient conditional stimuli presented together might become differentially associated with an unconditional stimulus (as in overshadowing; Reynolds (1961)); and why a stimulus that predicts the *absence* of an expected unconditional stimulus acquires a negative associative strength (as in inhibitory conditioning; Konorski (1948) and Rescorla and Lolordo (1968)). Furthermore, the model predicted correctly previously unknown phenomena such as over-expectation (Kremer, 1978; Rescorla, 1970).

The Rescorla–Wagner model explains a large collection of behavioral data with one elegant learning rule, however, it suffers from two major shortcomings. First, by treating the conditional and unconditional stimuli as qualitatively different, it does not extend to the important phenomenon of second order conditioning. In second order conditioning if stimulus B predicts an affective outcome (say, fruit juice, or electric shock) and stimulus A predicts stimulus B, then stimulus A also gains reward predictive value. This laboratory paradigm is especially important given the prevalence of second (or higher) order conditioning in every-day life, a prime example for which is the conditioning of humans to monetary outcomes, which are second order predictors of a wide range of affectively desirable unconditional stimuli such as food and shelter. The second shortcoming of the Rescorla–Wagner rule is that its basic unit of learning is a conditioning *trial* as a discrete temporal object. Not only does this impose an experimenter-oriented parsing of otherwise continuous events, but it also fails to account for the sensitivity of conditioning to the different temporal relations between the conditional and the unconditional stimuli *within* a trial (that is, whether they appeared simultaneously or serially, their order of appearance, and whether there was a time lag between them).

1.2. Temporal difference learning

To overcome these two problems, Sutton and Barto (1990) suggested the *temporal difference learning rule* as a model of prediction learning in Pavlovian conditioning. Temporal-difference (TD) learning is an extension of the Rescorla–Wagner model that also takes into account the timing of different events. *Prima facie* the distinctions between the two model are subtle (see below). However, the differences allow the TD model to account for higher order conditioning and make it sensitive to the temporal relationships within a learning trial (Sutton & Barto, 1990). As will be discussed in Section 2, the TD model is also more consistent with findings regarding the neural underpinnings of RL.

In TD learning, the goal of the learning system (the 'agent') is to estimate the values of different states or situations, in terms of the *future* rewards or punishments that they predict. For example, from a learning standpoint, the TD model assumes that the goal of a rat running in a novel arena is to learn the value of various positions in the arena in terms of obtaining any available rewards. One way to do this would be to estimate for each location the average total amount of reward that the rat could expect to receive in the future, when starting from that location. This departure from Rescorla and Wagner's framework, in which predictions are only of the immediately forthcoming reward, turns out to be key.

In order to formally introduce TD learning, let us depart for the moment from animal conditioning and human decision-making. Consider a dynamic process (called a *Markov chain*) in which different states $S \in \mathcal{S}$ follow one another according to some predefined probability distribution $P(S_{t+1}|S_t)$, and rewards are observed at each state with probability $P(r|S)$. As mentioned, a useful quantity to predict in such a situation is the expected sum of all future rewards, given the current state S_t, which we will call the *value* of state S_t, denoted $V(S_t)$. Thus

$$\begin{aligned} V(S_t) &= E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \middle| S_t\right] \\ &= E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \middle| S_t\right] \end{aligned} \quad (2)$$

where $\gamma \le 1$ discounts the effect of rewards distant in time on the value of the current state. The discount rate was first introduced in order to ensure that the sum of future rewards is finite, however, it also aligns well with the fact that humans and animals prefer earlier rewards to later ones, and such *exponential discounting* is equivalent to an assumption of a constant 'interest rate' per unit time on obtained rewards, or a constant probability of exiting the task per unit time. The expectation here is with respect to both the probability of transitioning from one state to the next, and the probability of reward in each state. From this definition of state values it follows directly that

$$\begin{aligned} V(S_t) &= E[r_t|S_t] + \gamma E[r_{t+1}|S_t] + \gamma^2 E[r_{t+2}|S_t] + \cdots && (3) \\ &= E[r_t|S_t] + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)\,(E[r_{t+1}|S_{t+1}] \\ &\quad + \gamma E[r_{t+2}|S_{t+1}] + \cdots) && (4) \\ &= P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t) V(S_{t+1}) && (5) \end{aligned}$$

(assuming here for simplicity that rewards are Bernoulli distributed with a constant probability $P(r|S_t)$ for each state). This recursive relationship or *consistency* between consecutive state values lies at the heart of TD learning. The key to learning these values is that the consistency holds *only* for correct values (ie, those that correctly predict the expected discounted sum of future values). If the values are incorrect, there will be a discrepancy between the two sides of the equation, which is called the *temporal difference prediction error*

$$\delta_t = P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)V(S_{t+1}) - V(S_t). \tag{6}$$

This prediction error is a natural 'error signal' for improving estimates of the function $V(S_t)$. If we substitute this prediction error for the 'surprise' term in the Rescorla–Wagner learning rule, we get

$$V(S_t)_{new} = V(S_t)_{old} + \eta \cdot \delta_t, \tag{7}$$

which will update and improve the state values until all prediction errors are 0, that is, until the consistency relationship between all values holds, and thus the values are correct.

However, returning to prediction learning in real-world scenarios, we note that this updating scheme (which is at the basis of a collection of methods collectively called "dynamic programming"; (Bellman, 1957)) has one major problem: it requires knowledge of the dynamics of the environment, that is, $P(r|S_t)$ and $P(S_{t+1}|S_t)$ (the "world model") must be known in order to compute the prediction error δ_t in Eq. (6). This is clearly an unreasonable assumption when considering an animal in a Pavlovian conditioning task, or a human predicting the trends of a stock. Werbos (1977) in his "heuristic dynamic programming methods", and later Barto, Sutton, and Watkins (1989) and Bertsekas and Tsitsiklis (1996), suggested that in a "model-free" case in which we can not assume knowledge of the dynamics of the environment, the environment itself can supply this information stochastically and incrementally. Every time an animal is in the situation that corresponds to state S_t, it can *sample* the reward probability in this state, and the probabilities of transitions from this state to another. As it experiences the different states repeatedly within the task, the animal will obtain unbiased samples of the reward and transition probabilities. Updating the estimated values according to these stochastic samples (with a decreasing learning rate or 'step-size') will eventually lead to the correct predictive values. Thus the stochastic prediction error

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t) \tag{8}$$

(where r_t is the reward observed at time t, when in state S_t, and S_{t+1} is the next observed state of the environment) can be used as an approximation to Eq. (6), in order to learn in a "model-free" way the true predictive state values. The resulting learning rule is

$$V_{new}(S_t) = V_{old}(S_t) + \eta(r_t + \gamma V_{old}(S_{t+1}) - V_{old}(S_t)). \tag{9}$$

Finally, incorporating into this learning rule the Rescorla–Wagner assumption that predictions due to different stimuli S_i comprising the state of the environment are additive (which is not the only way, or necessarily the most sensible way to combine predictions, see Dayan, Kakade, and Montague (2000)), we get for all S_i present at time t

$$V_{new}(S_{i,t}) = V_{old}(S_{i,t}) + \eta\left[r_t + \gamma \sum_{S_k @ t+1} V_{old}(S_{k,t+1}) - \sum_{S_j @ t} V_{old}(S_{j,t})\right], \tag{10}$$

which is the TD learning rule proposed by Sutton and Barto (1990). As detailed above, the formal justification for TD learning as a method for optimal RL derives from its direct relation to dynamic programming methods (Barto, Sutton, & Watkins, 1990; Sutton, 1988; Watkins, 1989). This ensures that using TD learning, animals can learn the optimal (true) predictive values of different events in the environment, even when this environment is stochastic and its dynamics are unknown.

Indeed this rule is similar, but not identical, to the Rescorla–Wagner rule. As in the Rescorla–Wagner rule, η is a learning rate or step-size parameter, and learning is driven by discrepancies between available and expected outcomes. However, one difference is that in TD learning time within a trial is explicitly represented and learning occurs at every timepoint within a trial. Moreover, in the specific tapped delay line representation variant of TD learning described in Eq. (10), stimuli create long-lasting memory traces (representations), and a separate value $V(S_{i,t})$ is learned for every timepoint of this trace (for instance, a stimulus might predict a reward exactly five seconds after its presentation). A second and more important difference is in how predictions, or expectations, are construed in each of the models. In TD learning, the associative strength of the stimuli (and traces) at time t is taken to predict not only the immediately forthcoming reward r_t, but also the future predictions due to those stimuli that will still be available in the next time-step $\sum_{S_j @ t+1} V(S_{j,t+1})$, with $\gamma \leq 1$ discounting these future delayed predictions.

1.3. Optimal action selection

The above holds whenever the probabilities of transitioning between different states of the environment are fixed, as in Pavlovian conditioning (in which the animal can not influence events by means of its actions) or in situations in which the animal has a fixed behavioral policy (Sutton, 1988). But what about improving action selection in order to obtain more rewards? That is, what about instrumental conditioning? Since the environment rewards us for our *actions*, not our *predictions* (be they correct as they may), one might argue that the ultimate goal of prediction learning is to aid in action selection.

The problem of optimal action selection is especially difficult in those (very common) cases in which actions have long-term consequences (such as in a game of checkers), or in which attaining outcomes requires a series of actions. The main problem, in these cases, is that of *credit assignment* (Barto et al., 1983; Sutton, 1978; Sutton & Barto, 1998) – how to figure out, when reaching the outcome (for instance, a win or a loss), what actions (perhaps in the distant past) were key to obtaining this outcome. The correct assignment of credit is crucial for learning to improve the behavioral policy: those actions that ultimately lead to rewards should be repeated, and those that lead to punishment should be avoided. This is true in the animal domain as well: when reaching a dead-end in a maze, how will a rat know which of its previous actions was the erroneous one? RL methods solve the credit assignment problem by basing action selection not only on immediate outcomes, but also on future value predictions such as those we discussed above, which embody predictions of long-term outcomes.

How does action selection then interact with state evaluation (for instance, using TD learning as above)? First, note that given predictive state values, the best action to select is the one that leads to the state with the highest value (e.g. McClure, Daw, and Montague (2003)). In fact, Samuel's 1959 checker player, the first notable application of TD learning (even prior to its conception in its modern form), used this method to select actions. However, this necessitates knowledge of how transitions between states depend on actions, that is, what is the probability of transitioning to each state, given a specific action. What if such knowledge is not available? For example, imagine deciding whether to buy or to sell a stock on the stock market – clearly this decision would be trivial if only you knew whether the stock's price would increase or decrease as a result of your (and the rest of the market's) actions. But what can a human or a rat do in the completely model-free case, ie, without knowledge of how different actions will influence the state of the environment?

1.3.1. Actor/Critic methods

In one of the first RL papers, which was inspired by neural-network models of learning, Barto et al. (1983) showed that the credit assignment problem can be effectively solved by a learning system comprised of two neuron-like elements. One unit, termed the "adaptive critic element (ACE)", constructed an evaluation of different states of the environment, using a temporal-difference-like learning rule from which the TD learning rule above was later developed. This evaluation was used to augment the external reinforcement signal and train through a trial-and-error process a second unit, the "associative search element (ASE)", to select the correct action at each state. These two elements were the precursors of the modern-day Actor/Critic framework for model-free action selection which has been closely associated with reinforcement learning and action selection in the brain.

The insight in the ASE–ACE model, first due to Sutton (1978), is that even when the external reinforcement for a task is delayed (as when playing checkers), a temporal difference prediction error can convey, at every timestep, a surrogate 'reinforcement' signal that embodies both immediate outcomes and future prospects, to the action just chosen. This is because, in the absence of external reinforcement (*ie*, $r_t = 0$), the prediction error δ_t in Eq. (8) becomes $\gamma V(S_{t+1}) - V(S_t)$, that is, it compares the values of two consecutive states and conveys information regarding whether the chosen action has led to a state with a higher value than the previous state (ie, to a state predictive of more future reward) or not. This means that whenever a positive prediction error is encountered, the current action has improved prospects for future rewards, and should be repeated. The opposite is true for negative prediction errors, which signal that the action should be chosen less often in the future. Thus the agent can learn an explicit *policy* — a probability distribution over all available actions at each state $\pi(S, a) = p(a|S)$, by using the following learning rule at every timestep

$$\pi(S, a)_{new} = \pi(S, a)_{old} + \eta_\pi \delta_t \tag{11}$$

where η_π is the policy learning rate and δ_t is the prediction error from Eq. (8).

Thus, in Actor/Critic models, a Critic module uses TD learning to estimate state values $V(S)$ from experience with the environment, and the same TD prediction error is also used to train the Actor module, which maintains and learns a policy π (Fig. 1). This method is closely related to policy improvement methods in dynamic programming (Sutton, 1988), and Williams (1992) and Sutton, Mcallester, Singh, and Mansour (2000) have shown that in some cases the Actor/Critic can be construed as a gradient climbing algorithm for learning a parameterized policy, which converges to a local maximum (see also Dayan and Abbott (2001)). However, in the general case Actor/Critic methods are not guaranteed to converge on an optimal behavioral policy (cf. Baird (1995) and Konda and Tsitsiklis (2003)). Nevertheless, some of the strongest links between RL methods and neurobiological data regarding animal and human decision making have been related to the Actor/Critic framework. Specifically, Actor/Critic methods have been extensively linked to instrumental action selection and Pavlovian prediction learning in the basal ganglia (e.g. Barto (1995), Houk, Adams, and Barto (1995) and Joel, Niv, and Ruppin (2002)), as will be detailed below.

1.3.2. State-action values

An alternative to Actor/Critic methods for model-free RL, is to explicitly learn the predictive value (in terms of future expected rewards) of taking a specific action at a certain state, that is, learning the value of the state-action pair, denoted $\mathcal{Q}(S, a)$. In his Ph.D. thesis, Watkins (1989) suggested $\mathcal{Q}$*-learning* as a modification of TD learning that allows one to learn such $\mathcal{Q}$-values (and brings TD learning closer to dynamic programming methods of 'policy iteration'; Howard (1960)). The learning rule is quite similar to the state-value learning rule above

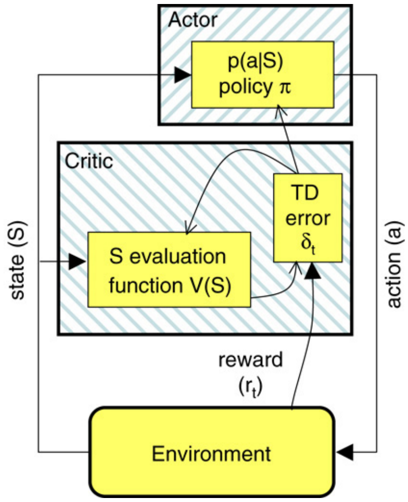


Fig. 1. Actor/Critic architecture: The state S_t and reinforcement signal r_t are conveyed to the Critic by the environment. The Critic then computes a temporal difference prediction error (Eq. (8)) based on these. The prediction error is used to train the state value predictions $V(S)$ in the Critic, as well as the policy $\pi(S, a)$ in the Actor. Note that the Actor does not receive direct information regarding the actual outcomes of its actions. Rather, the TD prediction error serves as a surrogate reinforcement signal, telling the Actor whether the (immediate and future expected) outcomes are better or worse than previously expected. Adapted from Sutton and Barto (1998).

$$\mathcal{Q}(S_t, a_t)_{new} = \mathcal{Q}(S_t, a_t)_{old} + \eta\delta_t \tag{12}$$

albeit with a slightly different TD prediction error driving the learning process

$$\delta_t = r_t + \max_a \gamma \mathcal{Q}(S_{t+1}, a) - \mathcal{Q}(S_t, a_t) \tag{13}$$

where the max operator means that the temporal difference is computed with respect to what is believed to be the best action at the subsequent state S_{t+1}. This method is considered 'off-policy' as it takes into account the best future action, even if this will not be the action that is actually taken at S_{t+1}. In an alternative 'on-policy' variant called SARSA (the acronym for state-action-reward-state-action), the prediction error takes into account the next chosen action, rather than the best possible action, resulting in a prediction error of the form:

$$\delta_t = r_t + \gamma \mathcal{Q}(S_{t+1}, a_{t+1}) - \mathcal{Q}(S_t, a_t). \tag{14}$$

In both cases, action selection is easy given $\mathcal{Q}$-values, as the best action at each state S is that which has the highest $\mathcal{Q}(S, a)$ value. That is, learning $\mathcal{Q}$-values obviates the need for separately learning a policy. Furthermore, dynamic programming results regarding the soundness and convergence of 'policy iteration' methods (in which a policy is iteratively improved through bootstrapping of the values derived given each policy; Howard (1960) and Bertsekas and Tsitsiklis (1996)) ensure that if the proper conditions on the learning rate are met and all state-action pairs are visited infinitely often, both $\mathcal{Q}$-learning and SARSA will indeed converge to the true optimal (in case of $\mathcal{Q}$-learning) or policy-dependent (in the case of SARSA) state-action values. Interestingly, recent electrophysiological recordings in non-human primates (Morris et al., 2006) and in rats (Roesch et al., 2007) suggest that dopaminergic neurons in the brain may indeed be conveying a prediction error that is based on state-action values (rather than state values, as in the Actor/Critic model), with the former study supporting a $\mathcal{Q}$-learning prediction error, and the latter a SARSA prediction error. Whether these results mean that the brain is not using an Actor/Critic scheme at all, or whether the Actor/Critic framework could be modified to use state-action values (and indeed, the potential advantages of such a scheme) is still an open question (Niv, Daw, & Dayan, 2006)

2. Neural correlates of reinforcement learning

In recent years, RL models such as those briefly described above have been applied to a wide range of neurobiological and behavioral data. In particular, the computational function of neuromodulators such as dopamine, acetylcholine, and serotonin have been addressed using the RL framework. Among these neuromodulatory systems, the dopamine system is the most studied, perhaps due to its implication in conditions such as Parkinson's disease, schizophrenia, and drug addiction as well as its long-suspected functions in reward learning and working memory. It is in elucidating the role of dopamine signals in the brain, that computational models of learning in general, and TD learning in particular, have had their most profound impact on neuroscience.

The link between dopamine and RL was made in the mid '90s. On the background of a dominant hypothesis that viewed dopamine as the brain's reward signal (Wise, Spindler, de Wit, & Gerberg, 1978; Wise, Spindler, & Legault, 1978), pioneering extracellular recordings in the midbrain of awake and behaving moneys for the lab of Wolfram Schultz showed that dopaminergic neurons did not simply signal the primary motivational value of rewarding stimuli such as food and water. In these experiments, recordings were done while the monkeys underwent simple instrumental or Pavlovian conditioning (Ljungberg, Apicella, & Schultz, 1992; Romo & Schultz, 1990; Schultz, Apicella, & Ljungberg, 1993). Surprisingly, although the recorded cells showed phasic bursts of activity when the monkey was given a rewarding sip of juice or a morsel of apple, if food delivery was consistently preceded by a tone or a light, after a number of trials the dopaminergic response to reward disappeared. Contrary to the "dopamine equals reward" hypothesis, the disappearance of the dopaminergic response to reward delivery did not accompany extinction, but rather it followed acquisition of the conditioning relationship – as the cells ceased to respond to rewards the monkeys began showing conditioned responses of anticipatory licking and arm movements to the reward-predictive stimulus. Indeed, not only the monkeys responded to the tone – the neurons now began responding to the tone as well, showing distinct phasic bursts of activity whenever the tone came on. This was also true for the difference between self-initiated reaching for reward, in which case dopamine neurons responded phasically to touching the reward, versus cue-initiated movements, in which case the neurons responded to the cue rather than to the reward. These results were extremely puzzling, as is evident by the conclusions of those early papers, which portray a handful of separate functions attributed to different types of dopaminergic responses, and reflect the dire need for a unifying theory.

2.1. *The reward prediction error hypothesis of dopamine*

And a unifying theoretical interpretation was not long to follow. In the mid '90s a number of theoreticians interested in computer science and computational neuroscience recognized the unmistakable fingerprint of reinforcement learning signals in these data, and suggested that the phasic firing of dopaminergic neurons reflects a *reward prediction error* (Barto, 1995; Montague, Dayan, Nowlan, Pouget, & Sejnowski, 1993; Montague, Dayan, Person, & Sejnowski, 1995; Montague, Dayan, & Sejnowski, 1994, 1996). Indeed, the hallmark of temporal difference prediction errors is that they occur only when motivationally significant events are *unpredicted*. This explains why dopaminergic neurons show burst firing to rewards early in training (when they were unexpected), but not later in training, after the animal has learned to expect reward on every trial. Similarly, early in training neutral cues that precede the reward should not cause a prediction error (as they themselves are not rewarding), but later in training, once they have acquired predictive value (ie, $V(cue) > 0$), an unexpected onset of such a cue should generate a prediction error (as $\delta_t = r_t + \gamma V(cue) - V(no\, cue) = \gamma V(cue) > 0$), and thus dopaminergic firing. Fig. 2 illustrates these effects in a simulation of TD learning, and, for comparison, in the activity of dopaminergic neurons (from Schultz et al. (1997)). The simulation is of a Pavlovian conditioning scenario in which a tone CS is followed two seconds later by a food US; the electrophysiological recordings are from an analogous instrumental task in which a cue signaled the availability of reward, provided the monkey responded correctly with a rapid reaching movement. Panels (a,d) illustrate the prediction error to the appetitive US early in training, and panels (b,e) show responses after training – now shifted to the time of the unexpected CS, rather than the US. Moreover, in trials in which the US is not delivered, a *negative* reward prediction error occurs at the precise time of the expected US delivery, as is illustrated by panels (c,e). The discrepancies between the simulation and the dopamine neuron firing patterns in terms of the magnitude and spread of the prediction errors at the time of the reward likely result from the temporal noise in reward delivery in the instrumental task, and the asymmetric representation of negative and positive prediction errors around the baseline firing rate of these neurons (Niv, Duff, & Dayan, 2005). Note that the prediction error to the CS occurs only if this cue is not itself predicted by earlier events. For instance, training with an earlier cue (CS2) that reliably precedes this CS, would result in the dopaminergic response shifting to CS2, that is, to the earliest possible cue that predicts the reward (Schultz et al., 1993). The fact that an unexpected cue that predicts reward generates a prediction error similar in all aspects to that generated by an unexpected reward, is the reason that second order conditioning can occur, with a predictive cue supporting new conditioning as if it were itself a reward.

The close correspondence between the phasic dopaminergic firing patterns and the characteristics of a temporal difference prediction error led Montague et al. (1996) to suggest the *reward prediction error hypothesis of dopamine* (see also Schultz et al. (1997)). Within this theoretical framework, it was immediately clear why dopamine is necessary for reward mediated learning in the basal ganglia. The link with RL theory provided a normative basis for understanding not only why dopamine neurons fire when they do, but also what the *function* of these firing patterns might be. If dopamine signals a reward prediction error, this could be used for prediction learning and for action learning in dopaminergic targets. Indeed, behaviorally the shift in dopaminergic activity from the time of reward to the time of the predictor (Takikawa, Kawagoe, & Hikosaka, 2004) resembles the shift of behavioral responding from the time of the US to that of the CS in Pavlovian conditioning experiments (Hollerman & Schultz, 1998; Schultz et al., 1997). Furthermore, there is physiological evidence for dopamine-dependent (or even dopamine-gated) plasticity in the synapses between the cortex and the striatum (Arbuthnott, Ingham, & Wickens, 2000; Reynolds, Hyland, & Wickens, 2001; Wickens, Begg, & Arbuthnott, 1996; Wickens & Kötter, 1995).

The above basic characteristics of phasic dopaminergic responding have since been replicated in many variants (e.g. Bayer and Glimcher (2005), Hollerman and Schultz (1998), Schultz (1998), Takikawa et al. (2004) and Tobler, Dickinson, and Schultz (2003)). In fact, recent work investigating the detailed quantitative implications of the prediction error hypothesis has demonstrated that the correspondence between phasic dopaminergic

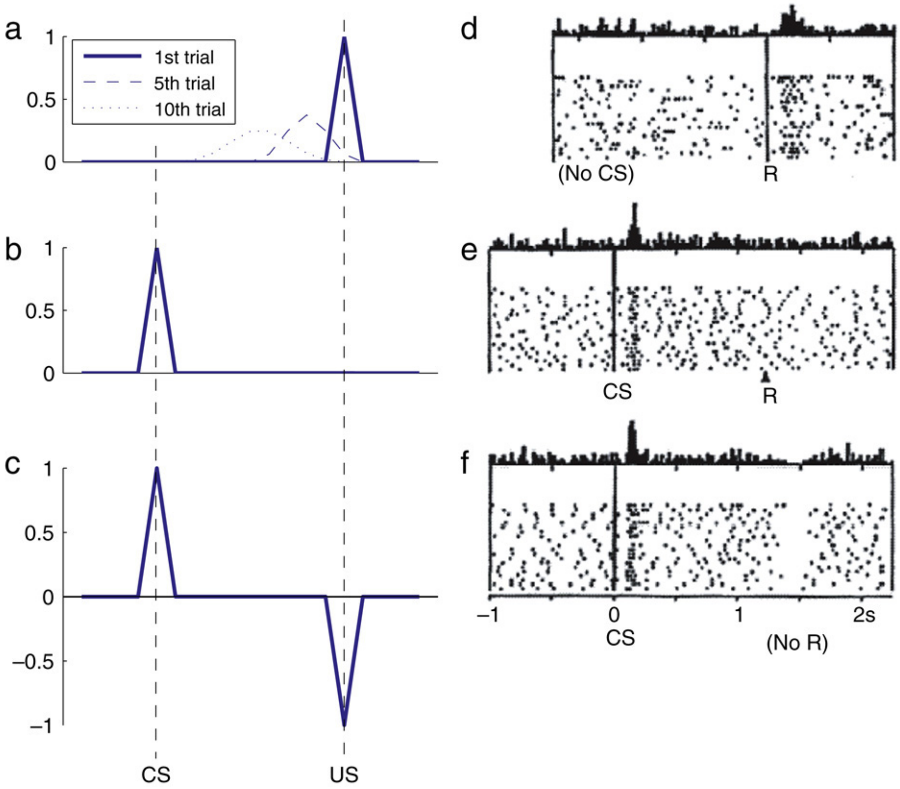


Fig. 2. (a–c) Temporal difference prediction errors in a Pavlovian conditioning task. A tone CS is presented at random times, followed 2 seconds later by a juice US. (a) In the beginning of training, the juice is not predicted, resulting in prediction errors at the time of the juice US. With learning, the prediction error propagates backward within the trial (trials 5 and 10 are illustrated; Niv et al., 2005) as predictive values are learned Eq. (9). (b) After learning, the now-predicted US no longer generates a prediction error. Rather, the unpredicted occurrence of the tone CS is accompanied by a prediction error. (c) The unexpected omission of the US causes a negative prediction error at the time in which the US was expected, as in this trial reality was worse than expected. In these simulations the CS was represented over time with the commonly used serial compound state representation (Kehoe, 1977; Sutton & Barto, 1990), and there was no discounting ($\gamma = 1$). Other representation schemes make different predictions for how the prediction error propagates backward, but do not differ in their predictions for the activity patterns in a fully learned task. (d–f) Firing patterns of dopaminergic neurons in monkeys performing an analogous instrumental conditioning task. Each raster plot shows action potentials (dots) with different rows representing different trials, aligned on the time of the cue (or the reward). Histograms show activity summed over the trials plotted below. (d) When a reward unexpectedly obtained, dopaminergic neurons respond with a phasic burst of firing. (e) After conditioning with a predictive visual cue (which, in this task, predicted a food reward if the animal quickly performed the correct reaching response), the reward no longer elicits a burst of activity, and the phasic burst now occurs at the presentation of the predictive cue. (f) When the food reward was unexpectedly omitted, dopaminergic neurons showed a precisely-timed pause in firing, below their standard background firing rate. Subplots (d–f) adapted with permission from Schultz et al. (1993).

firing and TD prediction errors goes far beyond the three basic characteristics depicted in Fig. 2. For instance, using general linear regression, Bayer and colleagues have rigorously shown that the contribution of previously experienced rewards to the dopaminergic response to the current reward is exactly according to an exponentially weighted average of past experience, as is implied by the TD learning rule (Bayer & Glimcher, 2005; Bayer, Lau, & Glimcher, 2007). Moreover, conditioned stimuli predicting probabilistic rewards or rewards of different magnitudes have been shown to elicit a phasic dopaminergic response that is proportional to the magnitude and/or probability of the expected reward (Fiorillo, Tobler, & Schultz, 2003; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004; Tobler, Fiorillo, & Schultz, 2005, Fig. 3a, b) and firing patterns in tasks involving probabilistic rewards are in accord with a constantly back-propagating error signal (Niv et al., 2005, Fig. 3b, c). With regard to delayed rewards, recent results from recordings in rodents show that dopaminergic activity to a cue predicting a delayed reward is attenuated in proportion to the delay (Fig. 4), as would be expected from a signal predicting the expected sum of *discounted* future rewards (Roesch et al., 2007). Impressively, even in sophisticated conditioning tasks such as blocking and appetitive conditioned inhibition, dopaminergic responses are in line with the predictions of TD learning (Tobler et al., 2003, 2005; Waelti, Dickinson, & Schultz, 2001). Finally, measurements of extracellular dopamine in behaving rodents using fast scan cyclic voltammetry have confirmed that phasic changes in the level of dopamine in target structures (specifically, in the striatum) also conform quantitatively to a prediction error signal (Paul Phillips, personal communication; see also Day, Roitman, Wightman, and Carelli (2007), Knutson, Delgado, and Philips (2008) and Walton, Gan, Barnes, Evans, and Phillips (2006)), despite the nonlinear relationship between dopamine neuron firing and actual synaptic discharge of the transmitter (Montague et al., 2004).

The prediction error theory of dopamine is a *computationally precise* theory of how phasic dopaminergic firing patterns are generated. It suggests that the input that dopaminergic neurons receive from their diverse afferents (which include the medial prefrontal cortex, the nucleus accumbens shell, the ventral pallidum, the central nucleus of the amygdala, the lateral hypothalamus, the habenula, the cholinergic pedunculopontine nucleus, the serotoninergic raphe and the noradrenergic locus coeruleus; Christoph, Leonzio, and Wilcox (1986), Floresco, West, Ash, Moore, and Grace (2003), Geisler and Zahm (2005), Matsumoto and Hikosaka (2007) and Kobayashi and Okada (2007)) conveys information about current motivationally significant events (r_t), and the predictive value of the current state $V(S_t)$, and that the circuitry in the dopaminergic nuclei uses this information to compute a temporal difference reward prediction error. Moreover, it suggests that dopamine provides target areas with a neural signal that is theoretically appropriate for controlling learning of both predictions and reward-optimizing actions.

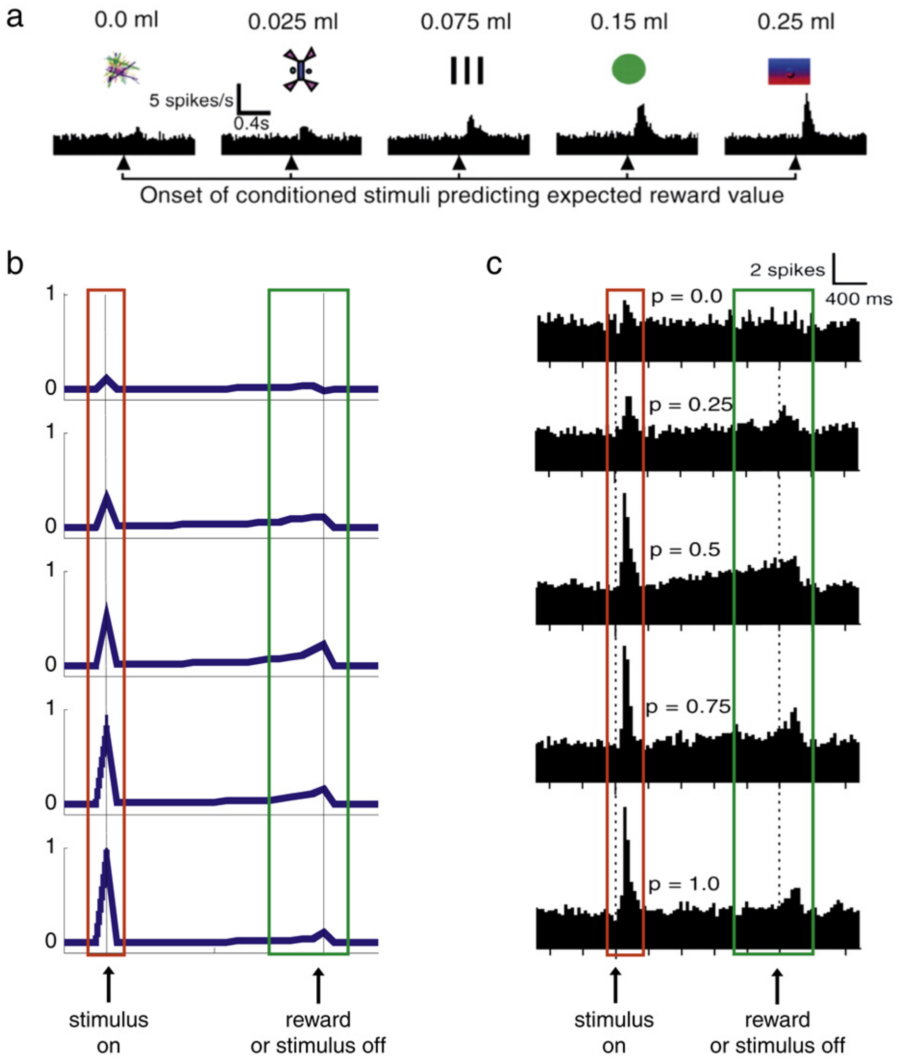


Fig. 3. Dopaminergic firing patterns comply with the predictions of TD learning. (a) Phasic responses to a cue predicting reward are proportional to the magnitude of the predicted reward (adapted with permission from Tobler et al., 2005). (b, c) When different cues predict the same reward but with different probabilities, the prediction error at the time of the cue is proportional to the predicted probability of reward (red (left) rectangles; compare panel b (simulation) to panel c (data)). However, due to the low baseline firing rate of midbrain dopaminergic neurons, negative prediction errors can only be encoded asymmetrically about the base firing rate, with a shallower 'dip' in firing rate to encode negative prediction errors as compared to the height of the 'peak' by which positive prediction errors are encoded. As a result, when rewards are probabilistic, averaging over rewarded and unrewarded trials will create an apparent ramp leading up to the time of the reward (green (right) rectangles; compare panel b (simulation) to panel c (data)). Panel b adapted with permission from (Niv et al., 2005), panel c adapted with permission from (Fiorillo et al., 2003).

Following the analogy between the dopamine signal and the temporal difference prediction error signal in Actor/Critic models (Joel et al., 2002), it has been suggested that dopaminergic signals originating in the ventral tegmental area and terminating in ventral striatal and frontal areas are used to train predictions, as in the Critic (Barto, 1995; Waelti et al., 2001), while a similar signal reported by dopaminergic neurons in the substantia nigra pars compacta to dorsal striatal target areas, is used to learn an action-selection policy, as in the Actor (Houk et al., 1995; Joel & Weiner, 1999; Miller & Wickens, 1991; Wickens & Kötter, 1995).

As should be the case when researching the basic characteristics of a neural signal, the studies mentioned above mostly used rather simple Pavlovian or instrumental tasks, in which trials include one unambiguous stimulus and one reward. Given the accumulation of positive results, it seems that the time is now ripe to test the reward prediction error theory of dopamine in more complex scenarios, for instance situations in which there are a number of conflicting predictive cues, tasks in which several actions are necessary to obtain an outcome, or tasks in which there are several possible outcomes to choose from. In these cases the theory is not as prescriptive – there are different ways to combine predictive cues, or to generate a prediction error that does or does not depend on the actual chosen action (ie, SARSA, $\mathcal{Q}$-learning and Actor/Critic that were detailed in Section 1.3, as well as others like advantage learning (Baird, 1993) that we did not detail), thus electrophysiological evidence is key to informing the RL theory and constraining the algorithm actually used by the brain.

Several studies have recently made progress in this direction. Morris et al. (2006) trained monkeys in a standard instrumental task in which cues predicted reward with different probabilities. In some trials, however, the monkeys were given a choice between two of these cues. Single unit recordings in the substantia nigra pars compacta showed that in these trials the cue-elicited dopaminergic firing matched best the prediction errors corresponding to the cue that would subsequently be chosen (even though the monkey could only signal its choice seconds later). This is contrary to the straightforward predictions of an Actor/Critic mechanism, and more in line with SARSA learning. Interestingly, recordings from the ventral tegmental area of rats performing a more dynamic odor-discrimination task (Roesch

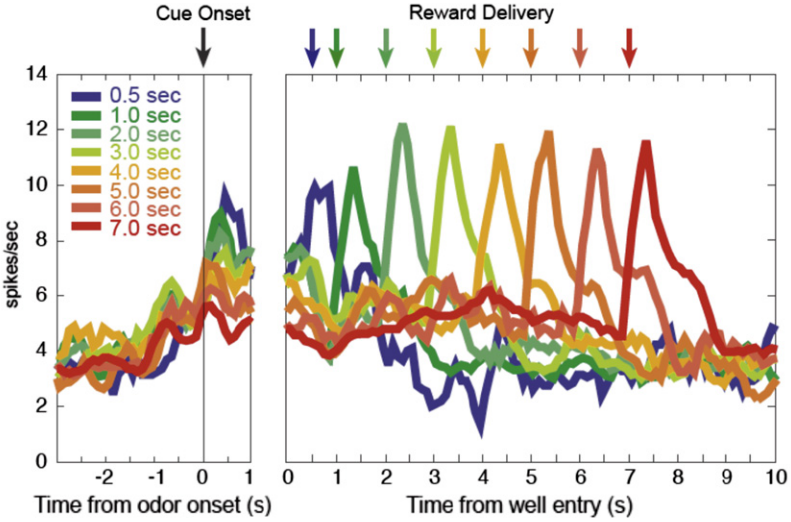


Fig. 4. Average firing rate of 19 dopaminergic neurons, recorded in rats performing an odor-discrimination task in which one of the odors predicted that a reward would be delivered in a food-well, with some delay. Color indicates the length of the delay preceding reward delivery from 0.5 to 7 s. Activity is aligned on odor onset (left) and food-well entry (right). Note that the response to the (not fully predicted) reward is similar in all trial types (with the earliest rewards perhaps better predicted, and thus accompanied by smaller prediction errors), but the response at the time of the predictive cue depends on the predicted delay of the reward, with longer predicted delays eliciting a smaller dopaminergic response. Adapted with permission from Roesch et al. (2007). (For a color version of this figure, the reader is referred to the web version of this article.)

et al., 2007) were taken to suggest that dopmainergic activity in choice scenarios was better described by $\mathcal{Q}$-learning. There are many differences between these two studies, including the animal species, the dopaminergic nuclei in which recordings were made, the task and the amount of training (or overtraining) of the animals, any of which could be invoked to explain their different results. However, the detailed activation patterns in the latter study, as well as results from a task in which monkeys engaged in a difficult random-dot motion discrimination task (Nomoto, Watanabe, & Sakagami, 2007), suggest that predictions (and thus prediction errors) can be sensitive to the information available at every timepoint, with stimuli represented before a choice is made, and chosen cues represented only later. These findings suggest a possible reconciling between the different results in terms of different representations in the tasks, and further highlights the need to study, from a theoretical point of view, as well as an experimental one, the effects of different state representations on TD learning.

As precise as the prediction error hypothesis of dopamine is, other open questions are abounding. Many of these will likely require modifications and enhancements of the currently highly simplified basic theory (Dayan & Niv, 2008). One extremely pressing issue is that of prediction of aversive events such as pain. Interestingly, dopaminergic neurons do not seem to be involved in the signaling or prediction errors for aversive outcomes (Mirenowicz & Schultz, 1996; Tobler et al., 2003; Ungless, Magill, & Bolam, 2004), although they do signal negative prediction errors due to the absence of appetitive outcomes (Bayer et al., 2007). Despite the behavioral similarities between the loss of a reward and the receipt of a punishment, these seem to be separated in terms of prediction learning, and it is currently far from clear what the substrate for aversive prediction learning might be Daw, Kakade, and Dayan (2002) and Nakamura, Matsumoto, and Hikosaka (2008).

We should also mention that there are alternative psychological theories regarding the role of dopamine in conditioned behavior (for a recent review, see Berridge (2007)). These include Redgrave and colleagues' 'incentive salience' (e.g. Horvitz (2000), Redgrave and Gurney (2006) and Redgrave, Prescott, and Gurney (1999)), Berridge and Robbinson's 'wanting' versus 'liking' (e.g. Berridge (2007) and Berridge and Robinson (1998)), and ideas about dopamine signaling uncertainty (Fiorillo et al., 2003). A discussion of the merits and pitfalls of the different theories is beyond the scope of this review. Moreover, such a discussion unfortunately involves the unsatisfactory comparison of qualitative suggestions to a quantitatively precise theory, rendering it difficult for any definitive conclusions to be reached. Nevertheless, it is our personal opinion that, in as far as these theories are indeed fundamentally different from the prediction error theory (which is not always clear), to date no alternative has mustered as convincing and multidirectional experimental support as the prediction error theory of dopamine.

2.2. *RL correlates in functional imaging of human decision-making*

Although animals can display complex decision-making behavior that is still well beyond our current understanding of the brain, ultimately we are interested in understanding human decision-making and how (and whether) it is related to the RL framework. While the characteristics of human conditioning are similar to those of animal conditioning, the possibility of instructing subjects verbally allows for much more elaborate paradigms in human experiments. Of course, our ability to measure neural processes in humans is much more limited. One technique that has been used extensively to study the underpinnings of RL in the human brain is functional magnetic resonance imaging (fMRI), in which correlates of neural activity can be measured non-invasively, albeit with a low signal-to-noise ratio (necessitating averaging over many trials or subjects) and poor temporal and spatial resolution (seconds and millimeters, respectively).

One advantage of fMRI is that it allows imaging of activity throughout the entire brain, rather than in only a small population of neurons. This also has disadvantages, in terms of the statistical analysis of the enormous volume of data collected even in a single experimental session. One might argue that fMRI thus places a premium on using precise computational models for data analysis. A model-driven analysis allows us to make precise hypotheses regarding *hidden variables* that control learning and decision-making, such as state values or prediction errors, and search for these in the brain. Using the computational model we can quantitatively specify the dynamics of the hidden variable even within a non-stationary task, and search the brain for a signal with a similar temporal profile. Identifying a neural correlate for such a signal advances our understanding of the brain in a way that would not be possible without the model, and can also lend powerful support for the model that gave rise to the specific values of the hidden variable.

A more pervasive disadvantage of fMRI investigations, however, is that the neural underpinnings of the measured blood oxygen level dependent (BOLD) signal are unclear. Some studies suggest that the BOLD signal in a brain area correlates with local field potentials in that area (signals that are themselves poorly understood, but are thought to be related to the overall dendritic synaptic activity within a volume of tissue), and thus reflects the input activity impinging on an area, rather than the (output) spiking activity of neurons within that area (Logothetis, 2003; Logothetis & Wandell, 2004). However, in these studies the correlation between BOLD and local field potentials was only slightly stronger than the (weak) correlation with spiking activity, making the results inconclusive. Furthermore, it is not clear that BOLD reflect the same underlying neural processes in all brain areas. For instance, because dopamine can directly affect the dilation and contraction of local blood vessels, it can affect BOLD measurements directly in brain areas in which extracellular concentration of dopamine is pronounced (Attwell & Iadecola, 2002; Krimer, Muly, Williams, & Goldman-Rakic, 1998). Dopamine is also known to affect local oscillatory activity (Costa et al., 2006), a prominent determinant of local field potentials, and thus perhaps BOLD. Indeed, these caveats go both ways: since dopamine is a major neuromodulator of interest in RL, its direct measurement is actually of major interest. However, despite its possible direct effect on BOLD signals, one cannot interpret BOLD measurements as reflecting dopaminergic activity per se (Knutson & Gibbs, 2007). Keeping these caveats in mind, we now turn to the specific use of RL models in fMRI studies of learning and decision making in the human brain.

The first fMRI studies to search for prediction errors in humans implicated the nucleus accumbens and the orbitofrontal cortex (Berns, McClure, Pagnoni, & Montague, 2001; Knutson, Adams, Fong, & Hommer, 2001; Pagnoni, Zink, Montague, & Berns, 2002), both major dopaminergic targets. O'Doherty, Dayan, Friston, Critchley, and Dolan (2003) and McClure, Berns, and Montague (2003) then used a hidden-variable analysis such as the one described above, to identify the neural correlates of model-derived TD prediction errors. These studies again implicated the nucleus accumbens (the ventral striatum) as well as the putamen (the dorsolateral striatum). O'Doherty et al. (2004) then showed that fMRI correlates of prediction error signals can be dissociated in the dorsal and ventral striatum according to whether active choice behavior is required in order to obtain reward (ie, instrumental conditioning) or not (Pavlovian conditioning). In the passive prediction-learning task the reward prediction error was evident only in the ventral striatum, while in the active choice task it was evident in both the ventral and the dorsolateral striatum. These findings supported a previously suggested mapping of an Actor/Critic architecture in the basal ganglia, according to which the ventral striatum includes a prediction-learning Critic, and the dorsal striatum hosts a policy-learning Actor (Joel et al., 2002, but see Section 2.3 for a more detailed parcellation of goal-directed and habitual instrumental learning in the dorsomedial and dorsolateral striatum, respectively).

Indeed, correlates of prediction errors in the dorsal and ventral striatum have now been seen in multiple studies (e.g. Li, McClure, King-Casas, and Montague (2006), Preuschoff, Bossaerts, and Quartz (2006) and Schönberg, Daw, Joel, and O'Doherty (2007)). As mentioned, although single unit recordings do not show prediction error encoding in the striatum, these results are in line with the fact that the striatum is a major target of dopaminergic influence. Indeed, dopaminergic manipulations (e.g. administration of dopamine enhancers (agonists)) or dopamine receptor blockers (antagonists)) in such tasks have been shown to influence both the BOLD measurement of prediction-error activity and learning and action selection (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006), and recent results show that better learners show a higher correlation of striatal BOLD with a reward prediction error (Schönberg et al., 2007). However, fMRI results cannot isolate dopaminergic activity from other activity in the brain (specifically, from the effects of other striatal afferents), and might not differentiate inhibitory from excitatory afferent activity, as is illustrated by the fact that BOLD correlates of *positive* prediction errors for pain and punishment have also been found in the striatum (Jensen et al., 2007; Menon et al., 2007; Seymour et al., 2004).

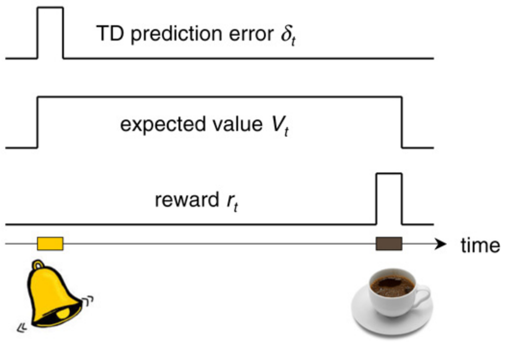


Fig. 5. Time course of different 'hidden variables' of interest in the TD model. The bell predicts a rewarding cup of coffee some time later. At the time of the cue, the phasic prediction error $\delta_t = r_t + V_{t+1} - V_t$ equals the magnitude of the predicted reward V_{t+1} (assuming here, for simplicity, $\gamma = 1$). The expected value signal corresponding to V_t also becomes positive at this time, and stays elevated until the time of the expected reward. At the time of the reward, a phasic cue might signal the occurrence of the reward, but no prediction error occurs if the reward was predicted. Figure adapted with permission from Niv and Schoenbaum (2008).

Note, also, that most fMRI analyses consist of searching for areas in the brain where the measured BOLD is correlated with some signal of interest. In particular, the assumption is that multiple signals in one brain area may be linearly multiplexed, and one can uncover the component signals via linear regression. As a result, it is not easy to distinguish between different correlated components of RL models, for instance, prediction errors and state values, especially at the time of the predictive cue (Fig. 5). This is because the prediction error at the time of the cue is $\delta_t = V(cue) - V(baseline)$, which, is obviously linearly related to $V(cue)$. Indeed, many studies have implicated the striatum in representing the anticipated value of outcomes (e.g. Delgado, Locke, Stenger, and Fiez (2003), Knutson et al. (2001) and Knutson, Fong, Bennett, Adams, and Hommer (2003)), and it is not always clear whether the measured activation is distinct from that attributable to a prediction error. In any case, electrophysiological data show that the striatum is definitely a viable candidate for representing state values (e.g. Samejima, Ueda, Doya, and Kimura (2005) and Schultz, Apicella, Scarnati, and Ljungberg (1992)).

Studies involving both gains and losses have further implicated the striatum in the anticipation of losses and not only gains, with decreases in BOLD signals correlated with the anticipated loss. Moreover, the degree of deactivation to losses compared to activation to gains ('neural loss aversion') in the nucleus accumbens and the prefrontal cortex was predictive of individual differences in behavioral loss aversion (Tom, Fox, Trepel, & Poldrack, 2007). Finally, outcome values themselves (as well as subjective preferences) have been associated with activations in areas such as the ventromedial prefrontal cortex and the orbitofrontal cortex (e.g. Knutson, Fong, Adams, Varner, and

Hommer (2001), Knutson et al. (2003), McClure et al. (2004) and O'Doherty, Deichmann, Critchley, and Dolan (2002)). These activations as well need to be convincingly dissociated from other potentially correlated signals such as TD errors.

While outlining the contribution of fMRI to elucidating the neural underpinnings of RL, it is clear from the above that fMRI results can reveal only so much. One way to overcome the different interpretational caveats is to synthesize results from electrophysiological recordings with those from fMRI. Another approach that is gaining popularity is the use of functional imaging in combination with pharmacological challenges (e.g. Knutson and Gibbs (2007) and Pessiglione et al. (2006)) or with radioligand labeled positron emission tomography (e.g. Zald et al. (2004)) to test in humans more directly the *causal predictions* and *pharmacological hypotheses* of RL models, respectively. In any case, the promise of model-driven analysis of imaging data has yet to be fully realized, and the link between computational models of learning and the brain does not end with the identification of the reward prediction error signal. For example, recent work has used such a 'hidden-variable' analysis within an RL framework to investigate the neural substrates of exploration (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and a hierarchical RL model has been used to demonstrate that the brain tracks the volatility (or rate of change) of the environment (Behrens, Woolrich, Walton, & Rushworth, 2007).

2.3. *Evidence for multiple reinforcement learning systems in the brain*

While dopamine is critical to many learning and behavioral processes in the brain, animals can learn to select actions correctly even in the absence of dopamine (Berridge, 2005, 2007). This is perhaps not so surprising, as converging lines of evidence suggest that animals and humans have a number of parallel decision-making systems at their disposal (Daw et al., 2005; Dickinson & Balleine, 2002), only a subset of which are dopamine dependent. Key to identifying these different systems is the fact that a certain behavior (for instance, simple lever-pressing by a rat) can have wholly different characteristics in different situations: early in training this behavior can show flexibility and sensitivity to changes in the task or in the animal's motivational state, while the same lever-pressing behavior can become inflexible and slow to adapt to any change after considerable training (e.g. Adams and Dickinson (1981), Adams (1982) and Killcross and Coutureau (2003)). The amount of training is not the only determinant of the degree of flexibility of learned behavior (e.g. Balleine, Garner, Gonzalez, and Dickinson (1995) and Dickinson, Nicholas, and Adams (1983)), but the link between over-training and the folk-psychological notions of "habits" has bestowed upon the inflexible form of behavior the name "habitual responding", while the more flexible instrumental actions are called "goal-directed" (Balleine, 2005; Balleine & Dickinson, 1998; Dickinson, 1985; Dickinson & Balleine, 2002). To complicate matters further, some forms of behavior seem wholly inflexible in that they are innately specified reactions to learned predictions (Dayan, Niv, Seymour, & Daw, 2006). These fall into the class of Pavlovian responses, and it is not clear whether they are also driven by two types of value predictions: flexible outcome-specific predictions (similar to those underlying goal-directed instrumental behavior), and less-flexible general affective predictions (as in habitual instrumental behavior).

This suggested multiplicity of neural controllers may be surprising – why not use the best controllers at all times? Careful consideration of the characteristics of real-world learning situations, together with the properties of different RL strategies, can offer insight into the advantages of combining different controllers as well as explain their different behavioral characteristics. Recall that the exposition of RL above began with a definition of predictive values (as the expected sum of future rewards), and then suggested different ways to learn or estimate such values. If different methods are available, perhaps the brain uses more than one? Daw et al. (2005) have suggested that since each method has different advantages and disadvantages, the brain should use each method in the circumstances for which it is best. They suggested to identify habitual action selection with behavior based on cached values – those values learned through prediction errors and slow trial-and-error experience with the environment. These are the classic "model-free" RL methods, for which action selection is easy, however, much training is needed in order to obtain accurate value estimates. Goal-directed behavior, on the other hand, was identified with online dynamic-programming-like computation of values through forward search or forward simulation of the consequences of actions using an estimated model of the environment (ie, estimated probabilities of transitions between states and reward probabilities at each state). These "model-based" methods are more accurate and adjust flexibly to changes in circumstances, however, their computation is costly in terms of neural resources and time, and so they should be used only sparingly – perhaps only in those situations in which there is not yet enough data to inform the "model-free" system. Daw et al. (2005) further postulated that the brain arbitrates between these parallel systems based on the uncertainty associated with their evaluations: when the two systems 'recommend' different courses of action, the recommendation that is most accurate is the one that should be followed. Assessing the accuracy of the evaluations of the two systems normatively depends on variables such as amount of training (which decreases uncertainty in the habitual system) and depth of the necessary forward search (which increases uncertainty in the goal-directed system).

These theoretical considerations align surprisingly well with both behavioral and neural data. Behaviorally, the circumstances that favor goal-directed behavior are those in which there is not yet sufficient experience for the model-free system to build on, such as early in training or when there are several actions leading to several outcomes (each of which has been sampled to a lesser extent). To the contrary, when a rather long sequence of possible events needs to be mentally simulated in order to evaluate a course of action in the model-based system, behavior tends to be habitual, determined by model-free evaluations instead. Other conditions favoring habitual responding are excessive training and simple scenarios in which only one action needs be evaluated. Even the fact that behavior on some schedules of reinforcement habitizes more readily than on other schedules (namely, interval and ratio schedules, respectively) can be understood within this framework: in interval schedules many behavioral policies lead to similar rates of reward, and thus policy learning can enjoy a large degree of generalization. In contrast, in ratio schedules different policies lead to different amounts of reward, and learning about one policy cannot generalize to another. This means that, given the same number of training sessions, the effective amount of learning experience per policy is smaller in a ratio schedule as compared to an interval schedule, and thus the uncertainty associated with the habitual system would be higher and behavior would remain goal-directed in ratio schedules.

Neurally, work from the lab of Bernard Balleine has implicated separate cortico-basal-ganglia loops (Joel & Weiner, 1994; Parent & Hazrati, 1993) in each of these evaluation and decision making systems (see Balleine (2005) for a review). Specifically, the so-called 'limbic loop', including areas such as the ventral striatum, the basolateral amygdala and the orbitofrontal cortex, has been associated with Pavlovian prediction learning and evaluation (Cardinal, Parkinson, Hall, & Everitt, 2002; Holland & Gallagher, 1999; Killcross & Blundell, 2002). The acquisition and expression of 'action-outcome' (forward model) associations in the goal-directed instrumental system has been localized to the 'associative loop' which includes the dorsolateral prefrontal cortex (or its homologue in rats, the prelimbic cortex) and the caudate nucleus (dorsomedial striatum in rats) (Balleine & Dickinson, 1998; Killcross & Coutureau, 2003; Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). Finally, 'stimulus-response' (cached policy) habitual behavior, which had previously been associated with the striatum in general, has recently been more specifically localized to the 'sensorimotor loop' originating in sensorimotor cortices, and involving the putamen (dorsolateral striatum in rats) (Yin, Knowlton, & Balleine, 2004).

The interactions between these interconnected loops (Joel & Weiner, 1994), and indeed the implementation of arbitration between the different systems, are less well understood. One candidate area for arbitration between goal-directed and habitual control is the rat infralimbic cortex (Killcross & Coutureau, 2003), but the story here is only beginning to unfold. Another avenue for future research, the flexible model-based action selection system is, as of yet, under-constrained both computationally and behaviorally. This complex system may be easier to study in humans, where the use of instructions can prevent the need for extensive training (and thus habitization of responding). Indeed, recent fMRI investigations in the lab of John O'Doherty have begun to see fruit from such a program. In one study, human goal-directed behavior was tested using the outcome devaluation protocol that has been developed for such studies in rats (Valentin, Dickinson, & O'Doherty, 2007). The neural results implicated the orbitofrontal cortex in the flexible evaluation of expected outcomes that is at the heart of goal-directed behavior. Another set of papers in which model-free TD learning was contrasted with model-based learning algorithms that exploit the higher order structure of a serial-reversal task, implicates the ventromedial prefrontal cortex and the anterior cingulate cortex in computations that are explicitly based on task structure (Hampton, Bossaerts, & O'Doherty, 2006; Hampton & O'Doherty, 2007).

2.4. *Tonic dopamine and the choice of response vigor*

Keen-eyed readers may have noticed that there is one aspect critically missing in our various accounts of reinforcement learning and decision-making in the brain. Indeed, it is something of a curiosity that although the tradition in animal experimentation is to investigate the determinants of *rates* of responding (as in Skinner's investigations of key-pecking in pigeons or leverpressing in rats, so called 'free-operant' experiments because the animal is free to choose when to respond and no trial structure is imposed on behavior), RL models of conditioning have concentrated exclusively on discrete choices of actions at pre-specified timepoints. Our issue is not only with laboratory paradigms: real-life decisions most often take place in continuous time, and one could argue that every choice of action, even that in a discrete trial setting, is accompanied by a choice of the *speed* or *vigor* with which that action will be performed. Such a decision gives rise to response rates in free operant behavior, to running times in mazes, and to reaction time data in discrete settings. It also interacts with the effects of motivation on behavior — a hungry rat running down a maze in search of food will run faster than a sated rat.

In this final subsection we will briefly discuss the application of RL methods to decisions about how fast (or with what vigor) to behave, and the neural implications of such a model. Despite the emphasis on discrete actions, RL theory does exist that deals with continuous time: this is *average reward* reinforcement learning in a *semi*-Markov decision process (Daw & Touretzky, 2002; Doya, 2000; Schwartz, 1993). Building on this theoretical framework, we have recently proposed an RL model of optimal rates of responding (Niv, Daw, & Dayan, 2005). In this model of instrumental conditioning, every choice of action is accompanied by a choice of a *latency* with which to perform that action. Furthermore, responding incur costs inversely proportional to the chosen latency, such that vigor is costly. Finally, the goal of the decision-maker is to obtain the highest possible net rate of rewards minus costs per unit time. The results of the model showed that the fundamental characteristics of free operant response rates could be explained as the consequences of the optimal choice of response rates in different tasks (Niv, 2007). Moreover, the model was used to derive a normative explanation for how motivational states should affect the rates of responding (Niv, Joel, & Dayan, 2006).

Importantly, the average reward RL framework highlights an important factor that determines optimal responding: the *net rate of rewards*, that acts as the *opportunity cost* of time. Consider, for instance, a rat pressing a lever in order to obtain food. Suppose that its presses had previously earned food at an average rate of four pellets per minute. When contemplating whether to devote five seconds to executing the next leverpress, the potential benefit of this action (ie, the probability of its generating reward, and the magnitude of this reward) should be weighed against both the costs of performing the action at this speed, and the opportunity cost of time, ie, the potential loss of (on average) 1/3 reward pellet due to devoting time to this action rather than continuing to behave according to the previous policy. Because the opportunity cost of time is similar for *all* actions, the model predicts that when the net reward rate is higher (for instance, due to a benevolent experimenter, or due to the rat being hungry, which renders each food pellet subjectively more valuable), all actions should optimally be performed faster (Niv, Daw, Joel, & Dayan, 2007; Niv et al., 2006).

How does this relate to neural reinforcement learning? Alongside the emphasis on only the discrete choice aspect of decision-making, the prediction error theory of dopamine also concentrated on only one aspect of dopaminergic activity and influence: the effect of *phasic* dopaminergic signaling on learning and plasticity. However, dopamine neurons operate in both a phasic and a tonic mode (Bergstrom & Garris, 2003; Floresco et al., 2003; Goto & Grace, 2005; Grace, 1991; Weiner & Joel, 2002), and affect not only synaptic plasticity, but also membrane potentials and neural excitability (which may be particularly sensitive to tonic levels of dopamine; Nicola, Surmeier, and Malenka (2000) and Schultz (2002)). Indeed, the effects of dopaminergic manipulations such as lesions, antagonism or agonism, are first and foremost seen in the vigor of ongoing behavior, rather than in learning processes. For instance, 6-hydroxydopamine injections into the ventral striatum that kill dopaminergic neurons projecting to that area, profoundly reduce the rate of instrumental responding (for a review see Salamone and Correa (2002)). As a result, dopamine in the striatum has been linked to invigorating Pavlovian and instrumental responding (Ikemoto & Panksepp, 1999; Salamone & Correa, 2002).

Combining these lines of evidence, we have suggested that the net rate of reward, the critical determinant of response rates across the board, might be represented by *tonic levels of dopamine* in the striatum (Niv et al., 2005, 2006). Different from phasic dopaminergic activity that changes on the timescale of milliseconds, and presumably exerts its main effect inside its target synapses, the tonic level of dopamine is the slowly-changing background level of the neuromodulator in the extrasynaptic fluid, hypothesized to change very slowly (on the order of minutes), bridging across events such as trials in an experiment. If tonic dopamine really does convey the net reward rate, it is now clear why higher levels of dopamine (for instance, as a result of amphetamine administration) would result in overall faster responding, and why dopamine depletion (as in Parkinson's disease) would induce lethargy. Recent work supporting this hypothesis has shown that the overall vigor of instrumental responding depends on the balance between the so-called direct and indirect pathways from the striatum to the output of the basal ganglia (Lobo, Cui, Ostlund, Balleine, & Yang, 2007), and that the lateral habenula controls the tonic levels of dopamine in the striatum, with manipulations exerting prolonged effects (more than one hour long) on the degree of locomotion of rats (Lecourtier, Defrancesco, & Moghaddam, 2008). Moreover, It conveniently turns out that if the tonic level of dopamine simply reflects spillover from phasic prediction error signals, averaged over a longer timeframe due to slow reuptake, it follows computationally that it would, by default, equal the net rate of obtained rewards. This 'tonic dopamine hypothesis' thus dovetails neatly both with the computational prediction error theory of phasic dopamine and with psychological theories about dopamine's role in energizing responses. It provides the first normative explanation for the critical role that tonic levels of dopamine play in determining the vigor of responding, and suggests a route by which dopamine could mediate the effects of motivation on response vigor.

3. Challenges and future directions

RL models are now used routinely to design and interpret a wide range of learning and decision-making experiments. However, one of the reasons that RL models have been successful is that they are highly simplified models, accounting for fundamental phenomena while eschewing the necessary complexities that accompany more detailed explanations. We have already discussed some possible extensions and fertile areas for future work throughout this review. In this last section, we highlight a few more theoretical challenges that await this area of active research.

The first challenge is hinted to by responses of dopamine neurons to stimuli not clearly related to reward prediction. It has long been known that novel stimuli cause phasic bursts in dopamine neurons (Schultz, 1998), although they are not (yet) predictive of any outcome, aversive or appetitive. However, new learning is not done on the background of a blank slate. It is reasonable to think that generalization to previously encountered stimuli would play a role in the initial appraisal of a novel stimulus. If the experimental (or the general ecological) scenario is such that animals have learned to expect that stimuli predict rewards (as is the case in many experimental situations), it is not surprising that new stimuli will be treated optimistically. Kakade and Dayan (2002) addressed this possibility directly, and furthermore suggested that the novelty responses can function as 'novelty bonuses' — quantities that are added to other available rewards ($r_t^{new} = r_t + novelty(S_t)$) and enhance exploration of novel stimuli. Kakade and Dayan showed how this simple idea can account in detail for the reported novelty responses of dopamine neurons (for instance, for the observation that the novelty burst is frequently followed immediately by a dip of the firing rate below baseline) yet still explain how they also communicate a reward prediction error. Recent fMRI work has demonstrated directly the existence of such additive novelty bonuses and their influence on choice behavior, in a situation in which novelty was explicitly not related to reward predictions (Wittmann, Daw, Seymour, & Dolan, 2008). The general issue which this line of work only begins to touch upon, is that of generalization: how does learning from one task affect subsequent learning. This fundamental question is still, for the most part, awaiting a normative computational account.

A second intriguing avenue of research deals with more complex tasks, for instance, those which have hierarchical structure. A quintessential example of this is the everyday task of making coffee, which comprises several high-level 'modules' such as 'grind beans', 'pour water', 'add sugar', each of which, in turn, comprises many lower-level motor actions. Hierarchical reinforcement learning (Barto & Mahadevan, 2003; Parr & Russell, 1998; Sutton, Precup, & Singh, 1999) is an active area of research exploring the ways in which RL systems can take advantage of the hierarchical structure of real-world tasks, in order to mix-and-match previously learned action modules. One question that frequently arises in these theoretical studies is, where do these previously learned modules come from, or, more importantly, how does an agent learn useful modules (e.g. Simsek and Barto (2004) and Stolle and Precup (2002)). In animal learning, at least part of this answer is clear: useful modules come from previous tasks. This again raises the issue of generalization — how to transfer learning from one task to another effectively. But, above and beyond the issue of learning of modules, the hierarchical RL framework raises many tantalizing questions regarding the neural implementation of hierarchical control (Botvinick, 2008; Botvinick, Niv, & Barto, 2008).

A third challenge is due to the nature of temporal difference learning, and specifically, its strong dependence on temporal representations. Behavioral results suggest that interval timing is extremely inaccurate, with the standard deviation of the prediction of an interval being proportional to the mean length of the interval (Gallistel & Gibbon, 2000; Gibbon, 1977). Simple variants of TD learning are extremely sensitive to timing noise, with even very small amounts of noise devastating the predictive power of the model, and resulting in pronounced prediction errors at the time of reward delivery, even in thoroughly learned tasks. The puzzle is whether the high degree of noise in behavioral timing is consistent with the temporal sensitivity displayed by neural prediction error signals. In any case, deriving such precisely-timed predictions despite considerable timing noise likely necessitates a more complex account for timing within a semi-Markov framework (Daw, Courville, & Touretzky, 2002, 2006)

The final area of research that we would like to mention here, has actually been waiting in the sidelines all along. Even simple experimental paradigms such as extinction (in which a once predicted reward ceases to appear) and conditioned inhibition (in which a cue predicts that an otherwise expected reward will not occur) do not yet have satisfactory computational models. The RL framework is exposed for its excessive simplicity by basic behavioral phenomena such as spontaneous recovery from extinction, or the fact that the inhibitory value of a conditioned inhibitor does not extinguish when this cue is presented alone. Temporal difference models that treat extinction merely as the unlearning of predictive values, can not explain spontaneous recovery. Similarly, modeling conditioned inhibitors as having negative predictive value cannot explain why this value is maintained even when it is consistently paired with no reward ("0") rather than a negative outcome. Clearly, conditioning is more than learning and unlearning of additively combined values. One recent model that suggested that negative prediction errors result in the inference of a new state S and new learning about this state, explained how both the original conditioned values and the new extinction knowledge can co-exist (Redish, Jensen, Johnson, & Kurth-Nelson, 2007). Further modeling work awaits.

4. Conclusions

To summarize, computational models of learning have done much to advance our understanding of decision making in the last couple of decades. Temporal difference reinforcement learning models have suggested a framework for optimal online model-free learning, which can be used by animals and humans interacting with the environment in order to learn to predict events in the future and to choose actions such as to bring about those events that are more desirable. Investigations into the decision-making behavior of both animals and humans support the existence of such a mechanism, controlling at least some types of decision-making behavior. The prediction error hypothesis of dopamine has further linked these algorithmic ideas to possible underlying neural substrates, specifically, to learning and action selection in the basal ganglia modulated by phasic dopaminergic signals. Converging evidence from a wide variety of recording and imaging methods supports this hypothesis. Neural investigations of the underpinnings of RL, in turn, have highlighted some holes in the current theory (e.g. dopaminergic novelty responses), and have suggested extensions to the RL framework (e.g. combining different RL controllers within one agent).

It thus seems that reinforcement learning has been most powerful (and unfortunately for neuroscience, almost unique), in tying together Marr's (1982) three levels: computation, algorithm and implementation, into one coherent framework that is used not only for gleaning understanding, but also for shaping the next generation of experimental investigations. Whether the theory can be elaborated to account for results of future experimentation without losing its simplicity and elegance, or whether it is eventually abandoned and replaced by a newer generation of computational learning theories, reinforcement learning has already left its permanent mark on the study of decision making in the brain.

Acknowledgments

The author wishes to thank P. Read Montague and Paul Glimcher for comments and contribution to earlier versions of this paper, and Michael Todd for comments on a previous draft. Parts of this work were published as a chapter in "Neuroeconomics: Decision making and the brain" (P. W. Glimcher, C. Camerer, E. Fehr, and R. Poldrack, editors). This work was supported by the Human Frontiers Science Program.

References

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

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *33B*, 109–121.

Arbuthnott, G. W., Ingham, C. A., & Wickens, J. R. (2000). Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy*, *196*(Pt 4), 587–596.

Attwell, D., & Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends in Neuroscience*, *25*(12), 621–625.

Baird, L. C. (1993). Advantage updating. *Tech. rep. no. WL-TR-93-1146*. Dayton, OH: Wright-Patterson Air Force Base.

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In A. Prieditis, & S. Russell (Eds.), *Proceedings of the 12th international conference on machine learning* (pp. 30–37). San Mateo, CA: Morgan Kaufman.

Balleine, B. W. (2005). Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology and Behaviour*, *86*(5), 717–730.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419.

Balleine, B. W., Garner, C., Gonzalez, F., & Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(3), 203–217.

Barto, A. G. (1994). Reinforcement learning control. *Current Opinion in Neurobiology*, *4*(6), 888–893.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge: MIT Press.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Systems Journal*, *13*, 44–77.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 834–846.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1989). Sequential decision problems and neural networks. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 686–693). Cambridge, MA: MIT Press.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 539–602). Cambridge, MA: MIT Press.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*(3), 1428–1439.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

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

Bergstrom, B. P., & Garris, P. A. (2003). 'Passive stabilization' of striatal extracellular dopamine across the lesion spectrum encompassing the presymptomatic phase of Parkinson's disease: A voltammetric study in the 6-OHDA lesioned rat. *Journal of Neurochemistry*, *87*(5), 1224–1236.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*(8), 2793–2798.

Berridge, K. C. (2005). Espresso reward learning, hold the dopamine: Theoretical comment on robinson et al. (2005). *Behavioral Neuroscience*, *119*(1), 336–341.

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology (Berl)*, *191*(3), 391–431.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Review*, *28*, 309–369.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Sc.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2008). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313–323.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioural Reviews*, *26*(3), 321–352.

Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *Journal of Neuroscience*, *6*(3), 613–619.

Costa, R. M., Lin, S.-C., Sotnikova, T. D., Cyr, M., Gainetdinov, R. R., Caron, M. G., et al. (2006). Rapid alterations in corticostriatal ensemble coordination during acute dopamine-dependent motor dysfunction. *Neuron*, *52*(2), 359–369.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2002). Timing and partial observability in the dopamine system. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*: *Vol. 14*. Cambridge, MA: MIT Press.

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

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603–616.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*(11), 2567–2583.

Day, J. J., Roitman, M. F., Wightman, R. M., & Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, *10*(8), 1020–1028.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. MIT Press.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, (3 Suppl), 1218–1223.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153–1160.

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive Affect Behavioural Neuroscience*, *3*(1), 27–38.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Learning, motivation and emotion*: *Vol. 3* (pp. 497–533). New York: John Wiley & Sons.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of instrumental training contingency on susceptibility to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *35B*, 35–51.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

Floresco, S. B., West, A. R., Ash, B., Moore, H., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience*, *6*(9), 968–973.

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

Geisler, S., & Zahm, D. S. (2005). Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *The Journal of Comparative Neurology*, *490*(3), 270–294.

Gibbon, J. (1977). Scalar Expectancy Theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325.

Goto, Y., & Grace, A. A. (2005). Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nature Neuroscience*, *8*, 805–812.

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience*, *41*(1), 1–24.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367.

Hampton, A. N., & O'Doherty, J. P. (2007). Decoding the neural substrates of reward-related decision making with functional mri. *Proceedings of the National Academy of Sciences USA*, *104*(4), 1377–1382.

Holland, P. C., & Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences*, *3*(2), 65–73.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge: MIT Press.

Howard, R. A. (1960). *Dynamic Programming and Markov processes*. MIT Press.

Ikemoto, S., & Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews*, *31*, 6–41.

Jensen, J., Smith, A. J., Willeit, M., Crawley, A. P., Mikulis, D. J., Vitcu, I., & Kapur, S. (2007). Separate brain regions code for salience vs. valence during reward prediction in humans. *Human Brain Mapping*, *28*, 294–302.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*, 535–547.

Joel, D., & Weiner, I. (1994). The organization of the basal ganglia-thalamocortical cicuits: open interconnected rather than closed segregated. *Neuroscience*, *63*, 363–379.

Joel, D., & Weiner, I. (1999). Striatal contention scheduling and the split circuit scheme of basal ganglia-thalamocortical circuitry: From anatomy to behaviour. In R. Miller, & J. Wickens (Eds.), *Conceptual advances in brain research: Brain dynamics and the striatal complex* (pp. 209–236). Harwood Academic Publishers.

Kacelnik, A. (1997). Normative and descriptive models of decision making: Time discounting and risk sensitivity. In G. R. Bock, & G. Cardew (Eds.), *Characterizing human psychological adaptations: Ciba Foundation symposium 208* (pp. 51–70). Chichester: Wiley.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4-6), 549–559.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell, & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 242–259). New York: Appleton-Century-Crofts.

Kehoe, E. J. (1977). Effects of serial compound stimuli on stimulus selection in classical conditioning of the rabbit nictitating membrane response. *Unpublished doctoral dissertation*, University of Iowa.

Killcross, S., & Blundell, P. (2002). Associative representationsof emotionally significant outcomes. In S. Moore, & M. Oaksford (Eds.), *Emotional cognition. from brain to behaviour*: *Vol. 44* (pp. 35–73). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal conrtex of rats. *Cerebral Cortex*, *13*, 400–408.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16), RC159.

Knutson, B., Delgado, M. R., & Philips, P. E. M. (2008). Representation of subjective value in the striatum. In P. W. Glimcher, C. Camerer, E. Fehr, & R. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain*. New York, NY: Academic Press.

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., & Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport*, *12*(17), 3683–3687.

Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fMRI. *Neuroimage*, *18*(2), 263–272.

Knutson, B., & Gibbs, S. E. B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology (Berl)*, *191*(3), 813–822.

Kobayashi, Y., & Okada, K.-I. (2007). Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. *Annals of the New York Academy of Science*, *1104*, 310–323.

Konda, V. R., & Tsitsiklis, J. N. (2003). On actor-critic algorithms. *SIAM Journal of Control Optimization*, *42*(4), 1143–1166.

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. New York: Cambridge University Press.

Kremer, E. F. (1978). The Rescorla–Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *4*(1), 22–36.

Krimer, L. S., Muly, E. C., Williams, G. V., & Goldman-Rakic, P. S. (1998). Dopaminergic regulation of cerebral cortical microcirculation. *Nature Neuroscience*, *1*(4), 286–289.

Lecourtier, L., Defrancesco, A., & Moghaddam, B. (2008). Differential tonic influence of lateral habenula on prefrontal cortex and nucleus accumbens dopamine release. *European Journal of Neuroscience*, *27*(7), 1755–1762.

Li, J., McClure, S. M., King-Casas, B., & Montague, P. R. (2006). Policy adjustment in a dynamic economic game. *PLoS ONE*, *1*(1), e103.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*(1), 145–163.

Lobo, M. K., Cui, Y., Ostlund, S. B., Balleine, B. W., & Yang, X. W. (2007). Genetic control of instrumental conditioning by striatopallidal neuron-specific S1P receptor Gpr6. *Nature Neuroscience*, *10*(11), 1395–1397.

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *Journal of Neuroscience*, *23*(10), 3963–3971.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the bold signal. *Annual Review of Physiology*, *66*, 735–769.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*, 1111–1115.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends in Neuroscience*, *26*(8), 423–428.

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., & Montague, P. R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron*, *44*(2), 379–387.

Menon, M., Jensen, J., Vitcu, I., Graff-Guerrero, A., Crawley, A., Smith, M. A., et al. (2007). Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: Effects of dopaminergic modulation. *Biological Psychiatry*, *62*(7), 765–772.

Miller, R., & Wickens, J. R. (1991). Corticostriatal cell assemblies in selective attention and in representation of predictable and controllable events. *Concepts in Neuroscience*, *2*(1), 65–95.

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, *379*, 449–451.

Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., & Sejnowski, T. J. (1993). Using aperiodic reinforcement for directed self-organization. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Advances in neural information processing systems*: *Vol. 5* (pp. 969–976). San Mateo, CA: Morgan Kaufmann.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, *377*, 725–728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1994). Foraging in an uncertain environments using predictive hebbian learning. In Tesauro, & J. D. Cowan (Eds.), *Advances in neural information processing systems*: *Vol. 6* (pp. 598–605).

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947.

Montague, P. R., McClure, S. M., Baldwin, P. R., Phillips, P. E. M., Budygin, E. A., Stuber, G. D., et al. (2004). Dynamic gain control of dopamine delivery in freely moving animals. *Journal of Neuroscience*, *24*(7), 1754–1759.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, *43*(1), 133–143.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057–1063.

Nakamura, K., Matsumoto, M., & Hikosaka, O. (2008). Reward-dependent modulation of neuronal activity in the primate dorsal raphe nucleus. *Journal of Neuroscience*, *28*(20), 5331–5343.

Nicola, S. M., Surmeier, J., & Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual Reviews in Neuroscience*, *23*, 185–215.

Niv, Y. (2007). The effects of motivation on habitual instrumental behavior. *Unpublished doctoral dissertation*, The Hebrew University of Jerusalem.

Niv, Y., Daw, N. D., & Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems: Vol. 18* (pp. 1019–1026). MIT Press.

Niv, Y., Daw, N. D., & Dayan, P. (2006). Choice values. *Nature Neuroscience*, *9*(8), 987–988.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, *191*(3), 507–520.

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*, 6.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, *10*(8), 375–381.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*(7), 265–272.

Nomoto, K., Watanabe, T., & Sakagami, M. (2007). Dopamine responses to complex reward-predicting stimuli. In *Society for Neuroscience abstracts*: *Vol. 33* (p. 749.5).

O'Doherty, J., Dayan, P., Friston, K., Critchley, H., & Dolan, R. (2003). Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron*, *38*, 329–337.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, *33*(5), 815–826.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98.

Parent, A., & Hazrati, L. N. (1993). Anatomical aspects of information processing in primate basal ganglia. *Trends in Neurosciences*, *16*(3), 111–116.

Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems: Vol. 10*. MIT Press.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*(3), 381–390.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neurosciences*, *22*(4), 146–151.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Rescorla, R. A. (1970). Reduction in effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation*, *1*, 372–381.

Rescorla, R. A., & Lolordo, V. M. (1968). Inhibition of avoidance behavior. *Journal of Comparative and Physiological Psychology*, *59*, 406–412.

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

Reynolds, G. S. (1961). Attention in the pigeon. *Journal of the Experimental Analysis of Behavior*, *4*, 203–208.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, *10*(12), 1615–1624.

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-intiated arm movements. *The Journal of Neurophysiology*, *63*, 592–606.

Salamone, J. D., & Correa, M. (2002). Motivational views of reinforcement: Implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behavioural Brain Research*, *137*, 3–25.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*, 210–229.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860–12867.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*(3), 900–913.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*(12), 4595–4610.

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

Schwartz, A. (1993). Thinking locally to act globally: A novel approach to reinforcement learning. In *Proceedings of the fifteenth annual conference of the cognitive science society* (pp. 906–911). Hillsdale, NJ: Lawrence Erlbaum Associates.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., et al. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664–667.

Simsek, O., & Barto, A. G. (2004). Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *21st international conference on machine learning*.

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *Journal of General Psychology*, *12*, 66–77.

Stolle, M., & Precup, D. (2002). Learning options in reinforcement learning. *Lecture Notes in Computer Science*, *2371*, 212–223.

Sutton, R. S. (1978). A unified theory of expectation in classical and instrumental conditioning. *Unpublished bachelors thesis*.

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

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

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

Sutton, R. S., Mcallester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems 12* (pp. 1057–1063). MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*, *92*, 2520–2529.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, *23*(32), 10402–10410.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–1645.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518.

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, *303*(5666), 2040–2042.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*(15), 4019–4026.

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

Walton, M. E., Gan, J. O., Barnes, S. D., Evans, S. B., & Phillips, P. E. M. (2006). Subsecond dopamine release in cost-benefit decision making. In *Society for Neuroscience abstracts*: *Vol. 32* (p. 71.2).

Watkins, C. J. C. H. (1989). Learning with delayed rewards. *Unpublished doctoral dissertation*, Cambridge University, Cambridge, UK.

Weiner, I., & Joel, D. (2002). Dopamine in schizophrenia: Dysfunctional information processing in basal ganglia-thalamocortical split circuits. In G. D. Chiara (Ed.), *Handbook of experimental pharmacology vol. 154/II, dopamine in the CNS II* (pp. 417–472). Berlin: Springer-Verlag.

Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook*, *22*, 25–38.

Wickens, J. R., Begg, A. J., & Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high frequency stimulation of cortex in vitro. *Neuroscience*, *70*(1), 1–5.

Wickens, J. R., & Kötter, R. (1995). Cellular models of reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 187–214). MIT Press.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256.

Wise, R. A., Spindler, J., de Wit, H., & Gerberg, G. J. (1978). Neuroleptic-induced "anhedonia" in rats: pimozide blocks reward quality of food. *Science*, *201*(4352), 262–264.

Wise, R. A., Spindler, J., & Legault, L. (1978). Major attenuation of food reward with performance-sparing doses of pimozide in the rat. *Canadian Journal of Psychology*, *32*, 77–85.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–973.

Yerkes, R. M., & Morgulis, S. (1909). The method of Pawlow in animal psychology. *Psychological Bulletin*, *6*, 257–273.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of the dosolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181–189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 505–512.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.

Zald, D. H., Boileau, I., El-Dearedy, W., Gunn, R., McGlone, F., Dichter, G. S., et al. (2004). Dopamine transmission in the human striatum during monetary reward tasks. *Journal of Neuroscience*, *24*(17), 4105–4112.