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 $\gamma \leq 1$ 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

$$V(S_t) = E[r_t|S_t] + \gamma E[r_{t+1}|S_t] + \gamma^2 E[r_{t+2}|S_t] + \cdots \quad (3)$$

$$= E[r_t|S_t] + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t) \left(E[r_{t+1}|S_{t+1}] + \gamma E[r_{t+2}|S_{t+1}] + \cdots \right) \quad (4)$$

$$= P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t) V(S_{t+1}) \quad (5)$$

(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*

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

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

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

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

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

(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

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

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

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

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

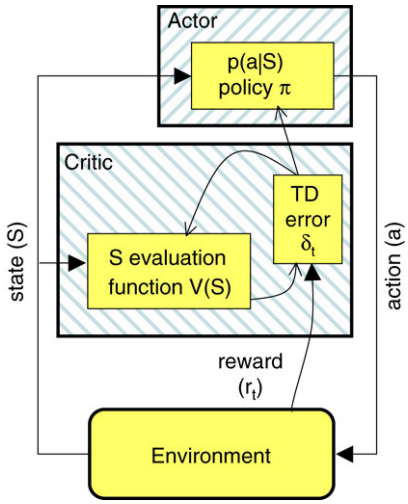

Fig. 1. Actor/Critic architecture: The state S_t and reinforcement signal r_t are conveyed to the Critic by the environment. The Critic then computes a temporal difference prediction error (Eq. (8)) based on these. The prediction error is used to train the state value predictions $V(S)$ in the Critic, 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).

(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

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

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

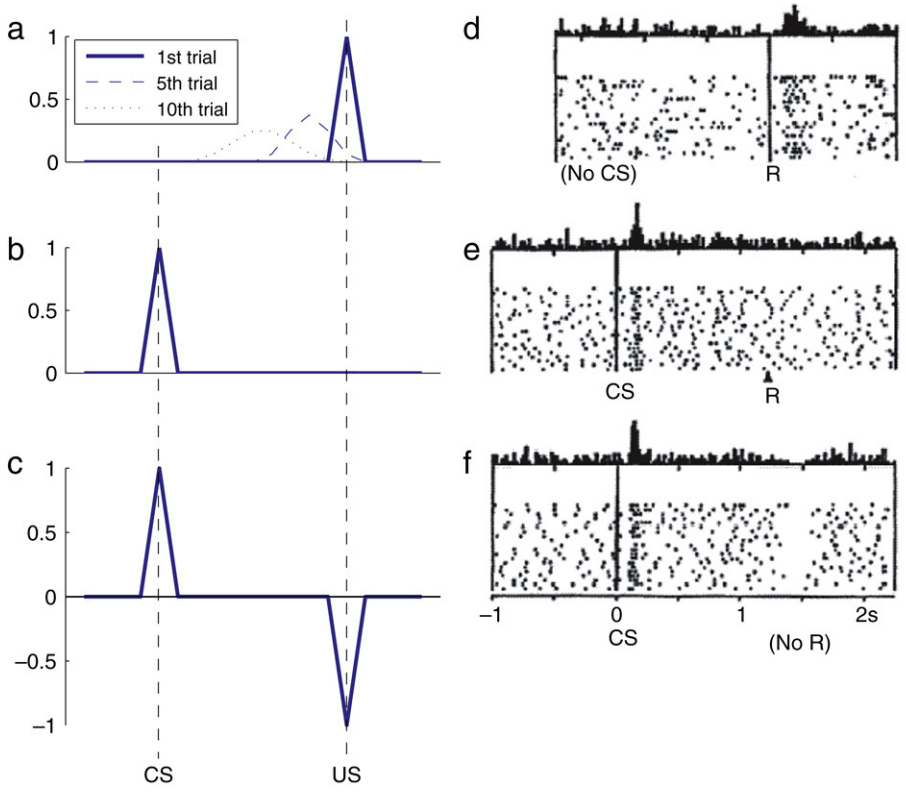

Fig. 2. (a–c) Temporal difference prediction errors in a Pavlovian conditioning task. A tone CS is presented at random times, followed 2 seconds later by a juice US. (a) In the beginning of training, the juice is not predicted, resulting in prediction errors at the time of the juice US. With learning, 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, 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