HFSP Journal

ISSN: 1955-2068 (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tfls19

Reinforcement learning: Computational theory and biological mechanisms

Kenji Doya

To cite this article: Kenji Doya (2007) Reinforcement learning: Computational theory and biological mechanisms, HFSP Journal, 1:1, 30-40, DOI: 10.2976/1.2732246/10.2976/1

To link to this article: https://doi.org/10.2976/1.2732246/10.2976/1

Reinforcement learning: Computational theory and biological mechanisms

Kenji Doya[1]

[1]Neural Computation Unit, Okinawa Institute of Science and Technology, 12-22 Suzaki, Uruma, Okinawa 904-2234, Japan

(Received 22 December 2006; accepted 29 March 2007; published online 8 May 2007)

Reinforcement learning **is a computational framework for an active *agent* to learn behaviors on the basis of a scalar *reward* signal. The agent can be an animal, a human, or an artificial system such as a robot or a computer program. The reward can be food, water, money, or whatever measure of the performance of the agent. The theory of reinforcement learning, which was developed in an artificial intelligence community with intuitions from animal learning theory, is now giving a coherent account on the function of the basal ganglia. It now serves as the "common language" in which biologists, engineers, and social scientists can exchange their problems and findings. This article reviews the basic theoretical framework of reinforcement learning and discusses its recent and future contributions toward the understanding of animal behaviors and human decision making. [DOI: 10.2976/1.2732246]**

CORRESPONDENCE
Kenji Doya: doya@oist.jp

What is the goal of our life? That is a difficult question to answer, but our life is a chain of actions to satisfy our needs or desires, many of which ultimately lead to our survival as individuals and proliferation as species. As a newborn or a novice sport player, our actions are initially random or awkward, but with repeated experience we become able to achieve the goals more efficiently and more reliably. Animal behavioral studies have described such processes of acquisition of goal-directed behaviors by the concepts of *reward* and *punishment*. A reward *reinforces* the action that causes its delivery (Thorndike, 1898). A punishment can be considered as a negative reward signal that reinforces an action that avoids its delivery. It is amazing, both in animal theaters and in the wild, how an animal can acquire a variety of complex behaviors by linking its actions to consequent positive and negative rewards. That gave a good motivation for artificial intelligence researchers to seek computer algorithms that allow a machine to acquire a variety of functions simply from reward signal, i.e., a scalar evaluation feedback.

The products of such studies are collectively called *reinforcement learning* algorithms and have been applied to a variety of control and optimization problems (Sutton and Barto, 1998). Since the late 1990s, neuroscientists became aware of interesting parallels between the key signals used in reinforcement learning algorithms and what they found in neural recording and brain imaging data. The collaborations of theoreticians and experimentalists contributed to a better understanding of the functions of, most notably, the basal ganglia and the neurotransmitter dopamine (Schultz *et al.*, 1997; Doya, 2000; Daw and Doya, 2006). The success has now interested psychiatrists, sociologists, and economists who are trying to understand how humans make good and bad decisions in the real world (Glimcher and Rustichini, 2004; Sanfey *et al.*, 2006).

In the rest of this article, I will review the basic concepts of reinforcement learning theory, how it has con-

HFSP Journal © HFSP Publishing $25.00
Vol. 1, No. 1, May 2007, 30–40 http://hfspj.aip.org/doi/10.2976/1.2732246/10.2976/1

tributed to the understanding of brain functions, and the future directions in which the fusion of reinforcement learning theory and other research disciplines may produce a big reward.

REINFORCEMENT LEARNING THEORY

The standard theory of reinforcement learning is defined for a *Markov decision process* (MDP), as shown in Fig. 1, where the action of an agent determines the state transition probability P(new state|state, action) and reward probability P(reward|state, action). In a simple case, the state is identical to the sensory input. The goal of reinforcement learning is to improve the agent's action probability P(action| state), or *policy*, to maximize total or average future rewards. More specifically, we consider the expected cumulative future reward

$$E[\text{reward}(t) + \gamma\text{reward}(t+1) + \gamma^2\text{reward}(t+2) + \cdots], \quad (1)$$

where $E[\ ]$ represents the expected value (average). The parameter γ specifies how far into the future the agent is concerned with [only immediate reward(t) for $\gamma=0$ and forever for $\gamma=1$] and is called the *temporal discount factor*. What makes reinforcement learning interesting (and difficult) is that an action(t) does not only affect the immediate reward(t), but also affects the next state($t+1$), which affects the availability of future reward($t+1$), reward($t+2$), and so on. Seen in the other direction, a given reward (t) may not be due to the immediately preceding action(t), but may also be due to the past action($t-1$), action($t-2$), …, i.e., all the past action sequence. The problem of identifying which past actions are responsible for a given reward is known as the *temporal credit assignment problem*, which is a major issue in reinforcement learning theory.

Actor-critic and temporal difference signal

In the early 1980s, Barto and Sutton came up with a reinforcement learning method, called *actor-critic*, that can deal with the temporal credit assignment problem (Barto *et al.*, 1983). The agent is composed of the actor that takes actions according to its policy and the critic that predicts the expected future reward and thereby helps the actor to improve its policy. More specifically, what the critic learns is the so-called *state value function*:

$$V(s) = E[\text{reward}(t) + \gamma\text{reward}(t+1) + \gamma^2\text{reward}(t+2) + \cdots|\text{state}(t) = s]. \quad (2)$$

The state value $V(s)$ predicts how much future reward the actor is going to get by following the current policy from the state s. How can this prediction be learned, and how can this prediction be useful? The virtue of taking exponential temporal discounting $(1, \gamma, \gamma^2, \ldots)$ is that we can estimate the value of the current state using a recursive relationship

$$V(\text{state}(t)) = E[\text{reward}(t) + \gamma V(\text{state}(t+1))].$$

This formula tells us that the cumulative reward expected from the current state(t) is the sum of the expected reward(t) acquired immediately and the expected reward from the next state($t+1$), with a discounting by γ. We can use any deviation from this relationship as the learning signal. The deviation is called the *temporal difference* (*TD*) *error*, or the *TD signal* (Barto *et al.*, 1983; Sutton, 1988)

$$\delta(t) = \text{reward}(t) + \gamma V(\text{state}(t+1)) - V(\text{state}(t)). \quad (3)$$

If the prediction of the critic is perfect, this should be zero in average. So the TD signal is used for learning the state value function of the critic, for example, by

$$V^{\text{new}}(\text{state}(t)) = V^{\text{old}}(\text{state}(t)) + \alpha_c\delta(t). \quad (4)$$

A marked feature of the actor-critic learning is that the same learning signal $\delta(t)$ is used also for the actor, despite being used in a slightly different way. Even if the prediction of the critic is the best possible, the TD error varies around zero if the environment or the policy is stochastic. Suppose $\delta(t)$ turns out to be positive; that means that the previous action resulted in more reward(t) or a higher-value state ($t+1$) than usually expected. Then it is appropriate to *reinforce* the action, i.e., to increase the probability of taking the same action again when faced in the same state. Thus the TD error serves as the effective reward signal which takes into account the prediction by the critic.

It took a while for the artificial intelligence and neural networks community to really appreciate the power of the framework, but with the interpretation of the TD algorithm as an on-line approximation of the dynamic programming (Watkins, 1989; Werbos, 1990) and the demonstration of its world-champion-class performance in backgammon (Tesauro, 1994), more and more researchers started to apply the TD learning framework to a variety of problems, such as robotic control, by extending the algorithms. Figure 2 is an example of an application of TD learning for a robot that learns to stand up.

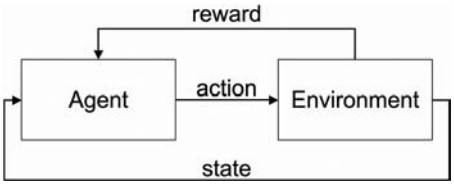

Figure 1. The problem setup of reinforcement learning. The agent observes the state of the environment and takes an action according to its policy P(action|state). The environment changes its state with probability P(new state|state, action) and gives a reward by P(reward|state, action) depending on the current state and the agent's action. The goal of the agent is to improve its policy so that it can get more rewards in the long run. Note that it is in general not optimal to maximize only the immediate reward(t) by action(t), as the next state($t+1$) can also affect the rewards acquired in the future.

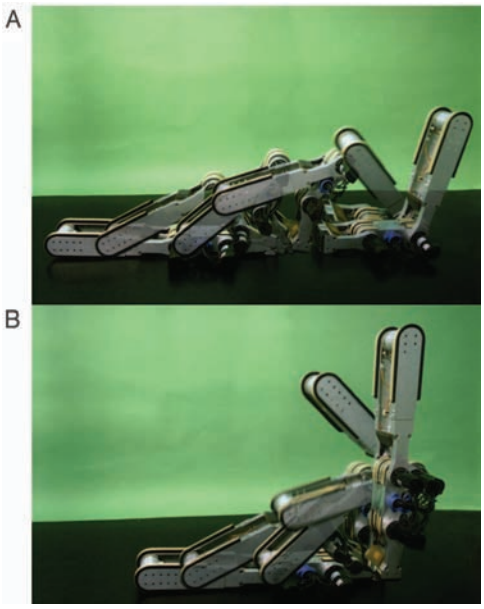

Figure 2. A robot that learns how to stand up by reinforcement learning (Morimoto and Doya, 2001). The robot is about 70 cm long and has two motors at its hip and knee joints. It observes its body state by joint angle sensors and a gyro sensor. A reward is given according to the height of its head and a punishment (negative reward) is given upon stumbling. The robot takes a hierarchical control scheme, and at its upper level, it learns an action value function that evaluates how good or bad the current state (joint angles and the position of its center of mass) is and which action (aimed angles of hip and knee) will lead to better or worse state. After hundreds of exploratory trials (A), first in computer simulation and then in real hardware, the robot learned a policy to allow it to reliably and smoothly stand up (B).

Q-learning and action values

Another popular reinforcement learning algorithm is called *Q*-learning (Watkins, 1989; Watkins and Dayan, 1992), in which the agent learns the so-called *action value function*

$$Q(s,a) = E[\text{reward}(t) + \gamma\text{reward}(t+1) + \gamma^2\text{reward}(t+2) + \cdots | \text{state}(t)=s, \text{action}(t)=a]. \quad (5)$$

The action value $Q(s,a)$ evaluates the goodness of an action at a given state, i.e., how much future reward the agent is going to get by taking an action a at the state s and then following the current policy. If the action value function has been learned for all the possible state-action pairs, the optimal policy is to select an action a that maximizes the action value $Q(s,a)$ for the given state s.

How can we learn the action value function? As in the actor critic, we can use a recursive relationship

$$Q(\text{state}(t),\text{action}(t)) = E[\text{reward}(t) + \gamma \max_a Q(\text{state}(t+1),a)].$$

This formula tells us that the goodness of an action taken now can be evaluated by the sum of the reward (t) acquired immediately and the discounted goodness of the best action to be taken at the subsequent state ($t+1$). This recursive relationship allows an agent to update its evaluation of action on line, without waiting indefinitely long for the rewards coming in the future. The TD signal for the action value function is given by

$$\delta(t) = \text{reward}(t) + \gamma \max_a Q(\text{state}(t+1),a) - Q(\text{state}(t),\text{action}(t)). \quad (6)$$

The action value for the experienced state-action pair is then updated by a learning rate α as

$$Q^{\text{new}}(\text{state}(t),\text{action}(t)) = Q^{\text{old}}(\text{state}(t),\text{action}(t)) + \alpha\delta(t), \quad (7)$$

while the action values for other state-action pairs are kept unchanged.

Figure 3 illustrates how the action value function can help an agent to find a behavior that maximizes the reward acquired in a long run (see the figure caption). In summary, a *Q*-learning agent repeats the following three steps: (1) pre-

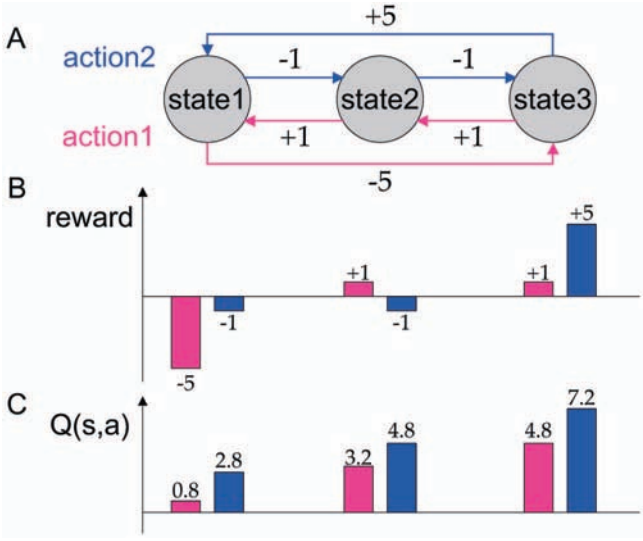

Figure 3. An example of a three-state Markov decision process. (A) Action 1 (red) causes state transition to the left with a small positive reward, but at the leftmost state 1, it cases a big negative reward. In contrast, action 2 (blue) causes state transition to the right with a small negative reward, but at the rightmost state 3, it causes a big positive reward. (B) The reward function reward (state, action). If the agent just learns from immediate rewards, it would take action 1 at state 2 for a small positive reward and then take action 2 at state 1 to avoid a larger loss, resulting in a cycle with no net gain. (C) The action value function Q(state, action) with the temporal discount factor γ=0.8. At state 2, the action value for action 2 is positive despite immediate negative reward because of the large positive reward expected from the next state 3. As the action value for action 2 is larger than that of action 1 at each of the three states, the agent would follow a three-step cycle with two small losses and a big gain.

dict the expected rewards from candidate actions, (2) select an action with the largest expected reward, and (3) update the prediction with the discrepancy to the outcome.

For more details of reinforcement learning algorithms and examples of their applications, please refer to the standard textbook by Sutton and Barto (1998).

REINFORCEMENT LEARNING IN THE BRAIN

Although the concept of reinforcement learning originates from animal learning, it is a question whether a reinforcement learning algorithm matches what is actually happening in the brain. A remarkable observation in the last decade was that there indeed seems to be a good parallel between the neurobiological processes in the brain and the computational steps of reinforcement learning algorithms.

Dopamine for the TD signal

The most notable is the findings by Schultz on the response properties of midbrain dopamine neurons (Schultz *et al.*, 1995; Schultz, 1998). His group recorded dopamine neuron activities while monkeys performed tasks like reaching for a food or pressing a lever for juice [Fig. 4(A)]. Although dopamine neurons initially responded to the rewards, when those rewards became fully predictable from preceding sensory cues, such as light and sound, their reward responses went away. Instead, dopamine neurons started to respond to reward-predictive sensory cues. If the reward is omitted after learning, dopamine neuron firing was suppressed at the timing when reward delivery is expected. These are interesting findings on their own, but most exciting for those who are familiar with reinforcement learning theory because it exactly matches what the TD error does.

In Eq. (3), if the cumulative predicted reward $V(\ )$ is zero for all states, the TD signal $\delta(t)$ is equal to the reward signal reward (t) [Fig. 4(B) top]. On the other hand, if a change of state (t) to state $(t+1)$ allows the critic to predict a positive future reward, the temporal difference $\gamma V(\text{state}(t+1)) - V(\text{state}(t))$ will make a positive pulse in the TD signal $\delta(t)$ even if reward(t) is not given yet. These responses exactly match the dopamine neuron's responses for an unpredicted reward and a reward predictive sensory cue. Furthermore, the TD signal $\delta(t)$ should stay zero when an already predicted reward is given, because actual receipt of the reward makes the future expectation $V(\text{state}(t+1))$ go down and the

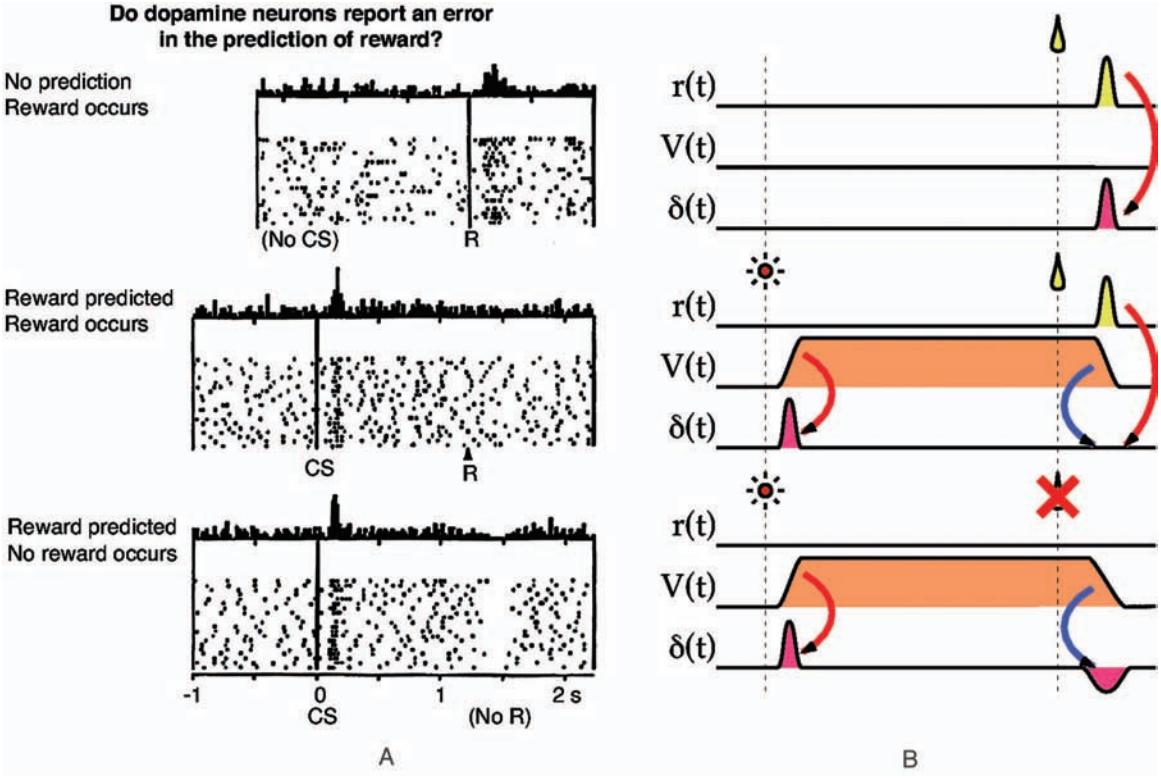

Figure 4. (A) Firing of dopamine neurons and its correspondence with the TD error. Top: before learning, or when a reward is delivered without a conditioned stimulus (CS) dopamine neurons respond to a reward (R) such as juice or food. Middle: after learning, dopamine neurons do not respond to rewards, but respond to a conditioned stimulus that allows the monkey to predict the reward. Bottom: when the predicted reward is omitted, firing of the dopamine neuron is suppressed. From Schultz *et al.* (1997). Reprinted with permission from AAAS. (B) The expected behaviors of the TD signal $\delta(t)=\text{reward}(t)+\gamma V(\text{state}(t+1))-V(\text{state}(t))$ in the three conditions, resembling the dopamine neuron response (see the text). Note the abbreviation $r(t)=\text{reward}(t)$ and $V(t)=V(\text{state}(t))$, and that we assumed $\gamma=1$ for simplicity (Doya, 2002).

negative pulse from the temporal difference of V should cancel the positive pulse for reward(t) [Fig. 4(B) middle]. If the predicted ted, only the negative pulse from temporal difference should appear in the TD signal [Fig. 4(B) bottom].

These and further parallels between the dopamine neuron activities and the TD signal (Waelti *et al.*, 2001; Satoh *et al.*, 2003; Nakahara *et al.*, 2004; Morris *et al.*, 2006) motivated proposals that the dopamine neurons and their major projection target, the striatum, may implement TD-type reinforcement learning (Barto, 1995; Houk *et al.*, 1995; Montague *et al.*, 1996; Schultz *et al.*, 1997).

Dopamine-dependent plasticity

Another important clue came from the study on the plasticity of the synapses from the cerebral cortex to the striatal neurons. A remarkable anatomical feature of the striatal projection neuron is that it has many synaptic spines and each of them receives glutamate input from the cortex and dopamine input from the midbrain dopamine neurons. Jeff Wickens and colleagues hypothesized that dopamine may control the learning of cortical input to the striatal neurons and tested it in experiments (Wickens *et al.*, 1996; Reynolds and Wickens, 2000; Reynolds *et al.*, 2001). In the famous Hebb's learning rule, a synapse is strengthened when a presynaptic input is followed by a postsynaptic neuron response, i.e., input times output. What Wickens and colleagues found was that the plasticity of cortico-striatal synapses is further weighted by the dopamine input: synaptic connection is potentiated when the presynaptic and postsynaptic activation is associated with increased dopamine input. On the other hand, when presynaptic and postsynaptic activation is not associated with dopamine input, the connection is depressed [Fig. 5(A)] (Reynolds *et al.*, 2001). Thus the same cortical input can result in potentiation or depression of the synapse depending on the level of dopamine input [Fig. 5(B)] (Reynolds and Wickens, 2002).

For what kind of learning can this three-term-rule, input times output times dopamine, be useful? Both in actor-critic and Q-learning mentioned above, the TD signal $\delta(t)$ is used for reinforcing the state-to-action mapping. Therefore, if the output of each striatal neuron encodes a particular action or action value and if the cortical input represents the current state, the three-term learning rule is exactly what we need in actor learning rule or action value learning [Eq. (7)].

Action value coding in the striatum

The striatum is a part of the basal ganglia circuit, which forms parallel loops from the cortex, through the striatum, the globus pallidus, and the thalamus, and back to the cortex. The hypothesis that the basal ganglia may implement TD learning motivated experimentalists to test the responses of striatal neurons to the changes in the predicted reward. In one of such early studies, Kawagoe *et al.* measured the activities of striatal neurons while a monkey made a saccade eye

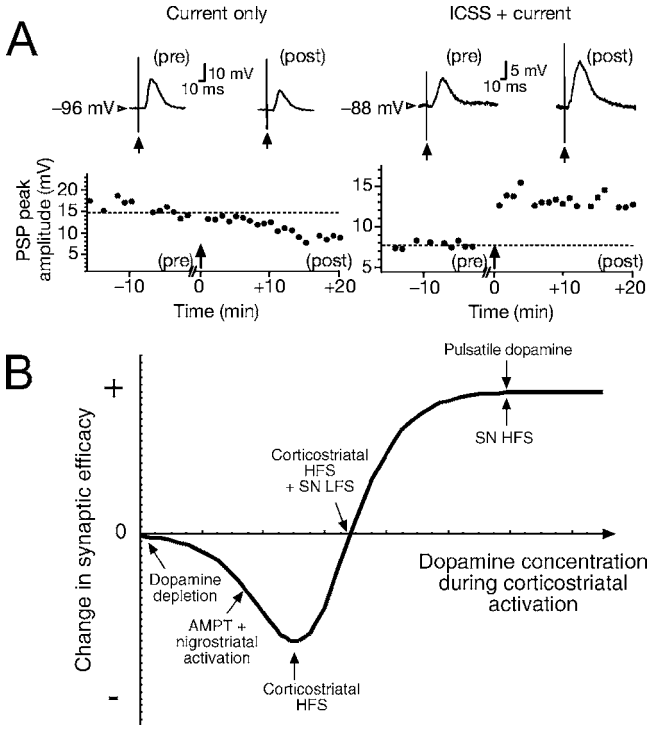

Figure 5. Dopamine-dependent plasticity of cortico-striatal synapses. (A) While a high-frequency stimulation of the cortical input to the striatum results in depression of the cortico-striatal synapses (left), simultaneous stimulation of cortical and dopamine fibers result in synaptic potentiation (right). Adapted by permission from Macmillan Publishers Ltd. (Reynolds *et al.*, 2001). (B) A summary diagram of dopamine-dependent plasticity of cortico-striatal synapses. Reprinted from Neural Networks (Reynolds and Wickens, 2002) with permission from Elsevier.

movement to one of four targets that lit up (Kawagoe *et al.*, 1998, 2004; Hikosaka *et al.*, 2006). In the "all-direction reward" condition, liquid reward was given after a saccade to any one of the lighted targets. In the "one-direction reward" condition, a large reward was given after a saccade to a lighted target in only one direction, e.g., the target to the left, and saccades to the other lighted target resulted in no or very small reward. This did not only affect the monkey's response (i.e., faster response to the rewarded target), but also the firing of striatal neurons before saccade [Fig. 6(A)]. In the all-direction reward condition, many of the striatal neurons had a prticular directional tuning (e.g., firing before a saccade to the right). On the other hand, in the one-direction reward condition, many neurons fired more strongly when the direction of the saccade being made was in the rewarded direction. This observation showed that the firing of the striatal neurons did not just represent the action being executed, but how much reward was expected after taking that action. This parallels the role of the action value function. Subsequently, Samejima *et al.* showed in a free choice task that many of the striatal neurons represent action-specific reward prediction [Fig. 6(B)] (Samejima *et al.*, 2005). These findings suggest

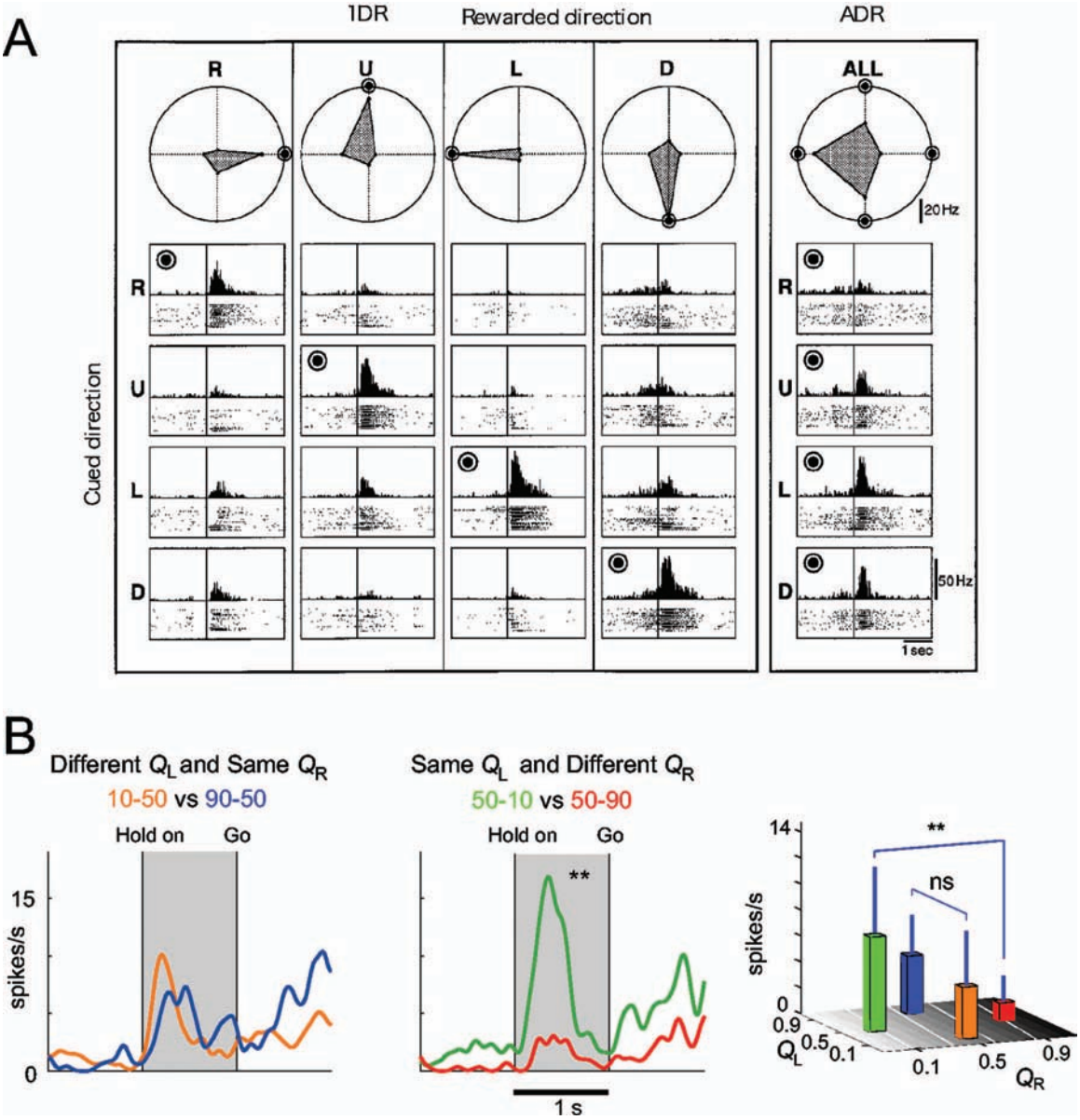

Figure 6. (**A**) **Reward-dependent firing of striatal neurons preceding actions.** A monkey makes a saccade to a target one second after it is lit. Many of the striatal neurons have direction tuning for the saccade target, but in experiment blocks where only the saccade to one of the four targets is rewarded, the direction tuning is modulated depending on whether the saccade is associated with a reward. Reprinted by permission from Macmillan Publishers Ltd. (Kawagoe *et al.*, 1998). (B) An example of action value-coding neurons in the stratum in stochastically rewarded free choice task (Samejima *et al.*, 2005). After the monkey turned a handle to the left or right, a liquid reward was given stochastically. The percentage of reward delivery for left and right actions were changed blockwise manner from four settings 10–50, 90–50, 50–10, 50–90. During one second before action initiation, the firing of this neuron changed according to the reward probability for turning the handle to the right, suggesting its coding of action value for the right handle turn. Note that the firing did not differ between 10–50 and 90–50 blocks, although the learned actions were the opposite in these blocks.

that the striatum plays a major role in linking the reward feedback to action selection by learning the action value function.

Basal ganglia as the reinforcement learning circuit

These above three findings, namely, TD signal represented by dopamine neurons, dopamine-dependent plasticity in the striatum, and action value represented by striatal neurons, totally changed our view on the function of the basal ganglia. The traditional view on the role of the basal ganglia was initiation of habitual behavioral repertoires. Now the basal ganglia are regarded as the major system for guiding behaviors by reward experience (reinforcement) and predicted rewards (motivation). Since the turn of the century, a remarkable

number of functional brain imaging experiments also found reward-related activities in the basal ganglia circuit.

There have been proposed a number of models on how TD learning is realized in the circuit of the basal ganglia (Houk *et al.*, 1995; Montague *et al.*, 1996; Suri and Schultz, 1998; Contreras-Vidal and Schultz, 1999; Doya, 1999, 2000; Daw and Doya, 2006). The common denominator of those models is that the TD error signal carried by dopamine is used for action learning. However, there have been different views on how the actor and the critic, or the action value and the state values are learned, how actions are selected, and how the TD error is calculated. Figure 7 is one of those models, proposing the action value coding in the striatum and action selection in the downstream of the basal ganglia (Doya, 2000).

It is worth noting that reinforcement learning is not the only paradigm studied in machine learning theory. One extreme is supervised learning, where the target output vector is given at each moment. Another is unsupervised learning where there is no explicit teaching signal. Reinforcement learning is between the two; learning from the scalar reward signal with possible delays. In fact there are many pieces of evidence suggesting that different brain structures are specialized for different learning paradigms, namely, the cerebellum for supervised learning, the basal ganglia for reinforcement learning, and the cerebral cortex for unsupervised learning (Houk and Wise, 1995; Doya, 1999, 2000).

It is also important to note that the basal ganglia are by no means the sole locus of reinforcement learning in the brain. Even the small brains of slugs or bees should have the basic capability for reinforcement learning. Lesion studies in rodents and other mammals showed that the amygdala and so-called "limbic system" are also involved in learning from reward and punishment (Balleine and Killcross, 2006). Reward-dependent activities are also found in a variety of cortical areas, such as the orbitofrontal cortex (Schultz *et al.*, 2000), the prefrontal cortex (Watanabe, 1996; Matsumoto *et al.*, 2003), and the parietal cortex (Platt and Glimcher, 1999; Dorris and Glimcher, 2004; Sugrue *et al.*, 2004). The differential roles of such distributed brain circuits in different types and stages of learning are the subject of active studies (Balleine and Killcross, 2006).

RECENT PROGRESS AND OPEN QUESTIONS

Reactive reinforcement or predictive planning

In the basic form of reinforcement learning, the agent just takes an action according to its policy and updates it after receiving a reward outcome in the form of a TD signal. However, when the goal of the agent changes, for example, with new fruits getting ripe in a tree or the agent's appetite satisfied, the appropriate action becomes different even if the dynamics of the environment stays the same. This is when learning of the environmental state transition rule, P(new state|state, action), as an *internal model* is useful (Doya, 1999; Kawato, 1999). If such an internal model is available, the agent can perform a simulation: if I take action(t) from current state(t), what new state($t+1$) will I end up in. If the reward for each state is also known, the agent can evaluate the goodness of a hypothetical action. By repeating this step many times, the agent can, in principle, estimate the goodness of a sequence of actions. This is a *tree-search* algorithm used in many classical artificial intelli-

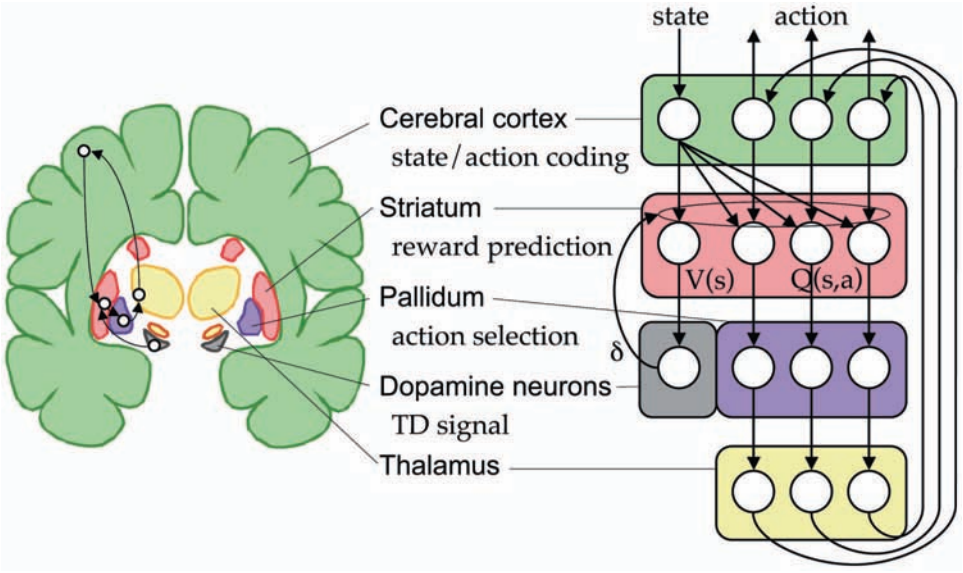

Figure 7. A schematic model of implementation of reinforcement learning in the cortico-basal ganglia circuit (Doya, 1999, 2000). Based on the state representation in the cortex, the striatum learns state and action value functions. The state value coding striatal neurons project to dopamine neurons, which sends the TD signal back to the striatum. The outputs of action value coding striatal neurons channel through the pallidum and the thalamus, where stochastic action selection may be realized.

gence programs, such as programs that play checker or chess. A recent model by Daw and colleagues proposes that a reactive, value-based algorithm and a predictive, tree-search algorithm can be switched on the basis of their relative reliability (Daw *et al.*, 2005). Their model replicates the results of "devaluation experiments" in which the value of the same food pellet changes by satiation.

How can such predictive planning be realized in the brain? A candidate is the network linking the parietal cortex, frontal cortex, and the striatum. The parietal, premotor, and prefrontal cortices are most frequently reported as the areas activated in imagery of body movement as well as abstract cognitive operations (Deiber *et al.*, 1998; Sawamura *et al.*, 2002; Hanakawa *et al.*, 2003). In combination with their interconnection with the cerebellum, those areas may store and update the predicted future states. The connections from those cortical areas to the striatum could be used for evaluation of the predicted states from hypothetical actions (Doya, 1999). A recently found shortcut pathway from the cerebellum to the striatum through thalamus (Hoshi *et al.*, 2005) may also be used for linking the internal models in the cerebellum and the value function in the striatum.

Bayesian inference in an uncertain environment

In reinforcement learning in realistic situations, the agent may not be certain about which "state" it is in. In such a case of partially observable states, the agent must estimate its current state, called the "belief state," by combining the current ambiguous observation with the prediction from the previous belief state and action using an internal model (Kaelbling *et al.*, 1998). The right method of integration is Bayesian inference and there is accumulating evidence suggesting that the cerebral cortex realizes such probabilistic inference (Knill and Pouget, 2004; Doya *et al.*, 2007).

In a recent functional brain imaging experiment, Yoshida and colleagues asked subjects to navigate in a computer maze with only limited vision (Yoshida and Ishii, 2006). A novel finding in their experiment was that while the medial prefrontal cortex was activated when there was a major inconsistency between the prediction and the sensory observation, the activity of the anterior prefrontal cortex was correlated with the uncertainty of the current state, showing the differential roles of subparts of the prefrontal cortex in working under uncertainty.

Parameter regulation and neuromodulators

Dopamine is one of the *neuromodulators*, a class of neurotransmitters that project diffusely from small nuclei in the brain stem to widespread areas of the brain and can have a complex, prolonged impact on recipient neurons (Marder and Thirumalai, 2002). Neuromodulators are supposed to broadcast scalar information. The notion that dopamine neurons are signaling the TD error prompt us to ask further questions: How are other parameters of reinforcement learning represented? What other neuromodulators are signaling? Doya proposed a set of hypotheses that serotonin regulates the temporal discount factor γ, noradrenaline regulates the noise parameter in stochastic action exploration, and acetylcholine regulates the learning rate α (Doya, 2002).

A classical notion about serotonin is that it is an opponent of dopamine (Daw *et al.*, 2002), encoding the prediction of future punishment. However, evidence is accumulating on the role of serotonin in controlling impulsive behaviors, including the inability to wait for delayed rewards (Winstanley *et al.*, 2006). Then how can serotonergic projection regulate the temporal discounting in the cortico-basal ganglia circuit? Tanaka and colleagues performed a functional brain imaging experiment using the Markov decision task shown in Fig. 3 (Tanaka *et al.*, 2004). Their analysis using a reinforcement learning model revealed a map of temporal discounting among the cortico-basal ganglia loops: while the ventral loop is specialized for prediction in a shorter time scale, the dorsal loop is specialized for prediction in a longer time scale (Fig. 8). We are now testing whether these parallel circuits are under differential regulation of serotonergic projection.

Regarding noradrenaline and acetylcholine, Yu and Dayan proposed a model in the context of Bayesian sensory discrimination (Yu and Dayan, 2005). Their proposal is that acetylcholine encodes expected uncertainty, e.g., when a subject starts learning in a novel environment, and that noradrenaline encodes unexpected uncertainty, e.g., when the learned environment suddenly changes. Their model successfully explained the results of pharmacological experiments in attention tasks.

Furthermore, what dopamine represents may not only be the TD signal. Responses of dopamine neurons to novel sensory stimuli have been reported (Redgrave and Gurney,

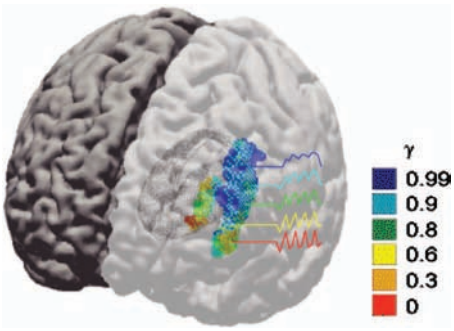

Figure 8. Parallel cortico-striatal pathways for reward prediction at different time scales (Tanaka *et al.*, 2004). Subjects learned the three-state Markov decision task as depicted in Fig. 3 in a MRI scanner. The time course of the state value function $V(\text{state}(t))$ of each subject was estimated by a reinforcement learning model using six different settings of the temporal discount factor γ (0, 0.3, 0.6, 0.8, 0.9, and 0.99). The voxels with significantly correlated activity with the time course of the value $V(\text{state}(t))$ or the TD signal $\delta(t)$ are shown using color codes for the discount factor γ. Ventro-dorsal maps of temporal discounting were found in the insular cortex with the value and in the striatum with the TD signal.

2006), which might be interpreted as the supplementary reward signal for promoting exploratory behaviors (Dayan and Sejnowski, 1996). Schultz and colleagues showed that tonic firing of dopamine neurons during the delay period before stochastic reward delivery is highest when the forthcoming reward is most uncertain: at 50% (Fiorillo *et al.*, 2003). It has been shown that the dopamine in the prefrontal cortex is necessary for maintaining working memory (Sawaguchi and Goldman-Rakic, 1994). Such novelty or uncertainty-related activation of dopamine may have a role in storing potentially important sensory cues in working memory for further learning.

Neuroeconomics and the social brain

In brain imaging experiments, money is the most convenient reward for human subjects. Incidentally in economics, a new trend called *experimental economics* has been surging, which put doubts on the classical assumption of rational humans and went out to analyze actual human behaviors in a variety of economic games. The marriage of neuroscience with experimental economics yielded a new field of *neuroeconomics*, which tries to understand the neural origin of economic and social behaviors (Barraclough *et al.*, 2004; Glimcher and Rustichini, 2004; Lee *et al.*, 2004; Lee *et al.*, 2005; Sanfey *et al.*, 2006; Soltani *et al.*, 2006).

While the majority of neuroscientists have been working on the function of single brains (or even single cells), many economists have been working on the theories and experiment on how multiple agents cooperate, exploit, trust, or compromise with each other. While traditional economics and game theory assumed perfect knowledge of the agents, neuroeconomists study how real human subjects learn and behave from actual interaction with other humans and ask what neural mechanisms are responsible for their behaviors.

CONCLUSION

Reinforcement learning is a theoretical framework that has promoted fruitful interaction with neuroscience, medicine, psychology, sociology, and economics. This is because the problem setup of reinforcement learning captures the basic features of animal and human behaviors. The development of a robust and flexible reinforcement learning algorithm may be a helpful model of understanding the sophisticated adaptive mechanisms in the brain. Also understanding how such algorithms can fail in certain conditions may shed light on the complex pathology of psychiatric disorders.

One fundamental question in reinforcement learning is where does the reward come from. In applications of reinforcement learning, engineers design reward functions based on their intuition of what should be done and what should be avoided. This designing process sometime takes a lot of trial and error. In the animal brain, the reward function should have been evolved to guarantee survival and reproduction. We are now trying to develop a more general theoretical framework in which the reward function, along with the temporal discounting parameter, are derived from a higher goal of survival and reproduction using a robotic platform as shown in Fig. 9 (Doya and Uchibe, 2005).

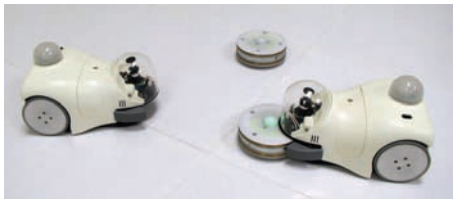

Figure 9. The Cyber rodent: robotic platform for assessing how the reward system should be designed and how it can be evolved (Doya and Uchibe, 2005). The marked features of Cyber Rodents are the capabilities to capture battery packs for survival and to exchange programs by infrared transmission for software reproduction.

Reinforcement learning theory is being applied to the understanding of how humans engage in cooperative social behaviors. But another interesting question is how billions of neurons in our brain can work cooperatively to run a complex society like the brain (Minsky, 1986; Houk, 2005). For example, almost all functional brain imaging studies rely on an assumption that when a specific function is required, a group of neurons in charge of that function become activated. However, we do not know how such dispatching of a neural circuit is actually realized. Is this just by a genetically prescribed design, or is this by exploration and consolidation by some kind of internal reward signal? Maybe concepts from economics and sociology may shed some light on how a complex society like the brain can work in such nice harmony.

REFERENCES

Balleine, BW, and Killcross, S (2006). "Parallel incentive processing: an integrated view of amygdala function." *Trends Neurosci.* **29**, 272–279.

Barraclough, DJ, Conroy, ML, and Lee, D (2004). "Prefrontal cortex and decision making in a mixed-strategy game." *Nat. Neurosci.* **7**, 404–410.

Barto, AG (1995). "Adaptive critics and the basal ganglia." In *Models of Information Processing in the Basal Ganglia*, Houk, JC, Davis, JL, and Beiser, DG (eds), pp 215–232, Cambridge, Mass.

Barto, AG, Sutton, RS, and Anderson, CW (1983). "Neuronlike adaptive elements that can solve difficult learning control problems." *IEEE Trans. Syst. Man Cybern.* **13**, 834–846.

Contreras-Vidal, JL, and Schultz, W (1999). "A predictive reinforcement model of dopamine neurons for learning approach behavior." *J. Comput. Neurosci.* **6**, 191–214.

Daw, ND, and Doya, K (2006). "The computational neurobiology of learning and reward." *Curr. Opin. Neurobiol.* **16**, 199–204.

Daw, ND, Kakade, S, and Dayan, P (2002). "Opponent interactions between serotonin and dopamine." *Neural Networks* **15**, 603–616.

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

Dayan, P, and Sejnowski, TJ (1996). "Exploration bonuses and dual control." *Mach. Learn.* **25**, 5–22.

Deiber, MP, Ibanez, V, Honda, M, Sadato, N, Raman, R, and Hallett, M (1998). "Cerebral processes related to visuomotor imagery and

generation of simple finger movements studied with positron emission tomography." *Neuroimage* **7**, 73–85.
Dorris, MC, and Glimcher, PW (2004). "Activity in posterior parietal cortex is correlated with the relative subjective desirability of action." *Neuron* **44**, 365–378.
Doya, K (1999). "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?" *Neural Networks* **12**, 961–974.
Doya, K (2000). "Complementary roles of basal ganglia and cerebellum in learning and motor control." *Curr. Opin. Neurobiol.* **10**, 732–739.
Doya, K (2002). "Metalearning and neuromodulation." *Neural Networks* **15**, 495–506.
Doya, K, Ishii, S, Pouget, A, and Rao, RPN (2007). "Bayesian brain: probabilistic approaches to neural coding." MIT Press, Cambridge, Mass.
Doya, K, and Uchibe, E (2005). "The Cyber Rodent project: exploration of adaptive mechanisms for self-preservation and self-reproduction." *Adaptive Behavior* **13**, 149–160.
Fiorillo, CD, Tobler, PN, and Schultz, W (2003). "Discrete coding of reward probability and uncertainty by dopamine neurons." *Science* **299**, 1898–1902.
Glimcher, PW, and Rustichini, A (2004). "Neuroeconomics: the consilience of brain and decision." *Science* **306**, 447–452.
Hanakawa, T, Immisch, I, Toma, K, Dimyan, MA, Van Gelderen, P, and Hallett, M (2003). "Functional properties of brain areas associated with motor execution and imagery." *J. Neurophysiol.* **89**, 989–1002.
Hikosaka, O, Nakamura, K, and Nakahara, H (2006). "Basal ganglia orient eyes to reward." *J. Neurophysiol.* **95**, 567–584.
Hoshi, E, Tremblay, L, Feger, J, Carras, PL, and Strick, PL (2005). "The cerebellum communicates with the basal ganglia." *Nat. Neurosci.* **8**, 1491–1493.
Houk, JC (2005). "Agents of the mind." *Biol. Cybern.* **92**, 427–437.
Houk, JC, Adams, JL, and Barto, AG (1995). "A model of how the basal ganglia generate and use neural signals that predict reinforcement." In *Models of Information Processing in the Basal Ganglia*, Houk, J. C., Davis, J. L., and Beiser, D. G. (eds), pp 249–270, MIT Press, Cambridge, Mass.
Houk, JC, and Wise, SP (1995). "Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action." *Cereb. Cortex* **2**, 95–110.
Kaelbling, LP, Littman, ML, and Cassandra, AR (1998). "Planning and action in partially observable stochastic domains." *Artif. Intell.* **101**, 99–134.
Kawagoe, R, Takikawa, Y, and Hikosaka, O (1998). "Expectation of reward modulates cognitive signals in the basal ganglia." *Nat. Neurosci.* **1**, 411–416.
Kawagoe, R, Takikawa, Y, and Hikosaka, O (2004). "Reward-predicting activity of dopamine and caudate neurons—a possible mechanism of motivational control of saccadic eye movement." *J. Neurophysiol.* **91**, 1013–1024.
Kawato, M (1999). "Internal models for motor control and trajectory planning." *Curr. Opin. Neurobiol.* **9**, 718–727.
Knill, DC, and Pouget, A (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation." *Trends Neurosci.* **27**, 712–719.
Lee, D, Conroy, ML, McGreevy, BP, and Barraclough, DJ (2004). "Reinforcement learning and decision making in monkeys during a competitive game." *Brain Res. Cognit. Brain Res.* **22**, 45–58.
Lee, D, McGreevy, BP, and Barraclough, DJ (2005). "Learning and decision making in monkeys during a rock-paper-scissors game." *Brain Res. Cognit. Brain Res.* **25**, 416–430.
Marder, E, and Thirumalai, V (2002). "Cellular, synaptic and network effects of neuromodulation." *Neural Networks* **15**, 479–493.
Matsumoto, K, Suzuki, W, and Tanaka, K (2003). "Neuronal correlates of goal-based motor selection in the prefrontal cortex." *Science* **301**, 229–232.
Minsky, M (1986). Society of Mind, Simon and Schuster, New York.
Montague, PR, Dayan, P, and Sejnowski, TJ (1996). "A framework for mesencephalic dopamine systems based on predictive Hebbian learning." *J. Neurosci.* **16**, 1936–1947.
Morimoto, J, and Doya, K (2001). "Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning." *Robotics and Autonomous Systems* **36**, 37–51.
Morris, G, Nevet, A, Arkadir, D, Vaadia, E, and Bergman, H (2006). "Midbrain dopamine neurons encode decisions for future action." *Nat. Neurosci.* **9**, 1057–1063.
Nakahara, H, Itoh, H, Kawagoe, R, Takikawa, Y, and Hikosaka, O (2004). "Dopamine neurons can represent context-dependent prediction error." *Neuron* **41**, 269–280.
Platt, ML, and Glimcher, PW (1999). "Neural correlates of decision variables in parietal cortex." *Nature (London)* **400**, 233–238.
Redgrave, P, and Gurney, K (2006). "The short-latency dopamine signal: a role in discovering novel actions?" *Nat. Rev. Neurosci.* **7**, 967–975.
Reynolds, JN, Hyland, BI, and Wickens, JR (2001). "A cellular mechanism of reward-related learning." *Nature (London)* **413**, 67–70.
Reynolds, JN, and Wickens, JR (2000). "Substantia nigra dopamine regulates synaptic plasticity and membrane potential fluctuations in the rat neostriatum, in vivo." *Neuroscience* **99**, 199–203.
Reynolds, JN, and Wickens, JR (2002). "Dopamine-dependent plasticity of corticostriatal synapses." *Neural Networks* **15**, 507–521.
Samejima, K, Ueda, Y, Doya, K, and Kimura, M (2005). "Representation of action-specific reward values in the striatum." *Science* **310**, 1337–1340.
Sanfey, AG, Loewenstein, G, McClure, SM, and Cohen, JD (2006). "Neuroeconomics: cross-currents in research on decision-making." *Trends Cogn. Sci.* **10**, 108–116.
Satoh, T, Nakai, S, Sato, T, and Kimura, M (2003). "Correlated coding of motivation and outcome of decision by dopamine neurons." *J. Neurosci.* **23**, 9913–9923.
Sawaguchi, T, and Goldman-Rakic, P S. (1994). "The role of D1-dopamine receptor in working memory: local injections of dopamine antagonists into the prefrontal cortex of rhesus monkeys performing an oculomotor delayed-response task." *J. Neurophysiol.* **71**, 515–528.
Sawamura, H, Shima, K, and Tanji, J (2002). "Numerical representation for action in the parietal cortex of the monkey." *Nature (London)* **415**, 918–922.
Schultz, W, Romo, R, Ljungberg, T, Mirenowicz, J, Hollerman, JR, and Dickson, A (1995). "Reward-related signals carried by dopamine neurons." In *Models of Information Processing in the Basal Ganglia*, Houk, JC, Davis, JL, and Beiser, DG (eds), pp 233–248, Cambridge, Mass.
Schultz, W (1998). "Predictive reward signal of dopamine neurons." *J. Neurophysiol.* **80**, 1–27.
Schultz, W, Dayan, P, and Montague, PR (1997). "A neural substrate of prediction and reward." *Science* **275**, 1593–1599.
Schultz, W, Tremblay, L, and Hollerman, JR (2000). "Reward processing in primate orbitofrontal cortex and basal ganglia." *Cereb. Cortex* **10**, 272–284.
Soltani, A, Lee, D, and Wang, XJ (2006). "Neural mechanism for stochastic behaviour during a competitive game." *Neural Networks* **19**, 1075–1090.
Sugrue, LP, Corrado, GS, and Newsome, W T. (2004). "Matching behavior and the representation of value in the parietal cortex." *Science* **304**, 1782–1787.
Suri, RE, and Schultz, W (1998). "Learning of sequential movements by neural network model with dopamine-like reinforcement signal." *Exp. Brain Res.* **121**, 350–354.
Sutton, RS (1988). "Learning to predict by the methods of temporal difference." *Mach. Learn.* **3**, 9–44.
Sutton, RS, and Barto, AG (1998). Reinforcement Learning, MIT Press, Cambridge, Mass.
Tanaka, SC, Doya, K, Okada, G, Ueda, K, Okamoto, Y, and Yamawaki, S (2004). "Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops." *Nat. Neurosci.* **7**, 887–893.
Tesauro, G (1994). "TD-Gammon, a self teaching backgammon program, achieves master-level play." *Neural Comput.* **6**, 215–219.
Thorndike, EL (1898). "Animal intelligence: an experimental study of the associate processes in animals." *Psychol. Rev.* **2**, 1–109.
Waelti, P, Dickinson, A, and Schultz, W (2001). "Dopamine responses comply with basic assumptions of formal learning theory." *Nature (London)* **412**, 43–48.
Watanabe, M (1996). "Reward expectancy in primate prefrontal neurons." *Nature (London)* **382**, 629–632.
Watkins, C JCH (1989). "Learning from delayed rewards." Ph.D. thesis, University of Cambridge.
Watkins, CJCH, and Dayan, P (1992). "Q-learning." *Mach. Learn.* **8**, 279–292.

Werbos, PJ (1990). "A menu of designs for reinforcement learning over time." In *Neural Networks for Control*, Miller, WT, Sutton, RS, and Werbos, PJ (eds), pp 67–95, MIT Press, Cambridge, Mass.
Wickens, JR, Begg, AJ, and Arbuthnott, GW (1996). "Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro." *Neuroscience* **70**, 1–5.
Winstanley, CA, Theobald, DE, Dalley, JW, Cardinal, RN, and Robbins, TW (2006). "Double dissociation between serotonergic and dopaminergic modulation of medial prefrontal and orbitofrontal cortex during a test of impulsive choice." *Cereb. Cortex* **16**, 106–114.
Yoshida, W, and Ishii, S (2006). "Resolution of uncertainty in prefrontal cortex." *Neuron* **50**, 781–789.
Yu, AJ, and Dayan, P (2005). "Uncertainty, neuromodulation, and attention." *Neuron* **46**, 681–692.