PERGAMON

Neural Networks 15 (2002) 495–506

Neural
Networks

www.elsevier.com/locate/neunet

2002 Special issue

Metalearning and neuromodulation

Kenji Doya*

ATR Human Information Science Laboratories, CREST, Japan Science and Technology Corporation, 2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan

Received 28 December 2001; accepted 19 April 2002

Abstract

This paper presents a computational theory on the roles of the ascending neuromodulatory systems from the viewpoint that they mediate the global signals that regulate the distributed learning mechanisms in the brain. Based on the review of experimental data and theoretical models, it is proposed that dopamine signals the error in reward prediction, serotonin controls the time scale of reward prediction, noradrenaline controls the randomness in action selection, and acetylcholine controls the speed of memory update. The possible interactions between those neuromodulators and the environment are predicted on the basis of computational theory of metalearning. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Metalearning; Neuromodulator; Dopamine; Serotonin; Noradrenaline; Acetylcholine; Reinforcement learning; Discount factor

1. Introduction

Some of the neurotransmitters that have spatially distributed, temporally extended effects on the recipient neurons and circuits are called *Neuromodulators* (Katz, 1999; Marder & Thirumalai, 2002; Saper, 2000). The best known examples of neuromodulators are *dopamine* (DA), *serotonin* (5-HT), *noradrenaline* (NA; also called *norepinephrine*, NE), and *acetylcholine* (ACh). Neuromodulators are traditionally assumed to be involved in the control of general arousal (Robbins, 1997; Saper, 2000). Recent advances in molecular biological techniques have provided rich data on the spatial localization and physiological effects of different neuromodulators and their receptors. This prompted us to build a more specific yet still comprehensive theory for the functions of neuromodulators. This paper proposes a computational theory on the roles of the earlier four major neuromodulators from the viewpoint that neuromodulators are media for signaling specific global variables and parameters that regulate distributed learning modules in the brain (Doya, 2000b).

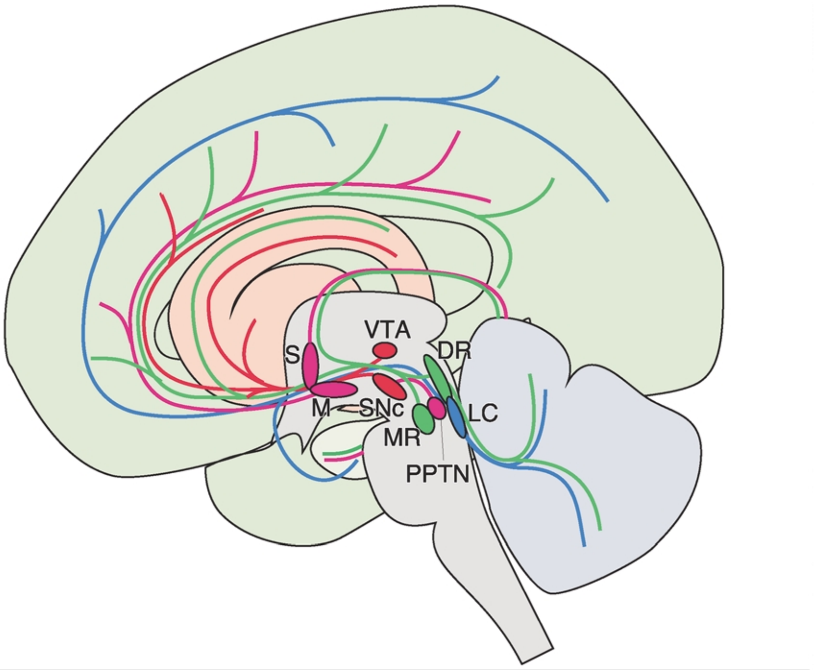
The computational theory for acquisition of goal-directed behaviors has been formulated under the name of *reinforcement learning* (RL) (Barto, 1995b; Doya, 2000c; Doya, Kimura, & Kawato, 2001; Sutton & Barto, 1998). The theory has been successfully applied to a variety of dynamic optimization problems, such as game programs (Tesauro, 1994), robotic control (Morimoto & Doya, 2001), and resource allocation (Singh & Bertsekas, 1997). In practical applications of reinforcement learning theory, a critical issue is how to set the parameters of the learning algorithms, such as the speed of learning, the size of noise for exploration, and the time scale in prediction of future reward. Such parameters globally affect the way many system parameters change by learning, so they are called *metaparameters* or *hyperparameters*.

In statistical learning theory, the need for setting the right metaparameters, such as the degree of freedom of statistical models and the prior distribution of parameters, is widely recognized. Theories of metaparameter setting have been developed from the viewpoints of risk-minimization (Vapnik, 2000) and Bayesian estimation (Neal, 1996). However, many applications of reinforcement learning have depended on heuristic search for setting the right metaparameters by human experts. The need for the tuning of metaparameters is one of the major reasons why sophisticated learning algorithms, which perform successfully in the laboratory, cannot be practically applied in highly variable environments at home or on the street.

Compared to current artificial learning systems, the learning mechanisms implemented in the brain appear to be much more robust and flexible. Humans and animals can learn novel behaviors under a wide variety of environments. This suggests that the brain has a certain mechanism for *metalearning*, a capability of dynamically adjusting its own metaparameters of learning. This paper presents a

* Tel.: +81-774-95-1251; fax: +81-774-95-1259.
E-mail address: doya@atr.co.jp (K. Doya).

PII: S0893-6080(02)00044-8



neuromodulator	origin of projection	major target area
dopamine (DA)	substantia nigra, pars compacta (SNc)	dorsal striatum
	ventral tegmental area (VTA)	ventral striatum
		frontal cortex
serotonin (5-HT)	dorsal raphe nucleus (DR)	cortex, striatum
		cerebellum
	median raphe nucleus (MR)	hippocampus
noradrenaline (NA)	locus coeruleus (LC)	cortex, hippocampus
(norepinephrine, NE)		cerebellum
acetylcholine (ACh)	Meynert nucleus (M)	cortex, amygdala
	medial septum (S)	hippocampus
	pedunculopontine tegmental	SNc, thalamus
	nucleus (PPTN)	superior colliculus

Fig. 1. Major neuromodulator systems that project diffusely to the cortex, the basal ganglia, and the cerebellum from brain stem nuclei. The dopaminergic system is shown in red, the serotonergic system in green, the noradrenergic system in blue, and the cholinergic system in magenta. The table shows the origins and targets of projections their abbreviations.

hypothesis stating that the ascending neuromodulatory systems (Fig. 1) are the media of metalearning for controlling and coordinating the distributed learning modules in the brain (Doya, 1999). More specifically, we propose the following set of hypotheses to explain the roles of the four major ascending neuromodulators (Doya, 2000b):

1. Dopamine represents the global learning signal for prediction of rewards and reinforcement of actions.
2. Serotonin controls the balance between short-term and long-term prediction of reward.
3. Noradrenaline controls the balance between wide exploration and focused execution.
4. Acetylcholine controls the balance between memory storage and renewal.

In order to state the above hypotheses in a more computationally well-defined manner, we first review the basic algorithms of reinforcement learning and the roles of major metaparameters. We then propose a set of hypotheses on how such metaparameters are regulated by the above neuromodulators. Finally, we discuss the possible neural mechanisms of metaparameter control and the possible interactions between neuromodulatory systems predicted from the hypotheses.

In this paper, our main focus is on the roles of neuromodulators within the circuit of basal ganglia, which have been suggested as the major locus of reinforcement learning (Doya, 2000a; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996). However, we also discuss how their roles can be generalized to other brain areas, including the cerebral cortex and the cerebellum.

2. Reinforcement learning algorithm

Reinforcement learning is a computational framework

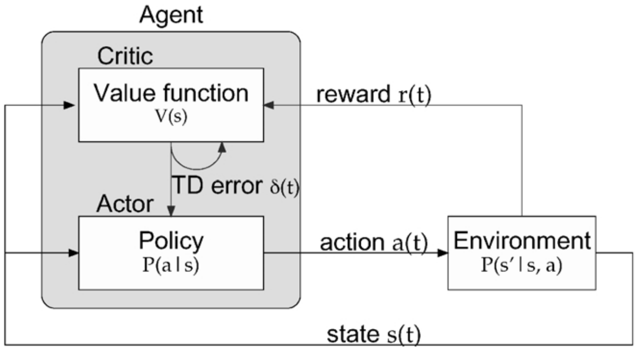


Fig. 2. A standard architecture for reinforcement learning, known as actor–critic. The agent has a two-part organization: the critic learns the state value function $V(s)$ and the actor learns the policy $P(a|s)$. The TD error $\delta(t)$ is used as the error signal for the learning of the critic and the reinforcement signal is used for the learning of the actor.

for an *agent* to learn to take an *action* in response to the *state* of the environment so that the acquired *reward* is maximized in a long run (Fig. 2) (Barto, 1995b; Doya, 2000c; Doya et al., 2001; Sutton & Barto, 1998). What makes reinforcement learning difficult yet interesting is that selection of an action does not only affect the immediate reward but also affect the future rewards through the dynamic evolution of the future states.

In order to outline a basic algorithm of reinforcement learning, here we consider a Markov decision problem (MDP), which assumes a discrete state, action, and time. The agent observes the state $s(t) \in \{s_1, \ldots, s_n\}$ and takes an action $a(t) \in \{a_1, \ldots, a_m\}$ according to its *policy*, which is given either deterministically as $a = G(s)$ or stochastically as $P(a|s)$. In response to the agent's action $a(t)$, the state of the environment changes either deterministically as $s(t+1) = F(s(t), a(t))$ or stochastically according to a Markov transition matrix $P(s(t+1)|s(t), a(t))$ for each action $a(t)$. The reward $r(t+1) \in \mathbf{R}$ is given deterministically as $r(t+1) = R(s(t), a(t))$, or stochastically according to $P(r(t+1)|s(t), a(t))$.

Given the earlier setup, the goal of reinforcement learning is to find an optimal policy that maximizes the expected sum of future rewards. A commonly used architecture for reinforcement learning is the actor–critic (Barto, 1995a; Barto, Sutton, & Anderson, 1983) (Fig. 2), which consists of two parts: (1) the critic, which learns to predict future rewards in the form of a state value function $V(s)$ for the current policy, and (2) the actor, which improves the policy $P(a|s)$ in reference to the future reward predicted by the critic. The actor and the critic learn either alternately or concurrently. For theoretical details of such learning algorithms, please refer to Singh, Jaakkola, Littman, and Szpesvari (2000).

2.1. State value function and TD error

The critic learns the state value function, which is defined as the cumulative future reward expected by following the current policy $P(a|s)$ from the state $s(t)$, i.e.

$$V(s(t)) = E[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \cdots]. \tag{1}$$

Here, the parameter γ $(0 \le \gamma \le 1)$ is the discount factor, which assigns less weight on the reward expected in the far future. Fig. 3 shows an example of state value function for a simple deterministic policy in a three-step decision task. The value function can guide the agent's behavior by signaling how good or bad state the agent is in based on the prediction of future reward.

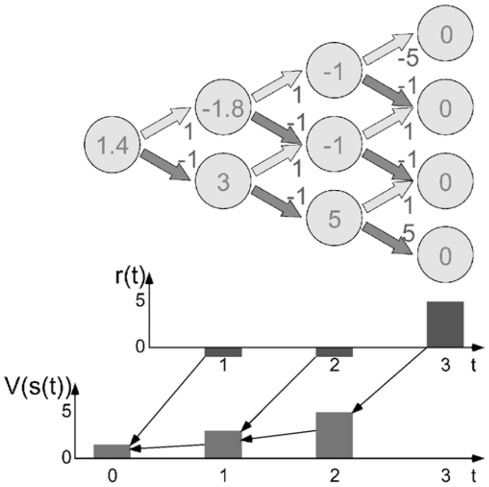


Fig. 3. Illustration of the state value function for a simple policy in a three-step decision task. The open circles represent the states $(s_1, \ldots, s_{10})$, the arrows the actions (a_1: going up; a_2 going down), the figure by the arrow the rewards $r = R(s, a)$. The figures in the circles show the value function with the discount factor $\gamma = 0.8$ for a deterministic policy of always taking the action a_2. The graphs below show the time courses of reward $r(t)$ and the state value function $V(s(t)) = r(t+1) + \gamma r(t+2) + \cdots = r(t+1) + \gamma V(s(t+1))$. For example, at the initial state s_1, the action value functions are $Q(s_1, a_1) = R(s_1, a_1) + \gamma V(s_2) = 1 + 0.8 \times (-1.8) = -0.44$ and $Q(s_1, a_2) = R(s_1, a_2) + \gamma V(s_3) = -1 + 0.8 \times 3 = 1.4$. Thus in reference to the future reward prediction by the value function, the agent can choose to take a_2, although its immediate reward is negative.

The value functions for adjacent states should satisfy the consistency condition

$$\begin{aligned} V(s(t-1) &= E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \cdots] \\ &= E[r(t) + \gamma V(s(t))]. \end{aligned} \tag{2}$$

Thus any deviation from this consistency condition, expressed as

$$\delta(t) = r(t) + \gamma V(s(t)) - V(s(t-1)), \tag{3}$$

should be zero on average. This signal, called the temporal difference (TD) error, is used as the error signal for learning the value function. The standard way is to correct the older estimate of value function $V(s(t-1))$ in proportion to the

TD error $\delta(t)$:

$$\Delta V(s(t-1)) \propto \delta(t). \tag{4}$$

2.2. Action value function and policy

Although the learning of the critic brings the TD error (3) close to zero, when the policy is stochastic, it fluctuates around zero according to the particular choice of the preceding actions. It can be seen from Eq. (3) that a positive TD error $\delta(t) > 0$ means that the agent acquired more reward $r(t)$ than expected, or reached a state $s(t)$ with a higher value $V(s(t))$ than expected. Thus the agent should increase the probability of taking the same action when it comes to the same state again, that is, to increase the probability $P(a(t-1)|s(t-1))$. The opposite holds for a negative TD error $\delta(t) < 0$. Thus the TD error $\delta(t)$ does not only serve as the error signal for the critic to learn the state value function $V(s)$, but also serves as the reinforcement signal for the actor to update the policy $P(a|s)$.

A common way of defining the policy is via the action value function $Q(s,a)$, which represents how much future rewards the agent would get by taking an action a at state s. For a given state s, the action values $Q(s,a_i)$ for the candidate actions a_i $(i = 1,\ldots,m)$ are compared and the one with a higher action value $Q(s,a_i)$ is selected with a higher probability. A typical way is so-called Boltzmann selection, in which the policy is given by

$$P(a_i|s) = \frac{\exp[\beta Q(s,a_i)]}{\sum_{j=1}^{m} \exp[\beta Q(s,a_j)]}. \tag{5}$$

Here, the parameter β, which is called the inverse temperature, controls the stochasticity of the policy. With $\beta = 0$, action selection is totally random, i.e. $P(a_i|s) = 1/m$. As β is increased, action selection becomes less random, and the probability of selecting the action with the highest action value becomes close to one. In the limit of $\beta \to \infty$, Eq. (5) becomes a deterministic, winner-take-all rule:

$$a(t) = \arg\max_a Q(s(t),a). \tag{6}$$

The action value function is defined as

$$Q(s(t),a) = E[r(t+1) + \gamma V(s(t+1))|a(t) = a], \tag{7}$$

where the immediate reward $r(t+1)$ and the state $s(t+1)$ depend on the particular choice of action a at state $s(t)$. After an action $a(t-1)$ is taken at a state $s(t-1)$ and the TD error $\delta(t)$ is calculated from the resulting reward $r(t)$ and state $s(t)$, the action value function for the state-action pair is updated by

$$\Delta Q(s(t-1),a(t-1)) \propto \delta(t), \tag{8}$$

2.3. Global learning signal and metaparameters

In implementing the earlier reinforcement learning algorithm, the value functions $V(s)$ and $Q(s,a)$ are often represented using a weighted sum of basis functions, such as

$$V(s) = \sum_j v_j b_j(s) \tag{9}$$

and

$$Q(s,a) = \sum_k w_k c_k(s,a). \tag{10}$$

Here, $b_j(s)$ $(j = 1,\ldots,N)$ and $c_k(s,a)$ $(k = 1,\ldots,M)$ are the basis functions, i.e. internal representation of states and actions of the agent, and v_j and w_k are the weight parameters. In order to update the value functions according to Eqs. (4) and (8), the weight parameters are updated as

$$\Delta v_j = \alpha\delta(t) b_j(s(t-1)) \tag{11}$$

and

$$\Delta w_k = \alpha\delta(t) c_k(s(t-1),a(t-1)), \tag{12}$$

where α is the learning rate for controlling how quickly the old memory is updated by experience. In this learning algorithm, the TD error $\delta(t)$ works as the global learning signal for updating the parameters v_j and w_k.

The way in which parameters like v_j and w_k change over time is also dependent on a number of global parameters, such as the learning rate α, the inverse temperature β, and the discount factor γ. These are called metaparameters or hyperparameters, since they are higher-level parameters that regulate the way a large number of ordinary parameters like v_j and w_k change by learning. In order to realize efficient learning, the setting of those metaparameters should match the characteristics of the environment and the task for the agent. In the following, we propose a set of hypotheses stating that distinct neuromodulators are used for signaling these global learning signal and metaparameters.

3. Hypothetical roles of neuromodulators

Now we restate our hypotheses on the roles of neuromodulators in terms of the global learning signal and metaparameters introduced in the above reinforcement learning algorithm (Doya, 2000b):

1. Dopamine signals the TD error δ.
2. Serotonin controls the discount factor γ.
3. Noradrenaline controls the inverse temperature β.
4. Acetylcholine controls the learning rate α.

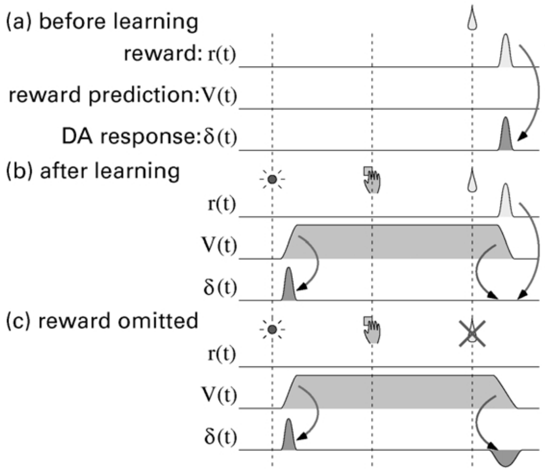


Fig. 4. Interpretation of the responses of midbrain dopamine neurons in the TD model, which assumes that the change in DA neuron firing relative to its baseline represents the TD error $\delta(t) = r(t) + \gamma V(t) - V(t-1)$. (a) Before learning, no reward is predicted, i.e. $V(t) \equiv 0$. Thus the TD error $\delta(t)$ is the same as the reward itself, $r(t)$. (b) After learning is complete, the predicted future reward $V(t)$ builds up immediately after the cue signal, causing the discounted temporal derivative $\gamma V(t) - V(t-1)$ to provide a positive pulse in the TD error even if there is no reward $r(t)$. At the time of reward delivery, as there will be no more reward for the time being, $V(t)$ drops to zero and the negative temporal derivative of $V(t)$ cancels out the positive reward signal $r(t)$. (c) If the promised reward is omitted, there is a negative response due to the drop in the predicted reward $V(t)$.

Below, we review the experimental findings and theoretical models that support these hypotheses.

3.1. Dopamine and reward prediction

As described in Section 2.3, the TD error δ is the essential learning signal for reward prediction and action selection. The midbrain dopaminergic system seems to be critically involved in both functions.

Schultz and colleagues (Schultz, 1998; Schultz, Apicella, & Ljungberg, 1993; Schultz, Dayan, & Montague, 1997) performed a series of experiments on the response of dopaminergic neurons in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA). They trained monkeys to press a lever after a light was turned on. Dopamine neurons responded to liquid reward early in learning, or when reward was given unexpectedly outside the task (Fig. 4(a)). After monkeys learned the task well, dopamine neurons responded to the cue light and did not respond to the reward itself (Fig. 4(b)). If the reward was omitted, dopamine neuron activity was depressed (Fig. 4(c)). Such changes in response closely resemble the behavior of the TD error $\delta(t)$ in the course of reward prediction learning (Houk et al., 1995; Montague et al., 1996; Suri, 2002) (Fig. 4).

In addition to the reward predictive response, dopamine is also known to be involved in action learning. Electric stimulation of the dopaminergic system has the effect of reinforcement, i.e. the animal learns to repeat the action that preceded the stimulus. Furthermore, most addictive drugs have the effect of increasing the release or blocking the reuptake of dopamine (Wise, 1996).

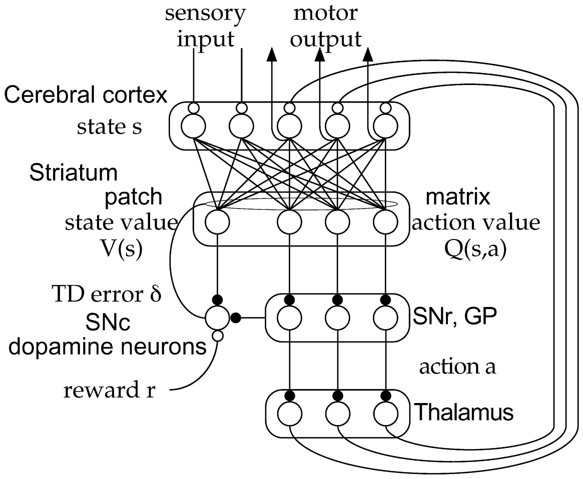


Fig. 5. Schematic diagram and hypothetical functions of the circuit of the basal ganglia. SNc and SNr: compact and reticular parts of substantia nigra. GP: globus pallidus.

At the cellular level, dopamine modulates the synaptic plasticity of cortical input to the striatum. In an ordinary Hebbian plasticity model, a synapse is potentiated after frequent stimulation that leads to response of the post-synaptic neuron. In the striatal neuron, however, the direction of plasticity is reversed with the level of dopamine (Reynolds & Wickens, 2001, 2002; Wickens, Beggs, & Arbuthnott, 1996). Thus, if the change in firing of dopamine neurons from the baseline encodes the TD error $\delta(t)$ and the cortical input represents the basis functions $b_j(s)$ and $c_k(s, a)$, the synaptic learning rules (11) and (12) can be implemented with this dopamine-dependent plasticity.

These observations strongly suggest that dopamine activity represents the TD error, which is used for learning of reward prediction and action selection.

3.1.1. Reinforcement learning model of the basal ganglia

It has been hypothesized that the neural circuit of the basal ganglia plays a major role in reinforcement learning (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). Fig. 5 shows hypothetical functions of the components of the basal ganglia circuit (Doya, 2000a). The striatum, the input part of the basal ganglia, learns the state value functions $V(a)$ and the policy $P(a|s)$, possibly in the form of action value functions $Q(s,a)$. The midbrain dopaminergic neuron represents the TD error δ, which is fed back to the striatum for the learning of the value functions thorough dopamine dependent plasticity of the cortico-striatal synapses.

More specifically, the cerebral cortex provides internal representation of the environmental states. Neurons in one of the two compartments of the striatum, the patch or striosome, represent the state value function $V(s)$. Their output is sent to the dopaminergic neurons in the compact

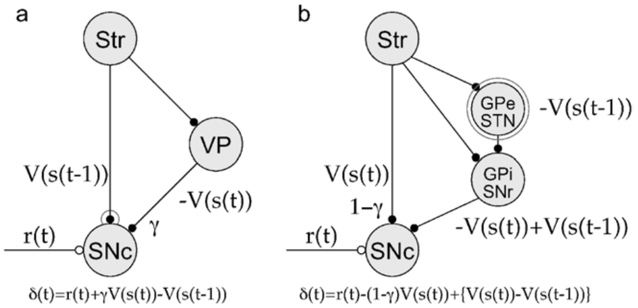


Fig. 6. Possible mechanisms for TD computation of the value function represented in the basal ganglia. The circles ○ represent excitatory connection while black dots • represents inhibitory connections. A large gray circle represents a source of possible time delay. SNc and SNr: compact and reticular parts of substantia nigra. Str: striatum. GPi and GPe: internal and external segments of globus pallidus. STN: subthalamic nucleus.

part of SNc to compute the TD error δ. Different groups of neurons in the other compartment, the matrix, represent the action value functions $Q(s, a_i)$ for different actions. Their outputs are sent to the reticular part of the substantia nigra (SNr) and the globus pallidus (GP), where competitive dynamics realize a Q-value-dependent action selection mechanism, such as Eqs. (5) or (6).

One of the major criticisms of this dopamine TD hypothesis is that the dopamine neurons respond to salient, non-rewarding stimuli (Horvitz, 2000; Redgrave, Prescott, & Gurney, 1999). Some of these responses can be explained by considering optimistic biases for promoting exploration, generalization in responses, and prediction using internal models (Kakade & Dayan, 2002; Suri, 2002).

Another critical issue in the reinforcement learning model of the basal ganglia is how to compute the 'temporal difference' of the value function $V(s)$, represented in the striatal output (Joel, Niv, & Ruppin, 2002). It has been proposed that the direct inhibitory connection from the striatum to SNc has long latency, providing $V(s(t-1))$, while the indirect double-inhibitory connection has a shorter latency, thus providing $V(s(t))$ (Fig. 6(a)) (Houk et al., 1995). Although slow GABA$_B$ receptors are known to exist in the direct inhibitory connection (Tepper, Martin, & Anderson, 1995), recent data suggest that both direct and indirect inputs are predominantly mediated by fast GABA$_A$ receptors (Paladini, Celada, & Tepper, 1999).

A different mechanism for TD computation is possible if we consider an alternative form of the TD error (Doya, 2000c):

$$\delta(t) = r(t) - (1-\gamma)V(s(t)) + (V(s(t)) - V(s(t-1))), \quad (13)$$

which is the sum of the reward, immediate inhibition, and TD of the value functions (Fig. 6(b)). This can be implemented if there is a delay mechanism in the indirect pathway, for example, within the recurrent circuit between GPe and STN. Other mechanisms have also been proposed (Brown, Bullock, & Grossberg, 1999; Joel et al., 2002) and further studies are necessary to clarify the neural circuit for computation of TD.

3.1.2. Dopaminergic modulation in the cerebral cortex

Although the striatum (including its ventral part, the nucleus accumbens) receives the strongest dopaminergic input in the forebrain, the cerebral cortex, especially the prefrontal cortex, receive dopaminergic projections from VTA. A possible role of dopamine in the cerebral cortex is to guide the acquisition of task-dependent internal representation of states and actions, which facilitate reward prediction and action selection in the basal ganglia.

In computation of value functions, as in Eqs. (9) and (10), the selection of the basis functions $b_i(s)$ and $c_k(s, a)$ critically affects the performance of learning. It has been shown that learning of task-dependent basis functions is possible in a biologically plausible way (without resorting to error backpropagation) using a global reward signal (Gullapalli, 1990; Mazzoni et al., 1991).

When the state of the environment is not perfectly observable (situation called partially-observable Markov decision problem, POMDP), the agent has to predict and update the invisible state of the environment in a form of working memory. An important problem is what to store in the working memory, since keeping track of all the invisible states of the environment is too demanding and wasteful. The neurons in the prefrontal cortex show sustained activity during working memory tasks. It has been shown that their responses are dependent on the amount and type of the reward associated with the memorized items (Tremblay & Schultz, 2000; Watanabe, 1996). It has also been shown that stable working memory requires activation of dopamine D1 receptor (Durstewitz & Seamans, 2002; Sawaguchi & Goldman-Rakic, 1994). Dopaminergic control of prefrontal working memory would be helpful for selective use of memory capacity for the sensory cues that are relevant for getting the reward.

3.2. Serotonin and time scale of reward prediction

The discount factor γ in the definition of the value function (1) determines how far into the future the agent should consider in reward prediction and action selection. The setting of the discount factor is particularly important when there is a conflict between the immediate and long-term outcomes (Fig. 7). In real life, it is often the case that one would have to pay some immediate cost (negative reward) in order to achieve a larger future reward, e.g. long travels in foraging or daily cultivation for harvest. It is also the case that one should avoid positive immediate reward if it is associated with a big negative reward in the future.

Although the discount factor γ has to be set large enough to achieve good behaviors over the long run, the closer γ approaches one, the more difficult it is to reliably predict the corresponding future reward (Baxter & Bartlett, 2000;

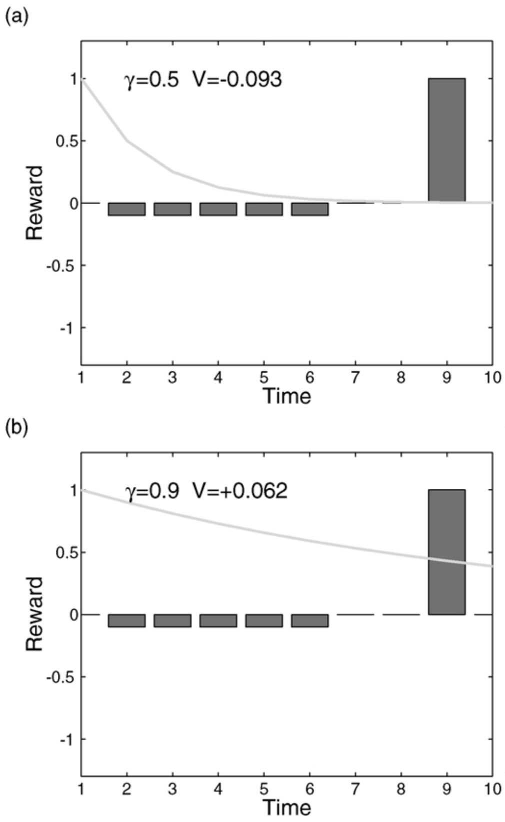


Fig. 7. The effect of discount factor γ in decision making. The gray lines show the weight γ^t for discounting. In a scenario where negative reward (cost) is expected before achieving a large positive reward, the cumulative future reward V becomes negative if γ is small (a) and positive if γ is large enough (b). Assuming that there is a baseline behavior where the expected reward is zero, such behavior is rejected as negligible with a small γ.

Littman, Cassandra, & Kaelbling, 1995). Furthermore, in many real-world tasks, rewards have to be acquired not too far in the future; an animal has to find food before it starves and has to find a mate before the reproductive season comes to an end. Thus, an appropriate value of the discount factor has to be chosen to match the demand of the task and the knowledge of the agent.

The serotonergic system has often been recognized as an opponent of the dopaminergic system (Daw, Kakade, & Dayan, 2002; Deakin & Graeff, 1991; Fletcher, 1995). However, the interactions between serotonin and dopamine appear to be more complex than those of simple opponents. Serotonin and dopamine in some cases work in a mutually facilitatory manner. For example, serotonin facilitates release of dopamine in the striatum (Sershen, Hashim, & Lajtha, 2000). Although genetic knockout of serotonin 1B receptor facilitates cocaine reinforcement (Rocha et al., 1998b; White, 1998), stimulation of serotonin 1B receptor enhances cocaine reinforcement (Parsons, Weiss, & Koob, 1998). While genetic knockout of dopamine transporter does not block cocaine reinforcement (Rocha et al., 1998a), knockout of both dopamine and serotonin transporters blocks cocaine reinforcement (Sora et al., 2001). It has been shown that not only the dopamine D1 receptor (Durstewitz & Seamans, 2002; Sawaguchi & Goldman-Rakic, 1994) but also the serotonin 2A receptor facilitate working memory in the prefrontal cortex (Williams, Rao, & Goldman-Rakic, 2002).

Such complex, context-dependent functions of serotonin may be better understood by assuming that serotonin controls the balance between short- and long-term consequence of actions by regulating the discount factor γ in reinforcement learning (Doya, 2000b).

In this hypothesis, a higher level of serotonin means a higher setting of the discount factor γ, which demands prediction of reward longer in the future. A low level of serotonin is often associated with impulsive behaviors, such as aggression (Buhot, 1997; Rahman, Sahakian, Cardinal, Rogers, & Robbins, 2001; Robbins, 2000). In experimental rats, depletion of the central serotonergic system resulted in impulsive choice of small immediate reward as opposed to larger, delayed reward (Mobini, Chang, Ho, Bradshaw, & Szabadi, 2000). Selective serotonin reuptake inhibitors (SSRI) and other serotonin enhancing drugs are also known to be effective for unipolar depression and bipolar disorders, although its therapeutical mechanism is still not fully understood (Wang & Licinio, 2001). The situation shown in Fig. 7(b), where the optimal policy is doing nothing with a small setting of the discount factor γ, can be a model of depression.

3.2.1. Possible mechanisms of the control of time scale

How can the discount factor γ be controlled in the above reinforcement learning model of the basal ganglia? In both Fig. 6(a) and (b), the relative balance of the direct and indirect pathways affects the discount factor; the contribution of the indirect pathway should be increased for larger γ. Thus if serotonin changes the balance of the effective strengths of the direct and indirect pathways, it could also change the effective discount factor. Both the striatum and the dopaminergic nuclei (SNc and VTA) receive serotonergic input from the dorsal raphe nucleus. An experiment comparing the serotonergic effects on the strengths of the direct and indirect pathways would clarify whether such a mechanism exists.

Another possible way of regulating the time scale of reward prediction is to activate or deactivate multiple reward prediction pathways. The cortico-basal ganglia circuit has a parallel loop organization, including cognitive and motor loops (Middleton & Strick, 2000). In addition to the basal ganglia, the amygdala is also involved in reinforcement learning. Furthermore, the cerebellum seems to provide internal models of the environmental dynamics (Kawato, 1999), which can be helpful in long-term prediction of the future events and rewards (Doya, 1999). It is possible that these different pathways are specialized for reward prediction and action selection in different time scales (Cardinal, Pennicott, Sugathapala, Robbins, & Everitt, 2001). If serotonin differentially enhances or suppresses the activities of these parallel

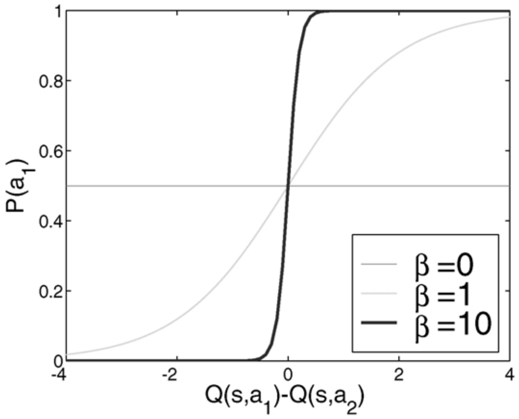


Fig. 8. The effect of inverse temperature β in action selection. Suppose there are two possible actions a_1 and a_2 at state s. The probability of taking action a_1 is plotted for the difference in the corresponding action value function $Q(s,a_1) - Q(s,a_2)$ with different values of inverse temperature $\beta = 0.1$, 1, and 10. Smaller β leads to more random action selection. Larger β leads to nearly deterministic selection of the action with the largest Q-value.

pathways, by using different receptors, transporters, and intracellular mechanisms (De Deurwaerde, Stinus, & Sampinato, 1998), it is possible for the serotonin to control the effective time scale of reward prediction.

3.3. Noradrenaline and randomness of action selection

While wide exploration by stochastic action selection facilitates learning of new behaviors, deterministic action selection such as Eq. (6) is favored in making the best use of what has already been learned. Thus, the randomness in action selection should be actively tuned in reference to the progress of learning and the urgency of the situation. This is known as the exploration-exploitation problem.

Fig. 8 shows how the different settings of the inverse temperature β affect the action selection probability defined by Eq. (5) for the case of two action choices. In this case, the probability of selecting one action is a sigmoid function of the difference in the Q-values for the two actions. The sigmoid curve becomes steeper with a larger value of β.

Noradrenaline has been known to be involved in the control of arousal and relaxation. The noradrenergic neurons in the locus coeruleus (LC) are activated in urgent situations (e.g. with aversive stimuli). More specifically, it was shown in an attention task of monkeys that the LC neuron activity is closely correlated with the accuracy of action selection; phasic response at the time of stimulus presentation is associated with a high accuracy of response (Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994). Such a change in the accuracy of response has been replicated in models in which noradrenaline sharpens the response tuning of neurons by increasing the input–output gain (Gilzenrat, Holmes, Rajkowski, Aston-Jones, & Cohen, 2002; Servan-Schreiber, Printz, & Cohen, 1990; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999; Usher & Davelaar, 2002).

These facts and models suggest that noradrenaline provides a means for dealing with the exploration-exploitation problem (Usher et al., 1999); a higher level of noradrenaline corresponds to a higher setting of the inverse temperature β, which results in reliable selection of the action with the largest predicted future reward. Interestingly, the use of amphetamine, which increase the level of noradrenaline by blocking its reuptake, results in stereotyped behaviors.

A basic question in the above hypothesis is how a stochastic choice is realized in the brain. Any neuron is subject to statistical opening and closing of ionic channels, but such small randomness could be quickly averaged away if the neuron operates in a stable attractor state. A possible mechanism for amplification of noise is chaotic or near-chaotic dynamics of the neuron and the circuit.

A marked feature of the neurons in GP is their high level of spontaneous firing. The inhibitory interactions of neurons with high spontaneous activity may realize an asynchronous, chaotic dynamic that serves as the roulette wheel for stochastic action selection. If the circuit in GP implements stochastic action selection (Fig. 5), the above hypothesis predicts that the randomness in action selection is subject to control by noradrenergic system. There is evidence showing that GP has a relatively high level of noradrenaline in the basal ganglia (Russell, Allin, Lamm, & Taljaard, 1992). Analysis of the changes in the randomness of GP neural firing is necessary for testing such a possibility.

3.4. Acetylcholine and memory update

It is obvious that learning becomes slow if the learning rate α is set too small. However, if it is set very large, what has already been learned could be quickly overwritten. Furthermore, if the learning rate is set too large, the learning process becomes unstable. For quick and accurate learning, the learning rate should be initially set large but gradually decreased. In a linear function approximator, modulation of learning rate in inverse proportion to the training time is often used. More adaptive methods for setting of learning rates have also been proposed (Murata, Kawanabe, Zienhe, Muller, & Amari, 2002; Sutton, 1992).

Acetylcholine appears to control the balance between the storage and update of memory at the both cellular and circuit levels (Hasselmo & Bower, 1993). Acetylcholine is known to modulate the synaptic plasticity in the hippocampus, the cerebral cortex, and the striatum (Partridge, Appasundaram, Gerhardt, Ronesi, & Lovinger, 2002; Rasmusson, 2000). Loss of cholinergic neurons in the Meynert nucleus is associated with memory disorders like Alzheimer's disease (Perry, Walker, Grance, & Perry, 1999).

In the above learning schemes (9) and (10), leaning is

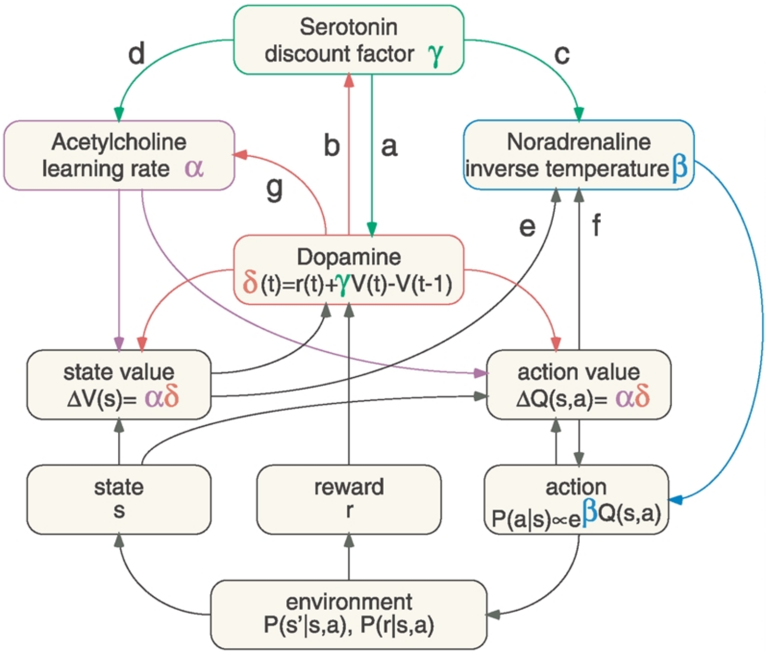


Fig. 9. Possible interactions between the neuromodulators representing the global learning signal and metaparameters, agent's experience in the form of value functions, and the state, action, and reward of the environment. See text for the particular ways of interactions shown by the arrows a through g.

achieved both by the plasticity of the output weights v_j and w_k, and the choice of the basis functions $b_j(s)$ and $c_k(s,a)$. The cholinergic neurons in the striatum, which shows a response timed to potentially rewarding sensory cues (Aosaki, Tsubukawa, Watanabe, Graybiel, & Kimura, 1994; Shimo & Hikosaka, 2001), may control the dopamine-dependent plasticity of cortico-striatal synapses (Partridge et al., 2002).

The cholinergic system may also modulate the information coding in the cortex and the hippocampus so that their response properties are not simply determined by the statistics of the sensory input but are also dependent on the importance of the sensory inputs. Hasselmo and colleagues proposed the role of acetylcholine in controlling the modes of operation in the hippocampus and the cortex: memory storage with a high level of cholinergic input and memory retrieval with a low level of cholinergic input (Hasselmo & Bower, 1993; Hasselmo & Schnell, 1994). Computational models have been proposed in which the cholinergic system controls the top-down and bottom up information flow based on the mismatch between the top-down prediction and bottom-up sensory inputs (Dayan & Yu, 2002; Hasselmo & Schnell, 1994; Yu & Dayan, 2002). Such control mechanisms of information flow would be helpful in action-based construction of internal representation of the environment, e.g. in our model the basis functions $b_j(s)$ and $c_k(s,a)$.

4. Dynamic interactions of neuromodulators

Based on the above hypotheses on the specific roles of neuromodulators in reinforcement learning, it is possible to theoretically predict how the activities of those modulators should depend on each other. Fig. 9 shows the possible interactions between the neuromodulators, the experience of the agent represented in the form of value functions, and the environment.

4.1. Dopamine as TD error

According to the definition of the TD error (3), the activity of dopamine $\delta(t)$ should depend on the value function $V(s)$ and the level of serotonin γ. Specifically, serotonin should have a facilitatory effect on dopamine when the predicted future reward $V(s(t))$ is positive, and an inhibitory effect when $V(s(t))$ is negative (Fig. 9(a)).

A different form of TD error is derived for average reward reward reinforcement learning (Mahadevan, 1996) by replacing the normalized value $(1-\gamma)V(s(t))$ in the discounted TD error (13) by the average reward $\bar{r}$

$$\delta(t) = r(t) - \bar{r} + V(s(t)) - V(s(t-1)). \qquad (14)$$

Based on this framework, Daw and colleagues proposed an alternate hypothesis that serotonin represents the predicted average reward $\bar{r}$, thus having an inhibitory effect on dopamine (Daw et al., 2002). A critical experiment to clarify which of these hypotheses are more appropriate is to compare the effects of serotonin on the activity of dopamine under different task conditions where predicted future reward $V(s)$ is either positive or negative.

4.2. Serotonin as discount factor

In general, learning to predict rewards in the far future is more difficult than to predict rewards in the near future. It has been shown that the value function learned with a large γ tends to have a large variance, and thus learning with a smaller γ can result in biased, but less variable estimate of future rewards (Baxter & Bartlett, 2000; Kakade, 2001; Littman et al., 1995). Since the high variability in the reward prediction is likely to be reflected in the variance of the TD error, it would possible to regulate the discount factor based on the variance in the TD error. Such a heuristic regulatory algorithm predicts that a high variability in the dopaminergic activity should have an inhibitory effect on the serotonergic system (Fig. 9(b)).

Moreover, in order to learn the long-term consequence of behaviors, the agent should not fix its behavior too quickly. Thus when the discount factor γ is large, both the inverse temperature γ and the learning rate α should be kept small. This predicts inhibitory effects of serotonin onto noradrenaline (Fig. 9(d)) and acetylcholine (Fig. 9(d)).

4.3. Noradrenaline as inverse temperature

In general, an agent has to resort to extensive, random search when it has little knowledge about the environment and should gradually focus its search as it acquires more knowledge about the environment. A typical method is to take an increasing schedule of the inverse temperature β, known as annealing.

A more adaptive way is to control the randomness in action selection depending on the performance of the agent. When the performance of an agent is close to its best desired, a better performance is given by reducing the randomness in action selection (Doya, 2000c; Gullapalli, 1990). On the other hand, when an agent is in a dangerous situation, it would not have a leisure of doing random search and should select the best possible action. Since the performance level is reflected in the level of state value function, the level of noradrenaline, representing the inverse temperature, should increase when the state value function is very high or very low (Fig. 9(e)).

In order to keep variability of actions at different states, Ishii and colleagues proposed a method of state-dependent control of the inverse temperature (Ishii, Yoshida, & Yoshimoto, 2002). Their model predicts that the inverse temperature encoded by noradrenaline is reduced when the action value function $Q(s,a)$ has a high variance for a given state (Fig. 9(f)).

4.4. Acetylcholine as learning rate

A number of methods for automatically tuning the learning rate parameter have been proposed (Murata et al., 2002; Sutton, 1992). One of those, known as the delta-bar-delta method, detects oscillations in the error signal, which means that the setting of the learning rate is too large. According to such a regulatory mechanism, frequent change in the direction of the TD error encoded as the dopaminergic activity would have an inhibitory effect on the learning rate represented by the cholinergic system (Fig. 9(g)).

5. Conclusion

This paper proposed a unified theory on the roles of neuromodulators in mediating the global learning signal and metaparameters of distributed learning mechanisms of the brain. We considered how such regulatory mechanisms can be implemented in the neural circuit centered around the basal ganglia. However, there are many other brain areas and functions that require further consideration, for example, the roles of the amygdala and hippocampus in reinforcement learning and the roles of neuromodulators in sensory processing. As we discussed in reference to the role of serotonin, the same global signal should have different effects on different neurons and circuits, depending on their particular functions. This may be one of the reasons for the variety of receptors and intracellular signaling mechanisms of neuromodulators (Marder & Thirumalai, 2002).

The proposed hypotheses enabled us to make specific predictions about the effects of the changes in a particular neuromodulatory system on the behavior of the animal, the dynamics of neural circuits, and the activity of other neuromodulators. The hypotheses also lead us to specific predictions as to how those neuromodulatory systems should respond to changes in the environment and the process of learning of the animal. Experimental tests of these predictions may force us to revise this simple theory, but it can nevertheless be helpful in delineating the complex functions of neuromodulators.

The neuromodulatory systems are regarded as biophysical substrates of motivation, emotional states, and personalities. The computational models of the roles of neuromodulators, as proposed in this and other papers in this special issue, could provide the theoretical basis for better understanding the mechanism of emotion, the appropriate therapy for psychiatric and behavioral diseases, and the design of more 'human-like' artificial agents.

Acknowledgments

The author is grateful to Peter Dayan, Barry Everitt, Takeshi Inoue, Sham Kakade, Go Okada, Yasumasa Okamoto, and Shigeto Yamawaki for valuable discussions on the roles of serotonin and also thanks Nicolas Schweighofer for his comments on the manuscript.

References

Aosaki, T., Tsubokawa, H., Watanabe, K., Graybiel, A. M., & Kimura, M. (1994). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensory-motor conditioning. *Journal of Neuroscience*, *14*, 3969–3984.

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience*, *14*, 4467–4480.

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

Barto, A. G. (1995b). Reinforcement learning. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 804–809). Cambridge, MA: MIT Press.

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

Baxter, J., & Bartlett, P. L (2000). Reinforcement learning in POMDP's via direct gradient ascent. *International Conference on Machine Learning*.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, *19*, 10502–10511.

Buhot, M. C. (1997). Serotonin receptors in cognitive behaviors. *Current Opinion in Neurobiology*, *7*, 243–254.

Cardinal, R. N., Pennicott, D. R., Sugathapala, C. L., Robbins, T. W., & Everitt, B. J. (2001). Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science*, *292*, 2499–2501.

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

Dayan, P., & Yu, A. J. (2002). *ACh, uncertainty, and cortical inference* (*vol. 14*). *Advances in neural information processing systems*, Cambridge, MA: MIT Press.

De Deurwaerde, P., Stinus, L., & Sampinato, U. (1998). Opposite changes of in vivo dopamine release in the rat nucleus accumbens and striatum that follows electrical stimulation of dorsal raphe nucleus: Role of 5-HT3 receptors. *Journal of Neuroscience*, *18*, 6528–6538.

Deakin, J. F. W., & Graeff, F. G. (1991). 5-HT and mechanisms of defense. *Journal of Psychopharmacology*, *5*, 305–315.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, *12*, 961–974.

Doya, K. (2000a). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739.

Doya, K. (2000b). Metalearning, neuromodulation, and emotion. In G. Hatano, N. Okada, & H. Tanabe (Eds.), *Affective minds* (pp. 101–104). Amsterdam: Elsevier.

Doya, K. (2000c). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 215–245.

Doya, K., Kimura, H., & Kawato, M. (2001). Computational approaches to neural mechanism of learning and control. *IEEE Control Systems Magazine*, *21*(4), 42–54.

Durstewitz, D., & Seamans, J. (2002). The computational role of dopamine D1 receptors in working memory. *Neural Networks*, *15*(4–6), 561–572.

Fletcher, P. J. (1995). Effects of combined or separate 5,7-dihydroxytryptamine lesions of thee dorsal and median raphe nuclei on responding maintained by a DRL 20s schedule of food reinforcement. *Brain Research*, *675*, 45–54.

Gilzenrat, M. S., Holmes, B. D., Rajkowski, J., Aston-Jones, G., & Cohen, J. D. (2002). Simplified dynamics in a model of noradrenergic modulation of cognitive performance. *Neural Networks*, *15*(4–6), 647–663.

Gullapalli, V. (1990). A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks*, *3*, 671–692.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neurosciences*, *16*, 218–222.

Hasselmo, M. E., & Schnell, E. (1994). Acetylcholine and memorylaminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modeling and brain slice physiology. *Journal of Neuroscience*, *14*, 3898–3914.

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

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

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation–exploration meta-parameter in reinforcement learning. *Neural Networks*, *15*(4–6), 665–687.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. Dopaminergic modulation in the basal ganglia. *Neural Networks*, *15*(4–6), 535–547.

Kakade, S. (2001). Optimizing average reward using discounted rewards. *Computational Learning*.

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

Katz, P. S. (1999). *Beyond neurotransmission: Neuromodulation and its importance for information processing*. Oxford, UK: Oxford University.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*, 718–727.

Littman, M., Cassandra, A., & Kaelbling, L. (1995). Learning policies for partially observable environments: Scaling up. In A. Prieditis, & S. Russel (Eds.), *Machine learning: Proceedings of the 12th International Conference* (pp. 362–370). San Francisco, CA: Morgan Kaufmann.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, *22*, 159–196.

Marder, E., & Thirumalai, V. (2002). Cellular, synaptic, and network effects of neuromodulators. *Neural Networks*, *15*(4–6), 479–493.

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences, USA*, *88*, 4433–4437.

Middleton, F. A., & Strick, P. F. (2000). Basal gagnlia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews*, *31*, 236–250.

Mobini, S., Chiang, T. J., Ho, M. Y., Bradshaw, C. M., & Szabadi, E. (2000). Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology*, *152*, 390–397.

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

Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*, 37–51.

Murata, N., Kawanabe, M., Ziehe, A., Muller, K.-R., & Amari, S. (2002). On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, *15*(4–6), 743–759.

Neal, R. M. (1996). *Bayesian learning for neural networks*. New York: Springer.

Paladini, C. A., Celada, P., & Tepper, J. M. (1999). Striatal, pallidal, and parsreticulata evoked inhibition of nigrostriatal dopaminergic neurons is mediated by GABA A receptors in vivo. *Neuroscience*, *89*, 799–812.

Parsons, L. H., Weiss, F., & Koob, G. F. (1998). Serotonin 1B receptor stimulation enhances cocaine reinforcement. *Journal of Neuroscience*, *18*, 10078–10089.

Partridge, J. G., Apparsundaram, S., Gerhardt, G. A., Ronesi, J., & Lovinger, D. M. (2002). Nicotinic acetylcholine receptors interact with dopamine in induction of striatal long-term depression. *Journal of Neuroscience*, *22*, 2541–2549.

Perry, E., Walker, M., Grance, J., & Perry, R. (1999). Acetylcholine in mind: A neurotransmitter correlate of consciousness? *Trends in Neurosciences*, *22*, 273–280.

Rahman, S., Sahakian, B. J., Cardinal, R. N., Rogers, R. D., & Robbins, T. W. (2001). Decision making and neuropsychiatry. *Trends in Cognitive Sciences*, *5*, 271–277.

Rasmusson, D. D. (2000). The role of acetylcholine in cortical synaptic plasticity. *Behavioural Brain Research*, *115*, 205–218.

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

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

Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of cortico-striatal synapses. *Neural Networks*, *15*(4–6), 507–521.

Robbins, T. W. (1997). Arousal systems and attentional processes. *Biological Psychology*, *45*, 57–71.

Robbins, T. W. (2000). Chemical neuromodulation of frontal-executive functions in humans and other animals. *Experimental Brain Research*, *133*, 130–138.

Rocha, B. A., Fumagalli, F., Gainetdinov, R. R., Jones, S. R., Ator, R., Giros, B., Miller, G. W., & Caron, M. G. (1998a). Cocaine self-administration in dopamine-transporter knockout mice. *Nature Neuroscience*, *1*, 132–137.

Rocha, B. A., Scearce-Levie, K., Lucas, J. J., Hiroi, N., Castanon, N., Crabbe, J. C., Nestler, E. J., & Hen, R. (1998b). Increased vulnerability to cocaine in mice lacking the serotonin-1B receptor. *Nature*, *393*, 175–178.

Russell, V. A., Allin, R., Lamm, M. C., & Taljaard, J. J. (1992). Regional distribution of monoamines and dopamine D1-and D2-receptors in the striatum of the rat. *Neurochemical Research*, *17*, 387–395.

Saper, C. B. (2000). Brain stem modulation of sensation, movement and consciousness. In E. Kandel, J. H. Schwartz, & T. M. Jessel (Eds.), *Principles of Neural Science* (4th ed) (pp. 889–909). New York: McGraw-Hill.

Sawaguchi, T., & 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. *Journal of Neurophysiology*, *71*, 515–528.

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

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

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

Sershen, H., Hashim, A., & Lajtha, A. (2000). Serotonin-mediated striatal dopamine release involves the dopamine uptake site and the serotonin receptor. *Brain Research Bulletin*, *53*, 353–357.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, *249*, 892–895.

Shimo, Y., & Hikosaka, O. (2001). Role of tonically active neurons in primate caudate in reward-oriented saccadic eye movement. *Journal of Neuroscience*, *21*, 7804–7814.

Singh, S., & Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), (*Vol. 9*) (pp. 974–980). *Advances in neural information processing systems*, Cambridge, MA: MIT Press.

Singh, S., Jaakkola, T., Littman, M. L., & Szpesvari, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, *38*, 287.

Sora, I., Hall, F. S., Andrews, A. M., Itokawa, M., Li, X. F., Wei, H. B., Wichems, C., Lesch, K. P., Murphy, D. L., & Uhl, G. R. (2001). Molecular mechanisms of cocaine reward: Combined dopamine and serotonin transporter knockouts eliminate cocaine place preference. *Proceedings of National Academy of Science, USA*, *98*, 5300–5305.

Suri, R. E. (2002). Td models of reward predictive responses in dopamine neurons. *Neural Networks*, *15*(4–6), 523–533.

Sutton, R. S. (1992). *Adapting bias by gradient descent: An incremental version of delta-bar-delta. Proceedings of AAAI 92*, Cambridge, MA: MIT Press, pp. 171–176.

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

Tepper, J. M., Martin, L. P., & Anderson, D. R. (1995). Gabaa receptor-mediated inhibition of rat substantia nigra dopaminergic neurons by pars reticulata projection neurons. *Journal of Neuroscience*, *15*, 3092–3103.

Tesauro, G. (1994). TD-Gammon, a self teaching backgammon program, achieves master-level play. *Neural Computation*, *6*, 215–219.

Tremblay, L., & Schultz, W. (2000). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*, 1877–1885.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, *283*, 549–554.

Usher, M., & Davelaar, E. (2002). Neuromodulation of decision and response selection. *Neural Networks*, *15*(4–6), 635–645.

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed). New York: Springer.

Wang, M., & Licinio, J. (2001). Research and treatment approaches to depression. *Nature Reviews Neuroscience*, *2*, 343–351.

Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature*, *382*, 629–632.

White, F. J. (1998). Cocaine and the serotonin saga. *Nature*, *393*, 118–119.

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

Williams, G. V., Rao, S. G., & Goldman-Rakic, P. S. (2002). The physiological role of 5-HT2A receptors in working memory. *Journal of Neuroscience*, *22*, 2843–2854.

Wise, R. A. (1996). Neurobiology of addiction. *Current Opinion in Neurobiology*, *6*, 243–251.

Yu, A. J., & Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks*, *15*(4–6), 719–730.