CHAPTER

26

c00026

The Basal Ganglia and the Encoding of Value

Kenji Doya and Minoru Kimura

OUTLINE

INTRODUCTION

s0010

p0010 What mechanism of the brain underlies our flexible learning of choice behaviors? According to the theory of reinforcement learning (Sutton and Barto, 1998), an adaptive agent learns behaviors by repeating the following three steps:

o0010 1. Predicting the *value* of each action candidate, or option (while the term *option* is commonly used for choice candidates in economics, the term *action* is generally used in reinforcement learning literature, while option often means a higher-level choice of a series of actions)

o0020 2. Selecting an action with the highest predicted value

3. Updating the value of the action by the difference between the prediction and the actual outcome. o0030

p0050 How can these steps of valuation, action selection, and prediction-error based learning be realized in the brain? Reward-predictive neuron firing has been reported from variety of cortical and subcortical areas, such as the orbitofrontal cortex (Schultz and Dickinson, 2000), the prefrontal cortex (Watanabe, 1996), the parietal cortex (Dorris and Glimcher, 2004; Sugrue *et al*., 2004), and the striatum (Kawagoe *et al*., 1998). Neural firing proportional to reward-prediction error has been reported for midbrain dopamine neurons (Schultz *et al*., 1997; Satoh *et al*., 2003; Bayer and Glimcher, 2005). Functional brain-imaging experiments also report reward-predictive and

prediction-error related activities in these areas (O'Doherty *et al*., 2004; Haruno and Kawato, 2006). The striatum receives a rich dopaminergic input and the cortico-striatal synapses show dopamine-dependent plasticity (Wickens *et al*., 1996; Reynolds *et al*., 2001). Based on these observations, we have proposed a schematic framework of how the above three steps of computation can be realized in the cortico-basal ganglia circuit, as depicted in Figure 26.1 (Doya, 2000, 2002, 2007; Daw and Doya, 2006).

p0060 Here we further address the following questions:

o0040 1. How are values for different actions evaluated in different timescales represented in the striatum?
o0050 2. What is the role of the thalamic neurons in action selection?
o0060 3. How does the dopamine neuron guide learning of reward prediction?

p0100 Based on our own findings and those of others, we propose a more elaborate model of value estimation, action selection and learning in the cortico-basal ganglia circuit.

s0020 ACTION-VALUE CODING IN STRIATAL NEURONS

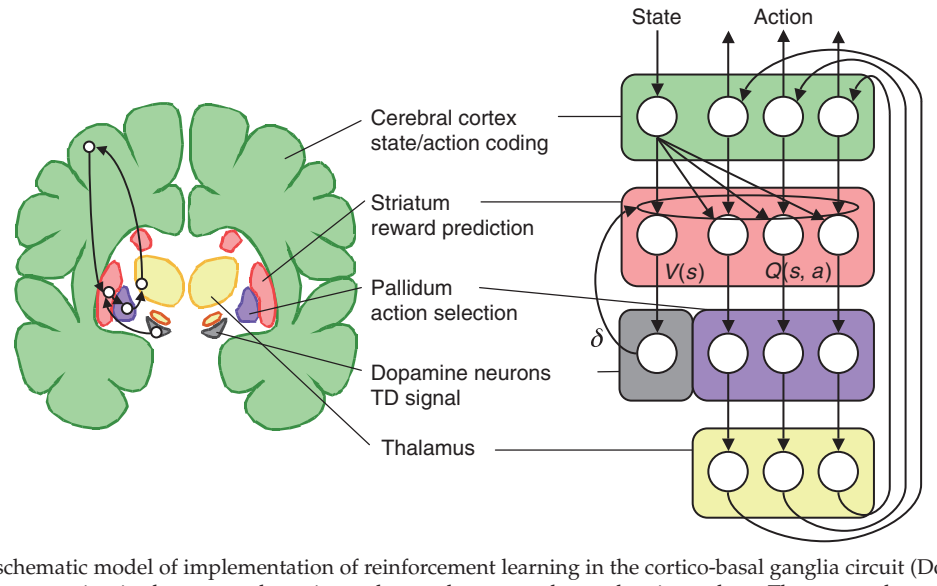
p0110 The most popular method in reinforcement learning is to learn the *action value*

$$Q(a) = E[r|a] \qquad (26.1)$$

which represents the reward r (amount $\times$ probability) expected when taking an action a. If the action values are learned for all possible actions, the obvious choice is to take the action a that gives the largest action value $Q(a)$. Does the brain use such a method and, if so, where in the brain are such action values represented?

p0120 In saccadic eye-movement experiments with monkeys, Hikosaka and colleagues showed that firing of striatal projection neurons before saccades was modulated by the amount of reward the monkey had learned to receive following a successful saccade to a given visual target (Kawagoe *et al*., 1998). Such action-specific, reward-predictive activities are reminiscent of the action value above. However, their saccade task did not involve a choice between multiple action candidates; only one target was presented in each trial. In order to test how striatal neurons are involved in a reward-based free-choice situation, we performed recording experiments in which a monkey chose one of two possible actions based on the varied probability of liquid rewards (Samejima *et al*., 2005).

p0130 Two macaque monkeys performed a reward-based, free-choice task of turning a handle to the left or the right. The monkeys held the handle in the center position, using their left hand, for a delay period of 1s, and then turned the handle to either the left (a = L) or the right (a = R). An LED on the selected side was illuminated in either green, notifying a large reward (0.2 ml water), or red, notifying a small reward (0.07 ml water). The probabilities of receiving a large reward



f0010 **FIGURE 26.1** (A schematic model of implementation of reinforcement learning in the cortico-basal ganglia circuit (Doya, 1999, 2000, 2007). Based on the state representation in the cortex, the striatum learns the state value and action values. The state-value coding striatal neurons project to dopamine neurons, which send the TD signal back to the striatum. The outputs of action-value coding striatal neurons channel through the globus gallidus and the thalamus, where stochastic action selection may be realized.

after turns to the left and to the right were fixed during a block of 30 to 150 trials, and varied between five different trial blocks. In the 90–50 block, for example, the probability of a large reward for a left turn was 90%, and for a right turn was 50%. In this case, by taking the small reward as the baseline ($r = 0$) and the large reward as unity ($r = 1$), the left action-value Q_L was 0.9 and the right action-value Q_R was 0.5. Four asymmetrically rewarded blocks, 90–50, 50–90, 50–10, and 10–50, and one symmetrically rewarded block, 50–50, were used. An important feature of this block design is that the neuronal activity related to action value can be dissociated from that related to action choice. Suppose a monkey chooses the action with the higher action-value after sufficient learning in a given trial block. While the monkey would be expected to prefer a left turn in both the 90–50 and 50–10 blocks, the action value Q_L for the left turn differs, at 0.9 and 0.5 respectively. Conversely, in the 90–50 and 10–50 blocks, while the monkey's choice behavior should be the opposite (i.e. a left turn in the former and a right turn in the latter), the action value Q_R remains the same at 0.5. Through analysis of choice and reward sequences of two monkeys, we verified that the action-value based model could predict their choice behavior very well.

p0140 We recorded 504 striatal projection neurons in the right putamen and caudate nucleus of the two monkeys. Here, we focus on the 142 neurons that displayed increased discharges during at least one task event, and had discharge rates higher than 1 spike/s during the delay period. We compared the average discharge rates during the delay period from two asymmetrically rewarded blocks. The comparison was based on the trials after the monkeys' choices had reached a "stationary phase" during each block, when the choice probability was biased toward the action with a higher reward probability in more than 70% of trials. In one-half of the neurons (72/142 in two monkeys), activity was modulated by either Q_L or Q_R. For example, the delay period discharge rate of some neurons was significantly higher in the 90–50 block than in the 10–50 block, but was not significantly different between the 50–10 and 50–90 blocks, for which the preferred actions differed. These neurons thus appear to encode the left action-value, Q_L, but not the action or choice itself. Other neurons showed significantly different discharge rates between the 50–10 block and the 50–90 block, but there was no significant difference between the 10–50 and 90–50 blocks. The firing rates of these neurons appear to code the right action-value, Q_R.

p0150 Through a multiple regression analysis of neuronal discharge rates with Q_L and Q_R as regressors, we found 24 (17%) "Q_L-type" neurons that had a significant regression coefficient to Q_L (t-test, $P < 0.05$) but not to Q_R, 31 (22%) "Q_R-type" neurons that correlated to Q_R but not to Q_L, and 16 (11%) differential action-value ("Δ-Q-type") neurons that correlated with the difference between Q_L and Q_R. One neuron (<1%) had significant coefficients that correlated to both Q_L and Q_R with the same sign. There were 18 motor-related ("m-type") neurons that had significant t-values only for the action being chosen. However, the discharge rates of most action-value neurons (19/24 in the Q_L-type, 24/31 in the Q_R-type) were not correlated significantly with the action being chosen. We concluded that, during a delay period before action execution, more than one-third of striate projection neurons examined (43/142) encoded action values, and that 60% (43/72) of all the reward value-sensitive neurons were action-value neurons.

p0160 Action-value coding in the striatum may be a core feature of information-processing in the basal ganglia. The striatum is the primary target of dopaminergic signals which regulate the plasticity (change in the strength) of cortico-striatal synaptic transmission (Calabresi *et al*., 1996; Reynolds *et al*., 2001), conveying signals of actions and cognition. Thus, the striatum may be the locus where reward value is first encoded in the brain.

p0170 A recent experiment compared action-value and chosen-action representations in the dorsal striatum and the internal globus pallidus (Pasquereau *et al*., 2007). These authors found action-value coding in both the striatum and the globus pallidus, but the number of neurons encoding the chosen action increased in the globus pallidus as each trial progressed toward the time of action initiation. This finding, along with our finding of relatively few action-coding neurons in the striatum, favors the view that action selection is realized downstream of the basal ganglia (Doya, 2000; Watanabe *et al*., 2003) rather than in the striatum itself (Houk *et al*., 1995; O'Doherty *et al*., 2004). Further studies on the neuronal activity before and after action selection from different stages of the cortico-basal ganglia loop are necessary to clarify where and how action selection is realized.

SHORT- AND LONG-TERM REWARD PREDICTION IN THE STRIATUM

s0030

p0180 In the previous section, we considered the case where a reward is given immediately after each action choice. In a more general scenario, an action can result in a reward after various delays, and thus it is not necessarily obvious which of the previous actions

is responsible for a given reward. A common way of resolving such a *temporal credit assignment problem* is to learn to predict the cumulative future rewards in the form of the *state value*

$$V(s) = E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots | s(t) = s] \quad (26.2)$$

where state s is defined by a sensory cue (or any other information that is useful for predicting the future outcome), E[] indicates the mean expected value under the current policy (state-to-action mapping), and γ is a parameter called the *temporal discount factor*. The state value is a measure of the long-term goodness of the given state s under the current policy; thus the increase or decrease in $V(s)$ can be regarded as a virtual reward signal. More precisely, the inconsistency in the prediction

$$\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t)) \quad (26.3)$$

is termed the *temporal-difference (TD) error*, and this can be utilized as the effective reward signal that takes into account the delayed rewards (Sutton and Barto, 1998).

p0190 The discount factor γ defines the temporal focus of this cumulative reward prediction: if $\gamma = 0$, only the immediate reward $r(t)$ is considered; if γ is set close to 1, long-delayed rewards are also taken into account. In essence, γ controls the temporal horizon of future reward estimation. The temporal discount factor γ is a critical parameter that determines the character of learned behaviors (Doya, 2002) – for example, a low setting of γ can lead to short-sighted, impulsive behaviors. In order to understand the brain's mechanism for reward prediction at different timescales, and its potential mechanism of regulation, we performed an fMRI experiment in which subjects learned to take small losses in order to acquire subsequent large rewards (Tanaka *et al*., 2004).

p0200 In the *Markov decision task* (Figure 26.2a), one of three states is presented to the subject visually, using three different figures, and the subject selects one of two actions by pressing one of two buttons. For each state, the subject's action affects not only the reward given immediately but also the state subsequently presented. In the SHORT condition, action a_1 gives a small positive reward $+r_1$ (20 yen average) and action a_2 gives a small negative reward $-r_1$ (-20 yen average) at all three states. The optimal behavior for maximizing the total outcomes is to collect small positive rewards by taking action a_1 at each state. In the LONG condition, while action a_2 at state s_3 gives a big bonus $+r_2$ (100 yen average), action a_1 at state s_1 results in a big loss $-r_2$ (-100 yen average). The optimal behavior is to receive small negative rewards to obtain a large positive reward by taking action a_2 at each state – the opposite to the optimal behavior in the SHORT condition. Thus, in the LONG condition, the subject has to select an action by taking into account both the immediate reward and the future reward expected from the subsequent state, while in the SHORT condition the subject needs to consider only the immediate outcome. In order to remove those vision and motor-related brain activities independent of reward processing, we introduced the NO reward condition, in which the reward was always zero and the subject was free to choose any button.

p0210 We first performed block-design analyses to assess the brain areas specifically involved in short- and long-term reward prediction. In the statistical comparison of brain activities during the SHORT vs NO conditions, a significant increase in activity was observed in the lateral orbitofrontal cortex (OFC), the insula, and the occipitotemporal area (OTA), as well as in the striatum, globus pallidus (GP), and medial cerebellum. These areas may be involved in reward prediction that only takes into account immediate outcome. In the LONG vs SHORT contrast, a robust increase in activity was observed in the ventrolateral PFC (VLPFC), insula, dorsolateral prefrontal cortex (DLPFC), dorsal premotor cortex (PMd), inferior parietal cortex (IPC), striatum, globus pallidus, dorsal raphe nucleus, lateral cerebellum, posterior cingulate cortex, and subthalamic nucleus. These areas are specifically involved in decision-making based on prediction of reward in multiple steps in the future, which was specifically required in the LONG condition but not in the SHORT condition. The results of these block-design analyses suggest differential involvement of brain areas in predicting immediate and future rewards. These results are also consistent with a more recent fMRI study using an inter-temporal choice task, which found activities in the lateral prefrontal and parietal cortex for delayed reward choice (McClure *et al*., 2004).

p0220 In order to further clarify the brain structures specific to reward prediction at different timescales, we estimated how much reward the subjects should have predicted on the basis of their behavioral data. We then used these trial-by-trial predictions to construct time-courses as the explanatory variables for a regression analysis. Specifically, we estimated the time-courses of reward prediction $V(s(t))$ and prediction error $\delta(t)$, as defined in equations (26.2) and (26.3), from each subject's performance data. In our Markov decision task, the minimum value of γ needed to find the optimal action in the LONG condition was 0.36, while any small value of γ was sufficient in the

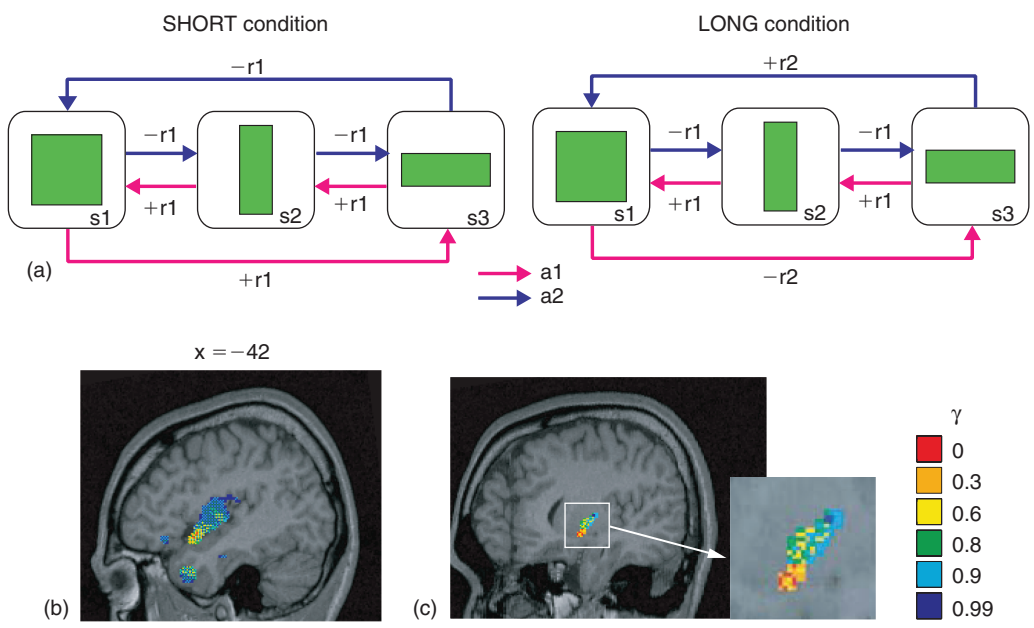


FIGURE 26.2 Experimental design of the Markov decision task. (a) At the beginning of each trial block, the condition is informed by displaying its character (e.g. "SHORT condition"). A fixation point is presented on the screen, and after 2 seconds one of three figures (square, vertical rectangle, or horizontal rectangle) is presented. As the fixation point vanishes after 1 s, the subject presses either the right or left button within 1 s. After a short delay (1 s), a reward for the current action is presented by a number, and the past cumulative reward is shown by a bar graph. A single trial takes 6 seconds. The rules of the reward and state transition for actions a_1 (red arrow) and a_2 (blue arrow) are shown in the SHORT and LONG conditions. The small reward r_1 is 10, 20, or 30 yen, with equal probability, and the large reward r_2 is 90, 100, or 110 yen. The rule of state transition is the same for all conditions; $s_3 \rightarrow s_2 \rightarrow s_1 \rightarrow s_3 \ldots$ for action a_1, and $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_1 \ldots$ for action a_2. Although the optimal behaviors are opposing (action a_1 in the SHORT condition and action a_2 in the LONG condition), the expected cumulative reward during one cycle of the optimal behavior is 60 yen in both the SHORT ($+20 \times 3$) and the LONG ($-20 - 20 + 100$) conditions. (b) Voxels with a significant correlation with reward prediction $V(t)$ in the insular cortex (height threshold of $P < 0.001$, uncorrected; extent threshold 4 voxels). (c) Voxels with a significant correlation with reward-prediction error $\delta(t)$ in the striatum. Different colors are used for different settings of the discount factor: $\gamma = 0$, red; $\gamma = 0.3$, orange; $\gamma = 0.6$, yellow; $\gamma = 0.8$, green; $\gamma = 0.9$, cyan; $\gamma = 0.99$, blue). Note the ventromedial to dorsolateral gradient with the increase in γ in both the insula and the striatum. f0020

SHORT condition. We hypothesized that different parts of the brain are specialized for reward prediction at different timescales. Accordingly, we estimated $V(s(t))$ and $\delta(t)$ using six different levels of the discount factor ($\gamma = 0$, 0.3, 0.6, 0.8, 0.9, and 0.99), and searched for voxels that had significantly correlated time-courses to those explanatory variables.

p0230 We observed a significant correlation with reward prediction $V(s(t))$ in the medial prefrontal cortex (mPFC) and the bilateral insula (Figure 26.2b), the left hippocampus, and the left temporal pole (the foremost part of the temporal lobe). The activities of the medial prefrontal cortex (mPFC), temporal pole, and hippocampus correlated with reward prediction $V(s(t))$ with a longer timescale ($\gamma = 0.6$). Furthermore, as shown in Figure 26.2b, using graded colors for different discount factors γ (red for $\gamma = 0$, blue for $\gamma = 0.99$), a graded map was produced showing activities for reward prediction at different timescales in the insula. While the activity in the ventromedial part correlated with reward prediction at a shorter timescale, the activity of the dorsolateral part correlated with reward prediction at a longer timescale. We also found significant correlation with reward-prediction error $\delta(t)$ with a wide range of timescales in the basal ganglia (Figure 26.2c). Again, a graded map was produced, which had a short timescale in the ventromedial part and a long timescale in the dorsolateral part.

p0240 The results of the block-design and performance-based regressor analyses suggest differential involvement of brain areas in action learning by prediction of rewards at different timescales. In the insula and the anterior striatum, activities were found both in block-design and performance-based regression analyses. The vertical anatomical shifts in the activated locus in the SHORT vs NO and LONG vs SHORT contrasts in each area are consistent with the ventro-dorsal maps of the discount factor γ found in the performance-based regression analysis. Correlation of the striatal activity with reward-prediction error $\delta(t)$ could be due to dopamine-dependent plasticity of cortico-striatal synapses (Reynolds and Wickens, 2002).

p0250 In summary, compared with the control task in which subjects simply learned to acquire immediate positive rewards, we found enhanced activity in the prefrontal, premotor, and parietal cortices, as well as in the dorsal striatum, lateral cerebellum, and the midbrain including the dorsal raphe nucleus. By reinforcement-learning model-based analysis using multiple temporal discounting parameters, we found a ventral-to-dorsal map of short-to-long timescales of reward prediction in the striatum and the insular cortex.

s0040 ## CENTROMEDIAN THALAMIC NEURONS

p0260 As we mentioned earlier in the chapter, the site and the mechanism of value-based action selection are still to be investigated. Thus we have examined the roles of the thalamo-striate projection in reward value-based action selection in the basal ganglia. The centromedian/parafascicular (CM/PF) complex of the thalamus (Steriade *et al*., 1997) has received little attention in the studies of action and cognition; however, its outputs direct mostly to the putamen and caudate nucleus as well as to the medial frontal cortex, and it receives topographically organized inputs from the output stations of the basal ganglia as well as from the reticular formation, superior colliculus, and pedunculo-pontine tegmental nucleus (Groenewegen and Berendse, 1994; Steriade *et al*., 1997; Matsumoto *et al*., 2001; Takada *et al*., 2001; Smith *et al*., 2004). There are two representative types of neurons in the CM/PF complex: one exhibits increase of their discharges after visual, auditory, and somatosensory stimuli at very short latency (SLF), while the other exhibits facilitatory responses at long latency (>200 ms, LLF) (Minamimoto and Kimura, 2002; Minamimoto *et al*., 2005). SLF neurons are mostly located in the PF, while LLF neurons are mostly located in the CM (Matsumoto *et al*., 2001).

p0270 We recorded from LLF neurons in the CM in an asymmetrically rewarded GO/NO-GO task in which two kinds of visual stimuli were presented, one for the GO response and the other for the NO-GO response (Minamimoto *et al*., 2005). Performance of the requested action, whether GO or NO-GO, was rewarded by a large amount of water, while performance of the other action was rewarded with a small amount of water. The action–outcome association was then altered in the next block. Monkeys performed the large-reward GO trials with shorter reaction times than they did the small-reward trials. This indicated that the monkeys assigned higher values to the large-reward actions and prepared for them (biased toward the actions). In addition, the rate of error trials, such as too long reaction times or initiation of incorrect actions, was higher in the small-reward trials.

p0280 Remarkably, the majority of the LLF neurons exhibited burst discharges selectively after the visual cue indicated that the small-reward option would be required, while they showed very little activity after of the visual cue indicated that the large-reward action could be required. This was true regardless of whether the action required was GO or NO-GO. The magnitude of the activity following small-reward action cues became larger, across trials, when the probability of a large-reward action cue increased. Thus, the critical nature of the activity was its specificity to the small-reward option among available actions when subjects were preparing to choose a large-reward option. These results raised the question of whether artificial activation of CM after a GO request could trigger the complementary process to the GO-action bias. Indeed, we found that electrical stimulation of CM after a large-reward GO request significantly slowed down the behavioral response, just as after a GO small-reward request. These results suggest a specific participation of CM in abolishing bias towards the large-reward option, and in pursuing the small-reward action, which are complementary processes to the response bias.

p0290 In this experimental situation, subjects had to complete small-reward trials in order to move on, and to obtain large rewards in later trials. Thus, when a small-reward instruction occurred a motivational conflict arose, because subjects had to perform voluntarily an option they did not want to choose. Response bias is not responsible for this situation, but another mechanism is necessary to cancel any premature motor programming through inhibiting the processes of response bias and reinforcing lower-valued but required action.

p0300 A recent study supported the involvement of the striato-pallidal system in this process by demonstrating that neurons encoding the values of chosen actions are more dominant and prevalent in the globus pallidus than in the striatum, where neurons dominantly encoding the expected reward probabilities of specific options of action (action values) are prevalent, irrespective of whether the action is to be selected or not (Pasquereau *et al*., 2007). A subset of neurons in the medial frontal cortex was selectively activated when saccadic eye movements which had been repeatedly triggered based on some rule to one of two previously rewarded directions were switched to the other direction after noticing reversal of action–outcome associations (Isoda and Hikosaka, 2007). The neuronal activation began and terminated mostly before the

movement onset in trials in which the monkey was asked to switch direction. Neurons in the cingulate motor area were activated selectively when subjects made switches from previously rewarded (but currently unrewarded) actions to previously unrewarded, currently rewarded actions. Further, this action switch was impaired by chemical inactivation of the cingulate motor area (Shima and Tanji, 1998). Therefore, the next critical issue for the studies of reward value-based decision-making, action selection, and monitoring and evaluation of ongoing or performed actions is to understand the roles played specifically by the striatum, striato-pallidal system, thalamo-striatal system, and medial frontal cortical areas.

s0050

MOTIVATION AND OUTCOME CODING IN DOPAMINE NEURONS

p0310

Dopamine (DA) neurons in the midbrain have been shown to encode prediction errors of probability and/or of magnitude of reward (Schultz, 1998; Satoh *et al*., 2003; Morris *et al*., 2004; Nakahara *et al*., 2004) or timing of expected reward (Bayer and Glimcher, 2005) through single-neuron recording experiments in monkeys performing behavioral tasks. On the other hand, a substantial body of evidence suggests involvement of DA systems in the processes of motivation (Robbins and Everitt, 1996; Koepp *et al*., 1998; Salamone and Correa, 2002; Wise, 2002), and in switching attentional and behavioral selections to salient stimuli that underlie associative learning (Spanagel and Weiss, 1999; Redgrave *et al*., 1999). It has been well documented that DA neurons show phasic activations with a wide variety of salient stimuli, including novel and high-intensity stimuli (Jacobs, 1986; Schultz and Romo, 1987; Ljungberg *et al*., 1992; Horvitz *et al*., 1997). Why do DA neurons encode reward expectation errors and motivation, and how are these signals integrated with the processes of decision-making and learning? We addressed these issues by examining the activity of DA neurons of monkeys that made a series of behavioral decisions based on trial-specific reward expectations (Satoh *et al*., 2003).

p0320

In a multi-step decision and action selection task, Japanese Monkeys depressed a start button with their hand after the button was illuminated. Three target buttons were then simultaneously switched on, and the monkeys released the start button and depressed one of the illuminated target buttons. If an incorrect button was depressed, a low-frequency acoustic tone occurred, and the monkeys chose one of the remaining two buttons in the next trial. If the correct button was depressed, a high-frequency tone occurred, and a small amount of reward water was delivered. The high- and low-tone sounds served as positive and negative reinforcers, respectively, following the behavioral decisions. Once the monkeys had identified the correct button, the same button was depressed correctly button in all subsequent trials. The monkeys received a reward three times, by selecting the same button during three consecutive trials. The trials in a single block were therefore divided into two time periods: the trial-and-error epoch and the repetition epoch. Five types of trials occurred – trials in which the monkeys chose the correct button at the first, second, or third choice in a single block (N1, N2, and N3, respectively) during the trial-and-error epoch, and at the first and the second trials during the repetition epoch (R1 and R2, respectively). Average probabilities of correct choices in N1, N2, N3, R1, and R2 trials were 20%, 50%, 85%, 93%, and 95%, respectively.

p0330

Following a few months of trials, the monkeys were found to be performing N1 trials with the longest reaction times and R2 trials with the shortest reaction times, among the five types of trials, after the start button had been illuminated. This suggested that the monkeys had developed trial type-specific levels of reward expectations. Furthermore, errors of reward prediction could be estimated – 80% in N1, 50% in N2, 15% in N3, 7% in R1 and 5% in R2 trials.

p0340

During these trials, DA neurons exhibited a maximal increase of discharge rate above their background level of 4–5 spikes/s after a positive reinforcer during N1 trials with a reward probability of 20%. Responses during N2 and N3 trials gradually decreased, and those during R1 and R2 trials were so small that it was difficult to detect increases of discharge rates above the background level. It was found that the magnitude of DA neuron responses reflected precisely the errors of reward expectation, as shown by an excellent fit of estimated values of reward expectation errors (REEs) to the DA responses. On the other hand, decrease of discharge rate of DA neurons after negative reinforcers exhibited a similar tendency – small responses in N1 trials and large responses in R1 and R2 trials – but the estimation of DA neuron responses by negative REEs was poor.

p0350

About a half of DA neurons also exhibited trial type-dependent activation after appearance of the start cue – the lowest in N1 trials and the highest in N2 or N3 trials. Thus, the dopamine neuron responses to the visual cue instructing the start of each trial had a tendency to be small when the reward probability of the trial was low, but large when the reward probability was high. However, the responses in the trials with the highest reward probability (>97% in R1 and/or

R2 trials) were smaller than those in either N2 or N3 trials. This was not consistent with previous observations of DA neuron responses of monkeys performing classical conditioning (Fiorillo *et al*., 2003) and instrumental conditioning paradigms (Morris *et al*., 2004, 2006). Probably, this was because subjects in our multi-step decision and action selection task expected rewards not only immediately after current trials but also after multi-step trials; this was in contrast to the behavioral situations employed in previous studies, in which only reward probabilities at current trials were predicted. On the other hand, the start cue responses could reflect the level of motivation, or willingness, to start an individual trial in order to obtain an expected reward after the trial. Reaction times of subjects in performing a required action have been used as a measure of motivational level at individual trials (Shidara *et al*., 1998; Watanabe *et al*., 2001; Kobayashi *et al*., 2002; Takikawa *et al*., 2002). We recorded the reaction times, as well as DA neuron discharges, of monkeys from the appearance of the start cue to performance of the action. DA neuron responses to the start cue changed considerably depending on the subject's reaction time. This suggested that DA neuron responses to the start cue are modulated by motivation. Importantly, the magnitude of DA neuron responses to the start cue was positively correlated with the magnitude of responses to the high-frequency tone (positive reinforcer) that occurred after a correct choice was made.

p0360 What is the functional role of the dual coding of incentive attribution to the start cue and of REEs in reward-based decision-making and learning? One possible and fascinating role is modulation of the effectiveness of REEs as a teaching signal by a motivation. For instance, the rate of learning could be faster when animals are highly motivated because of stronger activation of DA neurons (and thus larger amount of DA release) and slower when they are less motivated, even at identical REEs, as a consequence of an action. This suggests a new and richer model for DA neurons as teaching signals in reinforcement learning than is currently proposed. It is also consistent with the theory of classical conditioning, in which the rate of learning is assumed to be influenced by factors such as attention or motivation (Rescorla and Wagner, 1972; Dickinson, 1980; Niv *et al*., 2007). From a computational point of view, involvement of motivational processes in instrumental conditioning has recently been emphasized, and a new model of reinforcement learning has been put forward in which DA neurons transmit both reward-expectation error and impact of motivation (Dayan and Balleine, 2002).

CONCLUSION s0060

p0370 Figure 26.3 is an augmented schematic diagram incorporating our findings that LLF neurons in CM encode and transmit signals of anti-bias on selection mainly to the striatum, and that DA neurons encode and transmit signals of REEs, reward value, and motivation to the striatum, where adaptive action-value coding occurs based on the DA neuron signals. Reward-predictive and motivational coding of the dopamine neurons, as well as the complementary

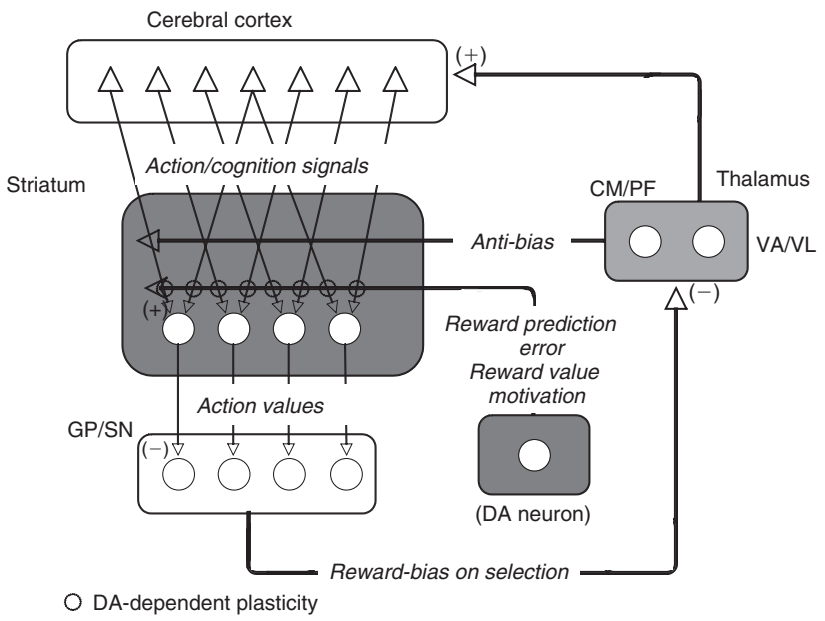


FIGURE 26.3 Schematic diagram showing how the basal ganglia encode reward values of external signals and actions, and how desirable actions are selected. The cortico-basal ganglia loop is composed of the cortico-basal ganglia-thalamo-cortical "external" loop, and the striato-pallido/nigro-thalamo-striate "internal" loop. GP/SN, globus pallidus and substantia nigra; CM/PF, centromedian parafascicular nuclei of intralaminar thalamus. f0030

activities of the CM thalamic neurons, may facilitate the acquisition of action-value coding of the striatal neurons. The mapping of multiple timescales in the ventro-dorsal axis of the cortico-striatal circuit may allow flexible decision-making in light of immediate and future costs and benefits.

References

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

Calabresi, P., Pisani, A., Mercuri, N.B., and Bernardi, G. (1996). The corticostriatal projection: from synaptic plasticity to dysfunctions of the basal ganglia. *Trends Neurosci.* 19, 19–24.

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

Dayan, P. and Balleine, B.W. (2002). Reward, motivation, and reinforcement learning. *Neuron* 36, 285–298.

Dickinson, A. (1980). *Contemporary Animal Learning Theory*. Cambridge: Cambridge University Press.

Dorris, M.C. and Glimcher, P.W. (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. (2007). Reinforcement learning: computational theory and biological mechanisms. *HFSP J.* 1, 30–40.

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

Groenewegen, H.J. and Berendse, H.W. (1994). The specificity of the "nonspecific" midline and intralaminar thalamic nuclei. *Trends Neurosci.* 17, 52–57.

Haruno, M. and Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks* 19, 1242–1254.

Horvitz, J.C., Stewart, T., and Jacobs, B.L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res.* 759, 251–258.

Houk, J.C., Adams, J.L., and 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, and D.G. Beiser (eds), *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press, pp. 249–270.

Isoda, M. and Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nat. Neurosci.* 10, 240–248.

Jacobs, B.L. (1986). Single unit activity of brain monoamine-containing neurons in freely moving animals. *Ann. NY Acad. Sci.* 473, 70–77.

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nat. Neurosci.* 1, 411–416.

Kobayashi, Y., Inoue, Y., Yamamoto, M. *et al*. (2002). Contribution of pedunculopontine tegmental nucleus neurons to performance of visually guided saccade tasks in monkeys. *J. Neurophysiol.* 88, 715–731.

Koepp, M.J., Gunn, R.N., Lawrence, A.D. *et al*. (1998). Evidence for striatal dopamine release during a video game. *Nature* 393, 266–268.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* 67, 145–163.

Matsumoto, N., Minamimoto, T., Graybiel, A.M., and Kimura, M. (2001). Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *J. Neurophysiol.* 85, 960–976.

McClure, S.M., Laibson, D.I., Loewenstein, G., and Cohen, J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science* 306, 503–507.

Minamimoto, T. and Kimura, M. (2002). Participation of the thalamic CM-Pf complex in attentional orienting. *J. Neurophysiol.* 87, 3090–3101.

Minamimoto, T., Hori, Y., and Kimura, M. (2005). Complementary process to response bias in the centromedian nucleus of the thalamus. *Science* 308, 1798–1801.

Morris, G., Arkadir, D., Nevet, A. *et al*. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43, 133–143.

Morris, G., Nevet, A., Arkadir, D. *et al*. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* 9, 1057–1063.

Nakahara, H., Itoh, H., Kawagoe, R. *et al*. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron* 41, 269–280.

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

O'Doherty, J., Dayan, P., Schultz, J. *et al*. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454.

Pasquereau, B., Nadjar, A., Arkadir, D. *et al*. (2007). Shaping of motor responses by incentive values through the basal ganglia. *J. Neurosci.* 27, 1176–1183.

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

Rescorla, R.A. and Wagner, A.R. (1972). Current research and theory. In: A.H. Black and W.F. Prokasy (eds), *Classical Conditioning*, Vol. II. New York, NY: Appleton Century Crofts, pp. 64–99.

Reynolds, J.N. and Wickens, J.R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* 15, 507–521.

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

Robbins, T.W. and Everitt, B.J. (1996). Neurobehavioural mechanisms of reward and motivation. *Curr. Opin. Neurobiol.* 6, 228–236.

Salamone, J.D. and Correa, M. (2002). Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behav. Brain Res.* 137, 3–25.

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

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.

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

Schultz, W. and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23, 473–500.

Schultz, W. and Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *J. Neurophysiol.* 57, 201–217.

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

Shidara, M., Aigner, T.G., and Richmond, B.J. (1998). Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J. Neurosci.* 18, 2613–2625.

Shima, K. and Tanji, J. (1998). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282, 1335–1338.

Smith, Y., Raju, D.V., Pare, J.F., and Sidibe, M. (2004). The thalamostriatal system: a highly specific network of the basal ganglia circuitry. *Trends Neurosci.* 27, 520–527.

Spanagel, R. and Weiss, F. (1999). The dopamine hypothesis of reward: past and current status. *Trends Neurosci.* 22, 521–527.

Steriade, M., Jones, E.G., and McCormick, C.D. (1997). *Organisation and Function*. Oxford: Elsevier Science.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science* 304, 1782–1787.

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

Takada, M., Tokuno, H., Hamada, I. *et al*. (2001). Organization of inputs from cingulate motor areas to basal ganglia in macaque monkey. *Eur. J. Neurosci.* 14, 1633–1650.

Takikawa, Y., Kawagoe, R., Itoh, H. *et al*. (2002). Modulation of saccadic eye movements by predicted reward outcome. *Exp. Brain Res.* 142, 284–291.

Tanaka, S.C., Doya, K., Okada, G. *et al*. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* 7, 887–893.

Watanabe, K., Lauwereyns, J., and Hikosaka, O. (2003). Neural correlates of rewarded and unrewarded eye movements in the primate caudate nucleus. *J. Neurosci.* 23, 10052–10057.

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

Watanabe, M., Cromwell, H.C., Tremblay, L. *et al*. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. *Exp. Brain Res.* 140, 511–518.

Wickens, J.R., Begg, A.J., and 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.

Wise, R.A. (2002). Brain reward circuitry: insights from unsensed incentives. *Neuron* 36, 229–240.