Behavioral/Systems/Cognitive

Role of Striatum in Updating Values of Chosen Actions

Hoseok Kim,[1] Jung Hoon Sul,[1] Namjung Huh,[1] Daeyeol Lee,[2] and Min Whan Jung[1]

[1]Neuroscience Laboratory, Institute for Medical Sciences, Ajou University School of Medicine, Suwon 443-721, Korea, and [2]Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut 06510

The striatum is thought to play a crucial role in value-based decision making. Although a large body of evidence suggests its involvement in action selection as well as action evaluation, underlying neural processes for these functions of the striatum are largely unknown. To obtain insights on this matter, we simultaneously recorded neuronal activity in the dorsal and ventral striatum of rats performing a dynamic two-armed bandit task, and examined temporal profiles of neural signals related to animal's choice, its outcome, and action value. Whereas significant neural signals for action value were found in both structures before animal's choice of action, signals related to the upcoming choice were relatively weak and began to emerge only in the dorsal striatum ~200 ms before the behavioral manifestation of the animal's choice. In contrast, once the animal revealed its choice, signals related to choice and its value increased steeply and persisted until the outcome of animal's choice was revealed, so that some neurons in both structures concurrently conveyed signals related to animal's choice, its outcome, and the value of chosen action. Thus, all the components necessary for updating values of chosen actions were available in the striatum. These results suggest that the striatum not only represents values associated with potential choices before animal's choice of action, but might also update the value of chosen action once its outcome is revealed. In contrast, action selection might take place elsewhere or in the dorsal striatum only immediately before its behavioral manifestation.

Introduction

A central issue in cognitive neuroscience is to understand the role of the basal ganglia (BG) in decision making. There are two major theories on this issue. One argues that the primary role of the BG is in action selection, whereas the other emphasizes its role in evaluating action outcomes. The action selection theory is supported by several lines of evidence. Anatomically, the BG receive converging inputs from virtually the entire cerebral cortex and send output projections back to the frontal cortex and subcortical motor structures (Alexander and Crutcher, 1990a; DeLong, 2000), suggesting a funneling function of the BG. Behaviorally, dysfunction of the BG can lead to various movement disorders (Albin et al., 1989) and lesions in the BG impair stimulus-response associations (Packard and Knowlton, 2002). Finally, physiological studies have found neuronal activity in the BG correlated with upcoming movement of the animal (Hikosaka et al., 1989; Alexander and Crutcher, 1990b; Apicella et al., 1992; Setlow et al., 2003; Nicola et al., 2004; Pasupathy and Miller, 2005). These features of the BG led to the proposal that its primary role is selecting a specific action out of multiple candidates that are provided by the cortex and relaying this information to the downstream motor structures (Alexander and Crutcher, 1990a; Hikosaka, 1994; Mink, 1996; Redgrave et al., 1999). Recently, a revised theory has been proposed that the BG contribute to action selection by biasing its activity toward an action with the most desirable outcome (Hollerman and Schultz, 1998; Kawagoe et al., 1998; Samejima et al., 2005; Hikosaka et al., 2006).

Evidence for the role of the BG in action evaluation is also widespread. The striatum is a major target of midbrain dopaminergic projections, which are known to carry reward prediction error (RPE) signals (Schultz, 1998), as well as projections from frontal cortical areas that carry signals related to the actions chosen by the animal and their values (Watanabe, 1996; Leon and Shadlen, 1999; Baeg et al., 2003; Barraclough et al., 2004). Hence, signals necessary to evaluate consequences of committed actions and update action values converge in the striatum. Neuroimaging studies have also found RPE and value signals in the striatum (for review, see O'Doherty, 2004; Delgado, 2007; Delgado et al., 2008), and behavioral studies have found impaired feedback-based learning, but intact non-feedback-based stimulus-response association, in Parkinson's disease patients (Shohamy et al., 2004).

Despite strong grounds for both selection and evaluation functions of the BG, their underlying neural processes are unclear. In fact, it still remains unresolved whether the primary role of the BG is in action selection, action evaluation, or both. We investigated these issues in this study by examining temporal dynamics of neural signals for animal's choice, its outcome, and action value in the striatum of rats performing a free-choice task. Because the dorsal and ventral striatum (DS and VS, respectively) have been proposed to serve different functions (DeLong, 2000; Cardinal et al., 2002; O'Doherty, 2004; Atallah et al., 2007; Balleine et al., 2007), we recorded single-neuron activity from both structures simultaneously.

Received June 10, 2009; revised Oct. 6, 2009; accepted Oct. 8, 2009.

This work was supported by a grant from Brain Research Center of the 21st Century Frontier Research Program, Korea Science and Engineering Foundation Grant R01-2008-000-10287-0, Korea Research Foundation Grant KRF-2008-314-H00006, and the Cognitive Neuroscience Program of the Korea Ministry of Science and Technology (M.W.J.).

Correspondence should be addressed to Min Whan Jung, Neuroscience Laboratory, Institute for Medical Sciences, Ajou University School of Medicine, Suwon 443-721, Korea. E-mail: min@ajou.ac.kr.

DOI:10.1523/JNEUROSCI.2728-09.2009

 0270-6474/09/2914701-12$15.00/0

Materials and Methods

Subjects

Experiments were performed with young male Sprague Dawley rats (~9–11 weeks old, 250–330 g, $n = 3$). Animals were individually housed in the colony room and initially allowed *ad libitum* access to food and water. Once behavioral training began, animals were restricted to 30 min access to water after finishing one behavioral session per day. Experiments were performed in the dark phase of 12 h light/dark cycle. The experimental protocol was approved by the Ethics Review Committee for Animal Experimentation of the Ajou University School of Medicine.

Behavioral task

Animals were trained in a dynamic two-armed bandit task on a modified figure 8-shaped maze (65 × 60 cm, width of track: 8 cm, 3-cm-high walls along the entire track except the central bridge) (Fig. 1*A*) (Huh et al., 2009). It was a free binary choice task with each choice associated with a different probability of reward that was constant within a block of trials, but changed across blocks. Although the probability of reward delivery was constant within a block, the reward was delivered stochastically in each trial, and no explicit sensory information on reward probability was available to animals. Hence, the probabilities of reward and the optimal choice could be discovered only by trial and error. Each animal was tested for a total of 4–18 sessions, and each session consisted of four blocks of trials. The number of trials in each block was 35 plus a random number drawn from a geometric distribution with a mean of 5, with the maximum number of trials set at 45. The following four combinations of reward probabilities were used in each session: 0.72:0.12, 0.63:0.21, 0.21:0.63, and 0.12:0.72. The sequence was determined randomly with the constraint that the richer alternative always changed its location at the beginning of a new block.

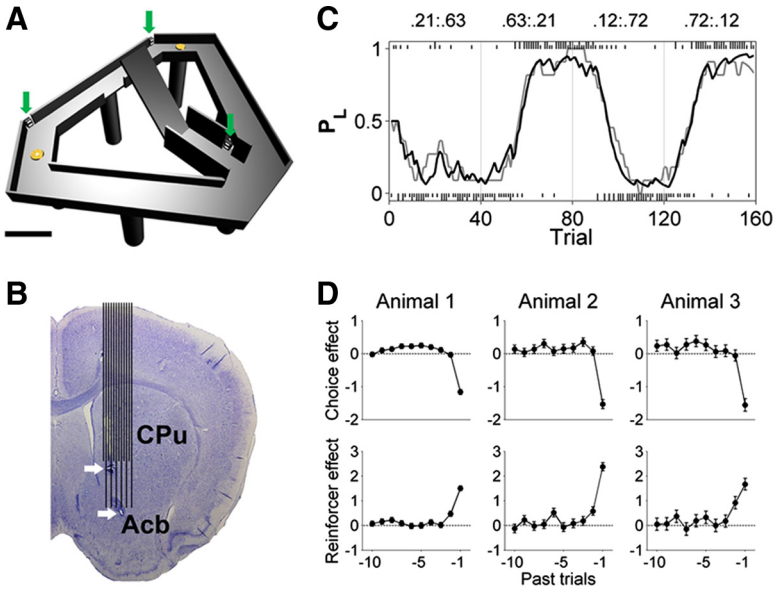

Figure 1. Behavioral task, recording sites and performance of animals. ***A***, Two-armed bandit task. Rats were tested on a modified figure 8-shaped maze to choose between two locations (yellow discs) that delivered water reward with different probabilities. Scale bar, 10 cm. Green arrows indicate the locations of photobeam detectors. ***B***, Unit signals were recorded in the DS and VS by implanting 12 tetrodes (schematically indicated by 12 vertical lines) in three rats. The photomicrograph shows a coronal section of the brain that was stained with cresyl violet. Two marking lesions, one in the DS and the other in the VS, are shown (white arrows). CPu, Caudate/putamen; Acb, nucleus accumbens. ***C***, An example of animal's choice behavior in a single behavioral session. The probability to choose the left arm (P_L) is plotted (moving average of 10 trials) across four blocks of trials (gray curve: actual choice of the animal, black curve: choice probability predicted by RL model). Tick marks denote trial-by-trial choices of the animal (upper, left choice; lower, right choice; long, rewarded trial; short, unrewarded trial). Block transitions are marked by vertical lines. Numbers on top indicate mean reward probabilities associated with left and right choices in each block. ***D***, Average regression coefficients from a logistic regression model showing the effects of past choices and rewards on animal's choice. The influence of past choices and past rewards (up to 10 trials) on the current choice was estimated by fitting a logistic regression model to the behavioral data for each animal. Error bars, 95% confidence intervals.

Behavioral stages

Each trial was divided into five stages, corresponding to (1) delay, (2) go, (3) approach to reward, (4) reward consumption, and (5) return to the center of the maze (Fig. 2*A*). Each trial began when the animal returned to the central section of the maze (Fig. 2*A*, region D) from either goal location (Fig. 2*A*, blue circles) with the connecting bridge elevated (Fig. 1*A*). After a delay for 3 s (delay stage), the connecting bridge was lowered allowing the animal to proceed to the upper branching point (go stage). The animal initiated locomotion as soon as the connecting bridge was lowered. To determine the onset of the approach stage, we first estimated, based on visual inspection, the vertical position in which the animal's horizontal position begins to diverge. Next, we aligned the animal's horizontal trajectory relative to the time when the animal reached this vertical position, and the onset of the approach stage was defined as the time when the animal's horizontal positions in the left-choice and right-choice trials became significantly different as determined by a *t* test at the significance level of 0.05 (Fig. 2*C*,*D*). Thus, the beginning of the approach stage was aligned to the first behavioral manifestation of animal's goal choice, which was determined separately for each behavioral session. The beginning of the reward stage was the time when the animal broke the photobeam that was placed 6 cm ahead of the water-delivery nozzle. In rewarded trials, breaking of the photobeam triggered opening of a solenoid value in ~20 ms, which delivered 30 μl of water in ~30 ms. It took ~200 ms for the animal to arrive at the water nozzle after breaking the photobeam. The outcome of the choice was revealed to the animal immediately after breaking of the photobeam, because opening of the solenoid valve produced a clicking sound (~50 db). The central connecting bridge was raised at the onset of the reward stage. The beginning of the return stage was the time when the animal crossed an invisible line 11 cm away from the water-delivery nozzle (i.e., exiting the reward area) (Fig. 2*A*). The durations of the reward stage in rewarded and unrewarded trials were 7.1 ± 0.8 and 2.5 ± 0.3 s, respectively. Hence, the animal stayed longer in the reward area when water was delivered. Nevertheless, the animals licked the water-delivery nozzle in most unrewarded trials as in rewarded trials. The beginning of the delay stage in the next trial was when the animal broke the central photobeam that was placed 13 cm from the proximal end of the maze (Fig. 2*A*). The average durations of the five behavioral stages were (mean ± SD) 3.0 ± 0.0 (delay), 1.2 ± 0.2 (go), 0.9 ± 0.3 (approach), 6.2 ± 3.0 (reward), and 2.7 ± 1.1 s (return). The delivery of water and the state (raised or lowered) of the central bridge were controlled by a personal computer using LabView software (National Instruments).

Unit recording

Activity of single units was recorded simultaneously from the left ($n = 1$) or right ($n = 2$) DS and VS (Fig. 1*B*). For the DS recording, unit activity was recorded from the dorsomedial striatum, and for the VS recording, activity was mostly recorded from the core of the nucleus accumbens. A microdrive array loaded with 12 tetrodes was lowered aiming the dorsomedial striatum (1.2 mm anterior, 1.7 mm lateral from bregma) with six tetrodes implanted in the DS [3.0 mm ventral (V) from the brain surface] and the other six implanted in the VS (6.0 mm V from the brain surface) under deep anesthesia with sodium pentobarbital (50 mg/kg body weight). Following 1 week of recovery from the surgery, tetrodes were gradually advanced for a maximum 320 μm per day for 2 d and then advanced 20–40 μm per day. The identity of unit signals was determined based on the clustering pattern of spike waveform parameters, averaged spike waveforms, baseline discharge frequencies, autocorrelograms, and

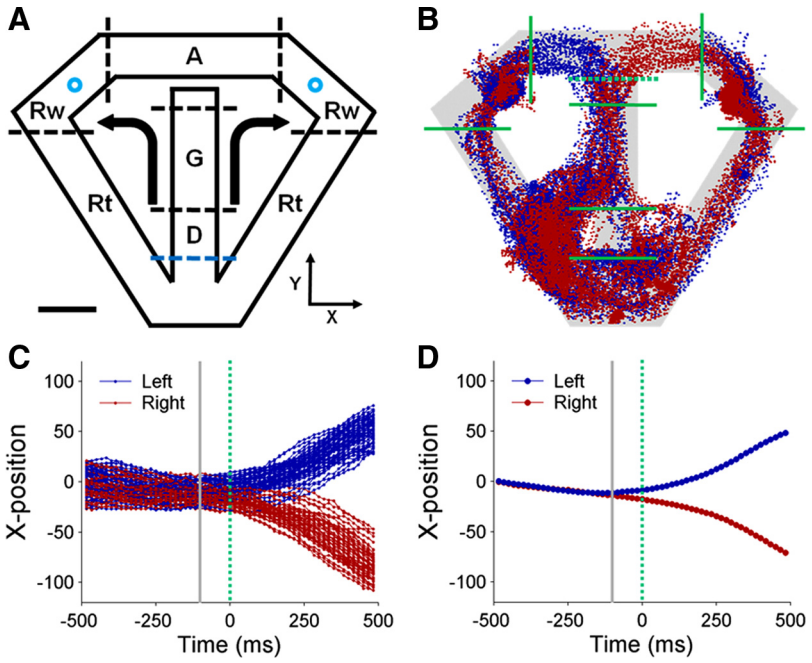

Figure 2. Behavioral stages and animal's locomotive trajectory. ***A***, The task was divided into five behavioral stages: delay (D), go (G), approach to reward (A), reward (Rw), and return (Rt) stages. Blue circles indicate the locations of water delivery. Dotted lines indicate approximate stage transition points. Each trial began when the animal crossed the blue dotted line (onset of the delay stage). Scale bar, 10 cm. ***B***, Movement trajectories during an example session. Each dot indicates animal's head position that was sampled at 60 Hz. Green solid lines indicate stage transitions. Beginning from the reward onset, blue and red traces indicate trials associated with left and right upcoming choice of the animal, respectively. Trials were decimated (3 to1) to enhance visibility. ***C***, ***D***, The time course of horizontal (*X*) coordinates of animal's position data near the onset of the approach stage during an example recording session shown in ***B*** (***C***, individual trials; ***D***, mean). Blue and red indicate trials associated with left and right goal choice, respectively. Green dotted line (0 ms) corresponds to the time when the animal reached a particular vertical position (horizontal dotted line in ***B***) determined by visual inspection to show clear separation in the animal's *X* positions according to its choice, whereas the gray line corresponds to the time when the difference in the *X* positions for the left- and right-choice trials first became statistically significant (*t* test, $p < 0.05$) within $\pm$ 0.5 s time window.

interspike interval histograms (Baeg et al., 2007). For those units that were recorded for two or more days, the session in which the units were most clearly isolated from background noise and other unit signals was used for analysis.

Tetrodes were fabricated by twisting four strands of polyimide-insulated nichrome wires (H.P. Reid Co.) together and gently heated to fuse the insulation (final overall diameter: ~40 μm). The electrode tips were cut and gold-plated to reduce impedance to 0.3–0.6 MΩ measured at 1 kHz. Unit signals were amplified with the gain of 5000–10,000, bandpass-filtered between 0.6 and 6 kHz, digitized at 32 kHz, and stored on a personal computer using a Cheetah data acquisition system (Neuralynx). Single units were isolated by examining various two-dimensional projections of spike waveform parameters, and manually applying boundaries to each subjectively identified unit cluster using custom software (MClust 3.4, A. D. Redish, University of Minnesota, Minneapolis, MN) (supplemental Fig. S1, available at www.jneurosci.org as supplemental material). Only those clusters that were clearly separable from each other and from background noise throughout the recording session were included in the analysis. The head position of the animals was recorded at 60 Hz by tracking an array of light-emitting diodes mounted on the headstage. Unit signals were recorded with the animals placed on a pedestal (resting period) for ~10 min before and after experimental sessions to examine the stability of recorded unit signals. Unstable units were excluded from the analysis. When recordings were completed, small marking lesions were made by passing an electrolytic current (50 μA, 30 s, cathodal) through one channel of each tetrode, and recording locations were verified histologically as previously described (Baeg et al., 2001) (Fig. 1*B*).

Logistic regression analysis of behavioral data

The effects of past choices and rewards on rat's current choice were examined by performing a trial-by-trial analysis of rat's choices using the following logistic regression model (Lau and Glimcher, 2005):

$$\log\left(\frac{p_L(i)}{p_R(i)}\right) = \sum_{j=1}^{N}\gamma_j^r(R_L(i-j) - R_R(i-j)) + \sum_{j=1}^{N}\gamma_j^c(C_L(i-j) - C_R(i-j)) + \gamma_0,$$

where $p_L(i)$ [or $p_R(i)$] is the probability of selecting the left (or right) goal in the *i*th trial. The variables $R_L(i)$ [or $R_R(i)$] and $C_L(i)$ [or $C_R(i)$] are reward delivery at the left (or right) goal (0 or 1) and the left (or right) goal choice (0 or 1) in the *i*-th trial, respectively. The variable N denotes the number of past trials that were included in the model ($N = 10$). The coefficients γ_j^r and γ_j^c denote the effect of past rewards and choices, respectively, and γ_0 is a bias term. The numbers of total trials used in the regression for the three animals were 2862, 1536, and 626.

Reinforcement leaning model

Action values were computed based on the Rescola–Wagner rule as previously described (Samejima et al., 2005; Seo and Lee, 2007). Briefly, we constructed a simple reinforcement learning (RL) model in which action values [$Q_L(t)$ and $Q_R(t)$] were updated based on RPE in each trial as the following: if $a(t)$ = left, RPE = $R(t) - Q_L(t)$, $Q_L(t+1) = Q_L(t) + \alpha \cdot$ RPE, $Q_R(t+1) = Q_R(t)$; if a(t) = right, RPE = $R(t) - Q_R(t)$, $Q_R(t+1) = Q_R(t) + \alpha \cdot$ RPE, $Q_L(t+1) = Q_L(t)$, where α is the learning rate, $R(t)$ represents the reward in the *t*th trial (1 if rewarded and 0 otherwise) and $a(t)$ indicates the selected action in the *t*th trial (left or right goal choice). Therefore, action value was updated only for the goal chosen by the animal. Note that "action value" is equivalent to "action value function" in RL theories (Sutton and Barto, 1998) and refers to the value associated with a particular choice rather than physical characteristics of the animal's motor outputs.

In the RL model, actions were chosen according the softmax action selection rule in which choice probability varied as a graded function of the difference in action values, $Q_L(t) - Q_R(t)$. Thus, the probability for selecting the left goal [$p_L(t)$] was defined as follows:

$$p_L(t) = \frac{1}{1 + \exp(-\beta(Q_L(t) - Q_R(t)))},$$

where β is the inverse temperature that determines the degree of exploration in action selection. The parameters α and β were estimated separately for each session using a maximum likelihood procedure (Seo and Lee, 2007). Their distribution is shown in supplemental Figure S2, available at www.jneurosci.org as supplemental material. Mean values of α for each animal were 0.29, 0.28, and 0.2, and those of β were 3.25, 2.46, and 2.67, respectively.

Although we only show the results from the analysis using the Rescola–Wagner rule (or Q-learning model) (Sutton and Barto, 1998), we also analyzed behavioral and neural data using several modified versions of the Q-learning model (Barraclough et al., 2004; Ito and Doya, 2009). All tested models accounted for animal's choice behavior quite well, and the pattern of value-related neural signals reported in the Results was similar regardless which Q-learning model was used to analyze the neural data (supplemental

Text, supplemental Fig. S3, available at www.jneurosci.org as supplemental material).

Analysis of neural data

Unit classification. The recorded units were classified into two groups based on average firing rate and spike width. Those units with mean discharge rate <7.2 Hz and spike width ≥0.22 ms were classified as putative medium spiny neurons (MSNs) and the rest were classified as putative interneurons (supplemental Fig. S1, available at www.jneurosci.org as supplemental material). Mean discharge rates of the putative MSNs and putative interneurons were 1.3 ± 0.8 and 9.7 ± 5.2 Hz, respectively, in the DS and 1.4 ± 0.9 and 9.2 ± 5.2 Hz, respectively, in the VS. Mean spike widths of the putative MSNs and putative interneurons were 0.37 ± 0.03 and 0.19 ± 0.06 ms, respectively, in the DS and 0.38 ± 0.03 and 0.23 ± 0.08 ms, respectively, in the VS. The majority of the analyzed units were putative MSNs (DS: $n = 153$, 76.1%; VS: $n = 128$, 77.6%). Otherwise noted, the results in the present study were obtained from the analyses that included both types of neurons; none of the conclusions changed when putative interneurons were excluded, however (data not shown).

Multiple regression analysis. To test how neural activity is modulated by behavioral factors, we applied multiple linear regression analyses in which the mean firing rate of a neuron during a particular time window of a trial was given by a linear function of various behavioral factors such as animal's choice, reward delivery, and estimated values. All trials in a given session (144–168, 157 ± 4.2 per session, mean ± SD) were subject to analysis. To focus on the time course of neural signals, spike rates were measured for windows with fixed durations that were aligned to the onset of each behavioral stage, including those trials in which successive windows overlapped temporally. Similar results were obtained, however, when neural data from the overlapping windows were excluded from the analysis (data not shown). Neural signals related to animal's choice, its outcome, and their interaction were examined using the following regression model for all behavioral stages (Fig. 3):

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot C(t-1) + a_3 \cdot C(t-2) + a_4 \cdot R(t) + a_5 \cdot R(t-1) + a_6 \cdot R(t-2) + a_7 \cdot X(t) + a_8 \cdot X(t-1) + a_9 \cdot X(t-2) + \varepsilon(t), \quad \text{(Model 1)}$$

where $S(t)$ denotes spike discharge rate, $C(t)$, $R(t)$, and $X(t)$ represent animal's choice (left or right), its outcome (0 or 1), and their interaction $[C(t) \times R(t)]$, respectively, in trial t, $\varepsilon(t)$ is the error term, and a_0–a_9 indicate the regression coefficients. It should be noted that the "choice" signal may represent movement direction or spatial location of the animal (Lavoie and Mizumori, 1994; Schmitzer-Torbert and Redish, 2004). We nevertheless use the term choice here to be consistent with our previous report (Kim et al., 2007) and for simplicity.

Value-related neural signals were examined using the following regression model for all behavioral stages (see Fig. 5; supplemental Fig. S3, available at www.jneurosci.org as supplemental material):

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot R(t) + a_3 \cdot X(t) + a_4 \cdot Q_L(t) + a_5 \cdot Q_R(t) + a_6 \cdot Q_c(t) + \varepsilon(t), \quad \text{(Model 2)}$$

where $Q_L(t)$ and $Q_R(t)$ denote action values associated with the left and right goal choices, respectively (Samejima et al., 2005), and $Q_c(t)$ denotes the action value of the goal chosen by the animal in trial t (chosen value) (Lau and Glimcher, 2008; Seo et al., 2009; Kim et al., 2009). This model was also used in examining whether choice, reward, and chosen value signals are concurrently conveyed by the same neurons in the reward stage. To confirm that the chosen value signals determined with the above model (see Fig. 5) indeed represent chosen rather than unchosen value, we compared the strengths of signals related to chosen and unchosen values using the following regression model:

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot R(t) + a_3 \cdot X(t) + a_4 \cdot Q_c(t) + a_5 \cdot Q_u(t) + \varepsilon(t), \quad \text{(Model 3)}$$

where $Q_c(t)$ and $Q_u(t)$ denote the values of the chosen and unchosen goal, respectively. We also examined neural signals for the sum of and the difference in action values ($\Sigma Q(t)$ and $\Delta Q(t)$, respectively), which would be more related to overall rate of reinforcement and animal's choice, respectively, than action values (Seo and Lee, 2007; Ito and Doya, 2009; Seo et al., 2009), using the following regression model (supplemental Fig. S4, available at www.jneurosci.org as supplemental material):

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot R(t) + a_3 \cdot X(t) + a_4 \cdot \Sigma Q(t) + a_5 \cdot \Delta Q(t) + a_6 \cdot Q_c(t) + \varepsilon(t), \quad \text{(Model 4)}$$

where $\Sigma Q(t) = Q_L(t) + Q_R(t)$ and $\Delta Q(t) = Q_L(t) - Q_R(t)$.

To examine whether striatal neurons encode chosen value consistently for both actions, we examined neural signals related to choice × chosen

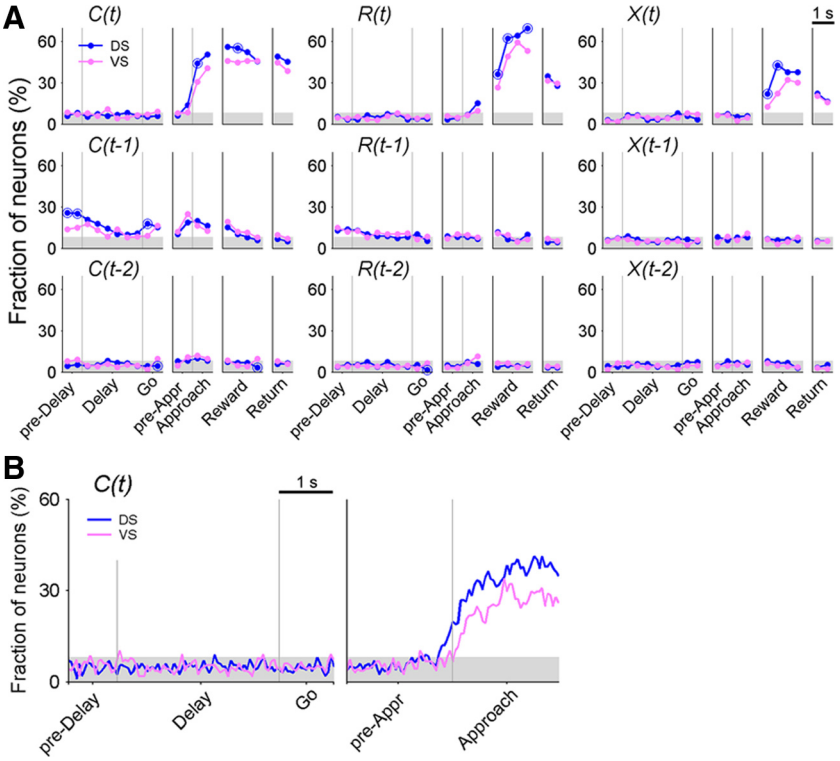

Figure 3. Striatal activity related to animal's choice and its outcome in the current and previous trials. ***A***, The graphs show fractions of neurons that were significantly modulated by animal's choice (*C*), its outcome (*R*), or their interaction (*X*) in the current (*t*) and previous trials ($t - 1$ and $t - 2$) in the regression Model 1 in non-overlapping 0.5 s time windows across different behavioral stages [pre-Delay: last 1 s of the return stage; Delay: the entire delay stage (3 s); Go: first 1 s; pre-Approach (pre-Appr): last 1 s of the go stage; Approach: first 1 s; Reward: first 2 s; Return: first 1 s]. Large open circles indicate that the fractions are significantly different between the DS and VS (χ^2 test, $p < 0.05$). Each vertical line indicates the beginning of a given behavioral stage. Values within the shaded areas are not significantly different from the significance level of $p = 0.05$ for the VS (binomial test). The threshold for the DS is slightly lower due to the larger number of neurons (data not shown). ***B***, The fraction of neurons that significantly modulated their activity according to the current choice [*C*(*t*)] is plotted at a higher temporal resolution (100 ms moving window advanced in steps of 50 ms).

value interaction $[C(t) \times Q_c(t)]$ by applying the following regression model to the spike data during the first 1 s of the reward stage:

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot R(t) + a_3 \cdot X(t) + a_4 \cdot Q_c(t) + a_5 \cdot Q_u(t) + a_6 \cdot C(t) \times Q_c(t) + \varepsilon(t). \quad \text{(Model 5)}$$

We also compared chosen value-related neuronal activity associated with left and right goal choice by applying the following regression separately to the left and right goal-choice trials using the spike data during the first 1 s of the reward stage (see Fig. 6):

$$S(t) = a_0 + a_1 \cdot R(t) + a_2 \cdot Q_c(t) + a_3 \cdot Q_u(t) + \varepsilon(t). \quad \text{(Model 6)}$$

Here, the chosen value is equivalent to the left (or right) action value for the left (or right) goal-choice trials.

The following regression models were used to determine whether neuronal activity in the reward stage is more correlated with RPE or updated chosen value. We first selected the neurons encoding both reward and chosen value using the following regression model based on the discharge rates during the first 1 s of the reward stage (see Figs. 8, 9):

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot Q_L(t) + a_3 \cdot Q_R(t) + a_4 \cdot R(t) + a_5 \cdot Q_c(t) + \varepsilon(t). \quad \text{(Model 7)}$$

Because RPE and updated value were computed based on the combination of the current choice and its outcome, their interaction term, $X(t)$, was omitted in this regression model. We then examined which of the following models better accounted for neuronal activity (see Figs. 8, 9):

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot Q_L(t) + a_3 \cdot Q_R(t) + a_4 \cdot \mathrm{RPE}(t) + \varepsilon(t), \quad \text{(Model 8)}$$

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot Q_L(t) + a_3 \cdot Q_R(t) + a_4 \cdot \mathrm{up}Q_c(t) + \varepsilon(t), \quad \text{(Model 9)}$$

where $\mathrm{RPE}(t)$ is the reward prediction error and $\mathrm{up}Q_c(t)$ is the updated chosen value. Namely, $\mathrm{up}Q_c(t) = Q_c(t) + \alpha \cdot \mathrm{RPE}(t)$, where α refers to the maximum likelihood estimate of the learning rate obtained for each session.

Coefficient of partial determination. The variance in neural activity accounted for by RPE or updated chosen value was quantified by the coefficient of partial determination (CPD) (Neter et al., 1996) as the following: $\mathrm{CPD}(X_2) = [\mathrm{SSE}(X_1) - \mathrm{SSE}(X_1, X_2)]/\mathrm{SSE}(X_1)$, where $\mathrm{SSE}(X_i)$ denotes the sum of squared errors in a regression model that includes X_i. To determine the CPD for RPE or updated chosen value, X_1 included $C(t)$, $Q_L(t)$, and $Q_R(t)$, and X_2 was either RPE or updated chosen value, $Q_c(t)$. Thus, CPD corresponds to the fraction of variance in neural activity that can be accounted for by adding either RPE or updated chosen value to the following regression model:

$$S(t) = a_0 + a_1 \cdot C(t) + a_2 \cdot Q_L(t) + a_3 \cdot Q_R(t) + \varepsilon(t). \quad \text{(Model 10)}$$

Permutation test. Because action values are updated iteratively in RL algorithms, the estimated action values are inevitably correlated between successive trials. Neural activity (dependent variable) was also correlated across successive trials (e.g., mean serial correlation = 0.052 ± 0.007 and 0.062 ± 0.007 for the neural data during the first and last 1 s of the delay stage), which could be due to a number of factors, such as a slow drift in the spike rates during the recording session. Regardless of its origin, such serial correlation in spike rates could potentially violate the independence assumption in the regression analysis and increase the amount of activity spuriously correlated with action values (Seo and Lee, 2008). We therefore used a permutation test to evaluate statistical significance of regression coefficients for the multiple regression analyses that contained action values. An additional analysis revealed that spike autocorrelation in the present study is largely due to systematic changes in neural activity between blocks, rather than trial-by-trial correlation in neural activity (supplemental Fig. S5, available at www.jneurosci.org as supplemental material). Thus, for the permutation test, the original block sequence was preserved and spike rates were randomly shuffled 1000 times across different trials within each block. For each of these trial-shuffled datasets, we repeated the same regression analysis using the original behavioral data (i.e., animal's choice, its outcome and action values). The p value for each regression coefficient was then determined by the frequency in which the magnitude of the original regression coefficient was exceeded by that of the regression coefficients obtained after trial shuffling.

Statistical analysis

Statistical significance of a regression coefficient was tested based on a t test (model 1 which does not include any value term) and a permutation test (other regression models including value terms). Significance of the fraction of neurons for a given variable was tested with a binomial test. A p value <0.05 was used as the criterion for a significant statistical difference unless noted otherwise. Data are expressed as mean ± SEM unless noted otherwise.

Results

Choice and locomotive behavior during two-armed bandit task

The animals began to choose the goal associated with a higher reward probability more frequently within 10–20 trials after a block transition, indicating that they quickly captured changes in relative reward probabilities and made goal choices accordingly (Fig. 1*C*). Consistent with the results from previous studies (Lee et al., 2004; Lau and Glimcher, 2005; Samejima et al., 2005; Huh et al., 2009), more quantitative analyses revealed that this was accomplished by a relatively simple learning algorithm that estimates the likelihood of reward for each choice (Fig. 1*C*). Similarly, a logistic regression analysis showed that as predicted by the RL model, the animals tended to make the same choice that was rewarded in recent trials, as indicated by positive coefficients related to reward, but tended to alternate their choice, as indicated by the negative coefficient related to the animal's choice in the previous trial ($t - 1$) (Fig. 1*D*). Thus, rat's choice in the present task was systematically influenced by the history of past choices and their outcomes.

The animals showed a stereotyped pattern of movement across five behavioral stages of the task (delay, go, approach, reward and return stages, Fig. 2*A*) as illustrated in Figure 2*B*. In this figure, trials were divided into two groups depending on the upcoming goal choice of the animal (blue: left choice, red: right choice) starting from the reward stage in the previous trial. Trajectories before the left-choice and right-choice were similar throughout the trial starting from the reward phase of the previous trial until they began to diverge on the distal portion of the central bridge, which was determined to be the onset of the approach stage (Fig. 2*B–D*). More quantitative analysis revealed that animal's movement trajectory between the reward and approach stages was not related to the upcoming choice of the animal (supplemental Fig. S6, available at www.jneurosci.org as supplemental material). On the other hand, animal's movement trajectory was different in the lower central section of the maze (early delay stage) depending on which goal the animal was returning from (supplemental Fig. 6*B*, available at www.jneurosci.org as supplemental material). Upon arrival at the central bridge, the animal extended its head to the left or right side of the central bridge (Fig. 2*B*) for the most of the delay period (late delay stage), during which animal's head position was independent of animal's previous goal choice. Thus, animal's head position was sig-

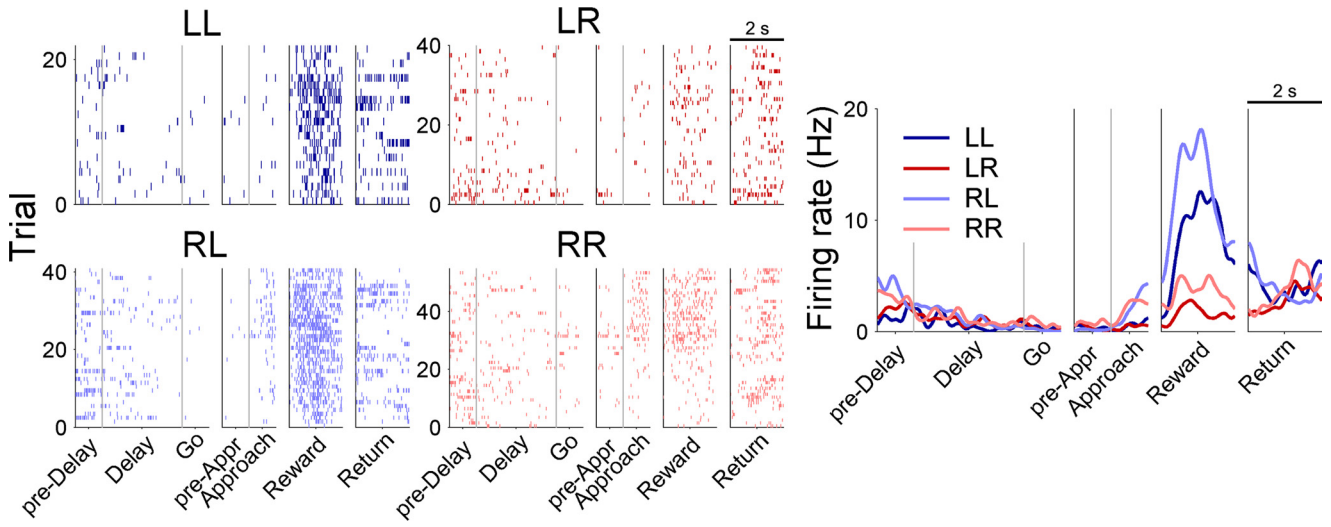

Figure 4. An example neuron in the VS that modulated its activity according to the animal's choices in the current and previous trials. Trials were grouped according to the sequence of the previous and current trial (L, left; R, right; e.g., RL, right and left choice in the previous and current trial, respectively). Left, Spike raster plots. Right, Spike density functions that were generated by applying a Gaussian kernel ($\sigma = 100$ ms) to the corresponding spike trains.

nificantly different depending on the previous goal choice only in the early delay stage (initial 1.1 ± 0.3 s, mean ± SD) (supplemental Fig. 6*B*, available at www.jneurosci.org as supplemental material).

Neural signals for choice and reward

A total of 201 and 165 single units with mean discharge rates >0.1 Hz were recorded from the DS and VS, respectively (Fig. 1*B*), of three rats. We related spike discharge rates to various behavioral variables by running multiple regression analyses using all trials in a given session. Because we were interested in neural signals related to action selection as well as action evaluation that might take place during different behavioral stages, all stages of the task were subjected to analysis. We first examined neural signals related to animal's choice, its outcome and their interaction in the current and previous trials using the regression Model 1 (Fig. 3*A*). Overall, the percentage of the neurons that significantly changed their activity according to the upcoming choice of the animal in the delay or go stage (i.e., before behavioral manifestation of animal's choice) was low. The analysis of the spike counts during the last 1 s of the delay stage (i.e., immediately before movement onset) showed that the number of neurons with significant effect of the upcoming choice was only 13 in the DS and 6 in the VS, which were not significantly higher than the values predicted by the significance level used (binomial test, $p = 0.208$ and 0.837, respectively). However, during the last 0.5 s of the go stage (i.e., 0.5 s before the onset of the approach stage), the number of neurons encoding the upcoming choice was significantly above chance level in the DS ($n = 27$, 13.4%; binomial test, $p < 0.001$), although it was not significant in the VS ($n = 13$, 7.9%; $p = 0.071$) (Fig. 3*A*).

We further examined the temporal dynamics of choice signals at a higher temporal resolution using a 100 ms time window that was advanced in 50 ms time steps. This analysis revealed that choice signal started to rise above chance level ~200 ms before the onset of the approach stage in the DS, but not in the VS (Fig. 3*B*). For the 200 ms time interval before the onset of the approach stage, there were 37 DS (out of 201, 18.4%, binomial test, $p < 0.001$) and 13 VS (out of 165, 7.9%, $p = 0.071$) neurons that significantly modulated their activity according to the upcoming choice of the animal, which were significantly different (χ^2 test, $p < 0.001$). In summary, upcoming choice signal was overall weak, but it started to strengthen ~200 ms before the behavioral manifestation of animal's choice in the DS.

Neural signals for current choice [$C(t)$] were greatly elevated after the animal made a goal choice (i.e., in the approach, reward and return stages) and then slowly decayed during the delay stage in the next trial [now corresponding to $C(t-1)$] in both structures (Fig. 3*A*). Of all the neurons conveying current choice signal during the first 1 s of the approach stage, the numbers of neurons increasing their firing rates more for the ipsilateral and contralateral choice (DS: $n = 51$ and 65, respectively; VS: $n = 31$ and 43, respectively) did not differ significantly (χ^2 test, DS: $p = 0.194$, VS: $p = 0.163$). Interestingly, the previous choice [$C(t-1)$] signal increased somewhat during the go, approach, and reward stages compared to the late delay stage, so that the signals related to the current and previous choices were encoded simultaneously (Figs. 3, 4), which is consistent with a previous study in the rat VS (Kim et al., 2007). Moreover, during the first 1 s period of the approach stage, both DS and VS neurons were more likely to change their activity according to the animal's previous choice when their activity was also significantly affected by the animal's current choice compared to when their activity was not related to the animal's current choice (χ^2 tests, $p = 0.013$ and 0.015, respectively).

Similar to the signals related to the animal's choice, the signals related to the reward [i.e., choice outcome, $R(t)$] were greatly elevated in the reward stage (i.e., after the outcome of a choice was revealed) and then slowly decayed until the delay stage in the next trial [now corresponding to $R(t-1)$] in both structures (Fig. 3*A*) as recently described in the rat VS (Ito and Doya, 2009) and monkey DS (Histed et al., 2009). During the first 1 s period of the reward stage, neuronal activity was lower in rewarded than unrewarded trials in the majority of cases (DS: 101 out of 123, 82.1%, binomial test, $p < 0.001$; VS: 62 out of 84, 73.8%, binomial test, $p < 0.001$), which is consistent with previous results obtained from the rat VS (Roitman et al., 2005; Ito and Doya, 2009).

To update action value, reward-related signals must be combined appropriately with the signals related to animal's choice of action. Indeed, a significant choice × reward interaction signal [$X(t)$] was frequently observed in the reward stage (Fig. 3*A*),

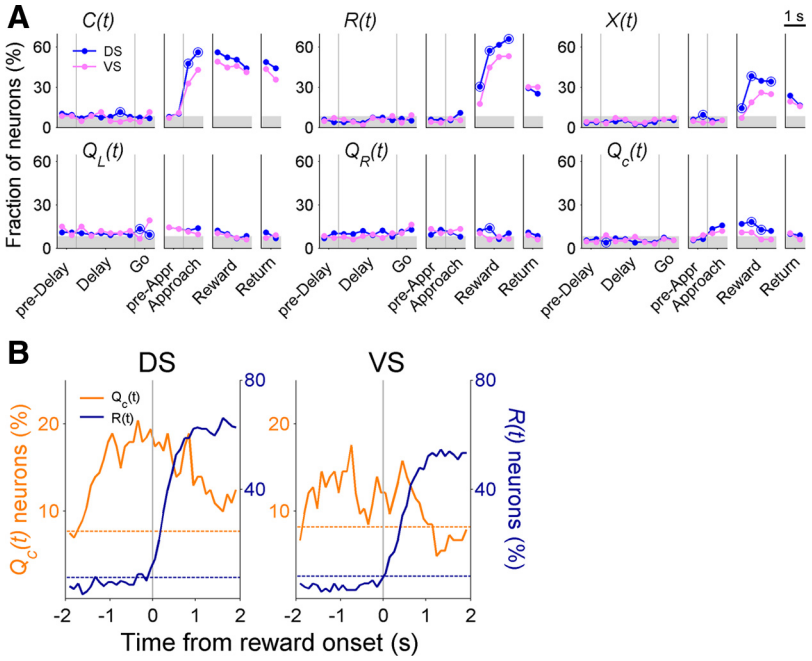

Figure 5. Neural signals related to action value and chosen value. ***A***, The graphs show fractions of neurons that significantly modulated their activity according to action value [$Q_L(t)$ and $Q_R(t)$] or chosen value [$Q_c(t)$] in the regression Model 2 tested with non-overlapping 0.5 s time windows. Results for the other variables (current choice, current reward, and their interaction) are also shown. Same format as in Figure 3. ***B***, The fractions of neurons that were significantly modulated by chosen value (orange) and current reward (blue) around the time of reward delivery are shown at a higher temporal resolution. The dotted lines show the minimum values that are significantly higher than the significance level of 0.05 (binomial test). Note that *y*-axis scales are different for the reward and chosen value signals.

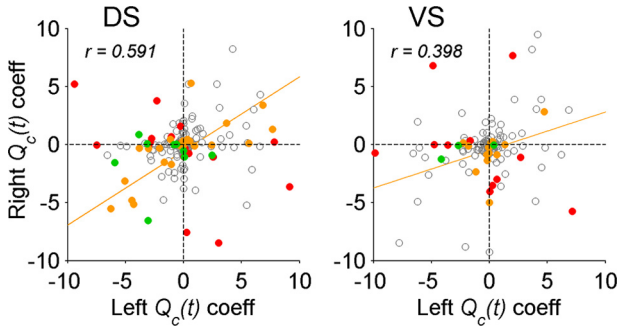

Figure 6. Relationship between regression coefficients related to chosen value in the left and right goal choice trials. Trials were divided according to animal's goal choice and discharge rates in the first 1 s of the reward stage were used for this analysis. The graphs show coefficients for chosen value [left and right $Q_c(t)$ coeff] in the regression Model 6. Orange and red circles denote neurons encoding chosen value (Model 2) and those with significant choice × chosen value interaction (Model 5), respectively, whereas green circles indicate neurons encoding chosen value (Model 2) and showing significant choice × chosen value interaction (Model 5). The neurons encoding chosen value (Model 2) were used to determine the best-fitting lines (orange lines) and to calculate the correlation coefficients shown, both of which were significantly different from 0 (DS: $p < 0.001$; VS: $p = 0.029$).

indicating that a large proportion of striatal neurons conveyed choice and reward signals conjunctively. Although many neurons showed significant interactions between choice and reward, the number of neurons that significantly modulated their activity according to the reward oppositely for different choices was relatively small (30.5 and 33.3% of all interaction-encoding neurons in the DS and VS, respectively). In other words, significant interaction between choice and reward arose largely when the magnitude of reward-related activity, rather than its polarity, differed for the two choices.

Neural signals related to value

We then tested whether the striatal activity is related to the subjective value of expected reward, using action values estimated by the RL model. We examined neural signals for action value, which is the value for one of the available actions (left goal-choice vs right goal-choice), as well as chosen value, which refers to the value of the action chosen in a given trial (Lau and Glimcher, 2008) using the regression Model 2. Neural signals for action value (Q_L or Q_R) were above chance level in the delay and go stages (i.e., before animal's choice was revealed) and persisted until the early reward stage in both structures. In contrast, neural signals for chosen value (Q_c) were weak in the delay and go stages, but arose above chance level in the approach and reward stages in both DS and VS (Fig. 5*A*). Analysis of spike counts during the last 1 s of the delay stage showed that 26 DS (12.9%) and 17 VS (10.3%) neurons significantly modulated their activity in relation to the value of at least one action (Q_L or Q_R; $p < 0.025$, $\alpha = 0.05$ was corrected for multiple comparisons), which were significantly higher than the significance level used (binomial test, $p < 0.001$ and $p = 0.004$, respectively). However, only 10 DS (4.9%) and 7 VS (4.2%) neurons significantly modulated their activity according to chosen value, which were not significantly above chance level, (binomial test, $p = 0.552$ and 0.723, respectively). In the regression analysis applied to the first 1 s of the approach stage that included both chosen and unchosen values as explanatory variables (Model 3), neurons encoded the value of chosen action more frequently than the value of unchosen action [DS: 38 (18.9%) and 18 (8.9%), respectively; VS: 22 (13.3%) and 17 (10.3%), respectively]. The number of chosen value-coding neurons was significantly larger than that of unchosen value-coding neurons when the DS and VS data were combined (60 vs 35; χ^2 test, $p = 0.010$), indicating that striatal neurons indeed conveyed neural signals for chosen action value after animal's choice of action. Thus, as previously reported in the monkey DS (Lau and Glimcher, 2008), action value and chosen value signals were preferentially observed before and after animal's overt choice of action, respectively.

In our binary choice task, it was not necessary for the animal to maintain two separate action values, because the richer alternative always changed its spatial position after a block transition. Keeping track of only one action value was sufficient for near-optimal performance in our task. An alternative possibility is that the animal kept track of the difference in action values (ΔQ) rather than encoding individual action values. We therefore examined neural signals related to the sum of and the difference in action values instead of individual action values (Model 4). The analysis showed that the strengths of neural signals for the sum of and the difference in action values were similar to those of individual action values (supplemental Fig. S4, available at www.jneurosci.org as supplemental material).

To examine whether striatal neurons encode chosen value consistently for both actions, we examined neural signals for choice ×

chosen value interaction (Model 5). In the first 1 s of the reward stage, there were 29 (14.4%) DS and 18 (10.9%) VS neurons that conveyed significant choice × chosen value interaction signals (see Fig. 7), which were significantly above chance (binomial test, DS: $p < 0.001$; VS: $p = 0.002$). Of these interaction-coding neurons, the polarity of chosen value-related activity differed for the two choices in some neurons (21 DS and 11 VS neurons), whereas only its magnitude differed for the two choices in the others (8 DS and 7 VS neurons; Model 6) (Fig. 6). The latter neurons encoded the value for a particular action (or goal) only when the same action (or goal) was chosen by the animal. Among the neurons that modulated their activity significantly according to the chosen value during the first 1 s of the reward stage (DS: $n = 35$; VS: $n = 22$; Model 2), 11 DS (31.4%) and 5 VS (22.7%) neurons showed choice × chosen value interactions (Model 5), respectively, suggesting that for some striatal neurons, the signals related to chosen value were modulated by the animal's choice. On the other hand, for the same chosen value-encoding neurons (35 DS and 22 VS neurons), regression coefficients for the chosen value for the left and right goal choice (Model 6) were significantly correlated (Fig. 6), indicating that striatal neurons tended to encode chosen values consistently for both actions.

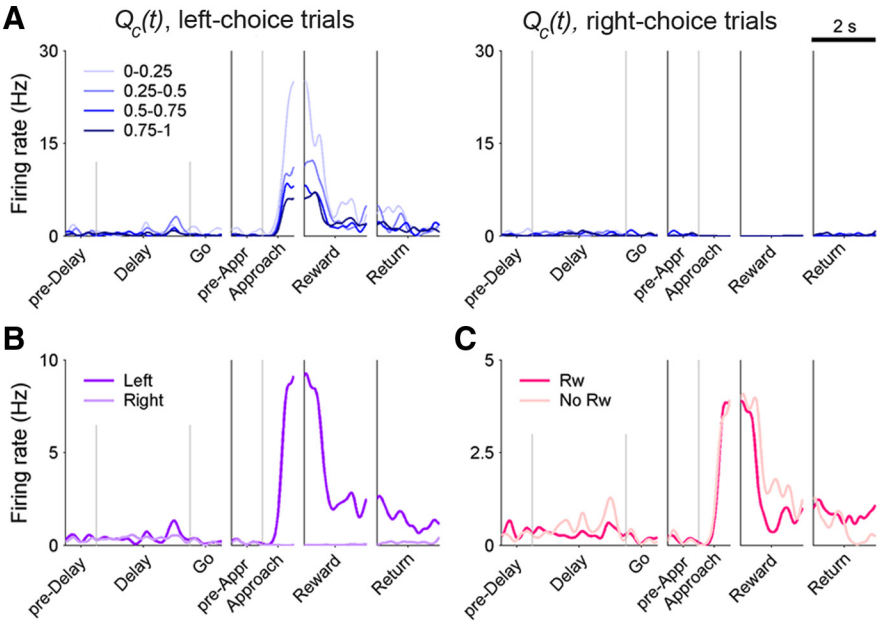

Figure 7. An example neuron in the DS that modulated its activity according to animal's choice, its outcome as well as chosen value in the reward stage. ***A***, Spike density functions (Gaussian kernel width = 100 ms) for different levels of chosen value are shown for the left and right goal-choice trials. This neuron also modulated its activity significantly according to choice × chosen value interaction (Model 5). The trials were divided into four groups according to the level of associated chosen value, and then spike density functions of the four groups were plotted in different colors. Neural activity in the approach and reward stages decreased with the left action value when the animal chose the left goal, and therefore encoded the chosen value in such trials. ***B***, Spike density functions for the left versus right goal choice trials. ***C***, Spike density functions for rewarded (Rw) versus unrewarded (No Rw) trials.

If the striatum is involved in updating action value, the signals necessary for this process, namely the chosen value and reward signals, might temporally overlap in the striatum immediately after the onset of the reward stage. Therefore, to examine more closely the time course of signals related to chosen value at the time of reward delivery, we applied the same analysis to neural activity during the 4 s period starting 2 s before the reward delivery using a 0.5 s sliding window that was advanced in 0.1 s time steps (Model 2). This analysis revealed that the signal for chosen value reached its peak near the onset of the reward stage and then subsided gradually (Fig. 5*B*). On the other hand, reward-related signal arose after the beginning of the reward stage and peaked ~1 s after its onset in both DS and VS. Hence, chosen value signal was uploaded before reward signal arrived in the striatum, and they overlapped briefly at the beginning of the reward stage.

Neural signals for updated value versus RPE

The above analyses showed that striatal neurons carry chosen action value signals in addition to choice and reward signals in the reward stage. We examined whether these signals are conveyed by the same striatal neurons by analyzing neuronal activity during the first 1 s of the reward stage (Model 2). In the DS, of 35 chosen-value-coding neurons, 16 encoded the current reward $[R(t)]$ as well. In the VS, of 22 chosen-value-coding neurons, 14 also encoded the current reward. Among the neurons that encoded both chosen value and reward, 10 DS and 9 VS neurons additionally encoded the current choice $[C(t)]$ of the animal, indicating that some neurons concurrently conveyed the information on the selected action, its outcome, and the value of the chosen action (Fig. 7). It should be emphasized that the likelihood of false negative (type II error) increases when multiple tests are applied conjunctively. Therefore, the fraction of neurons encoding the animal's action, its outcome, and chosen value is likely to be much higher than the values reported here.

The above analysis showed that all the necessary ingredients to compute RPE and updated value for the chosen action converge onto the same striatal neurons during the reward stage. The mere convergence of these multiple signals does not indicate, however, whether such signals are combined to compute RPE or update chosen action value. We therefore examined whether neural activity in the striatum is more correlated with RPE or updated chosen value $[upQ_c]$. If a given neuron encodes either of these quantities, its activity should be systematically affected by chosen value as well as the outcome of animal's choice (i.e., reward), since RPE and updated chosen value are determined by the difference and weighted sum of the reward and chosen value, respectively. Therefore, this analysis was applied to those striatal neurons that significantly modulated their activity according to both the current reward and chosen value (Model 7). Of 33 such neurons, 16 (12 DS and 4 VS) and 17 (7 DS and 10 VS) were better accounted for by the model containing RPE (Model 8) and updated chosen value (Model 9), respectively, suggesting that the signals related to RPE and updated chosen value coexist in the striatum. Examples of neurons encoding RPE or updated chosen value, their population spike density functions, and the time course of the CPD for RPE or updated chosen value are shown in Figure 8. It is notable that the time course of RPE signals in the striatum was not as brief as that of midbrain dopamine neurons (Schultz, 1998). This might reflect different types of RPE signals across the two brain regions, but it might simply reflect reward signals $[R(t)]$ that were highly correlated with RPE signals.

To confirm further whether striatal signals seemingly related to updated value or RPE were computed by combining the signals

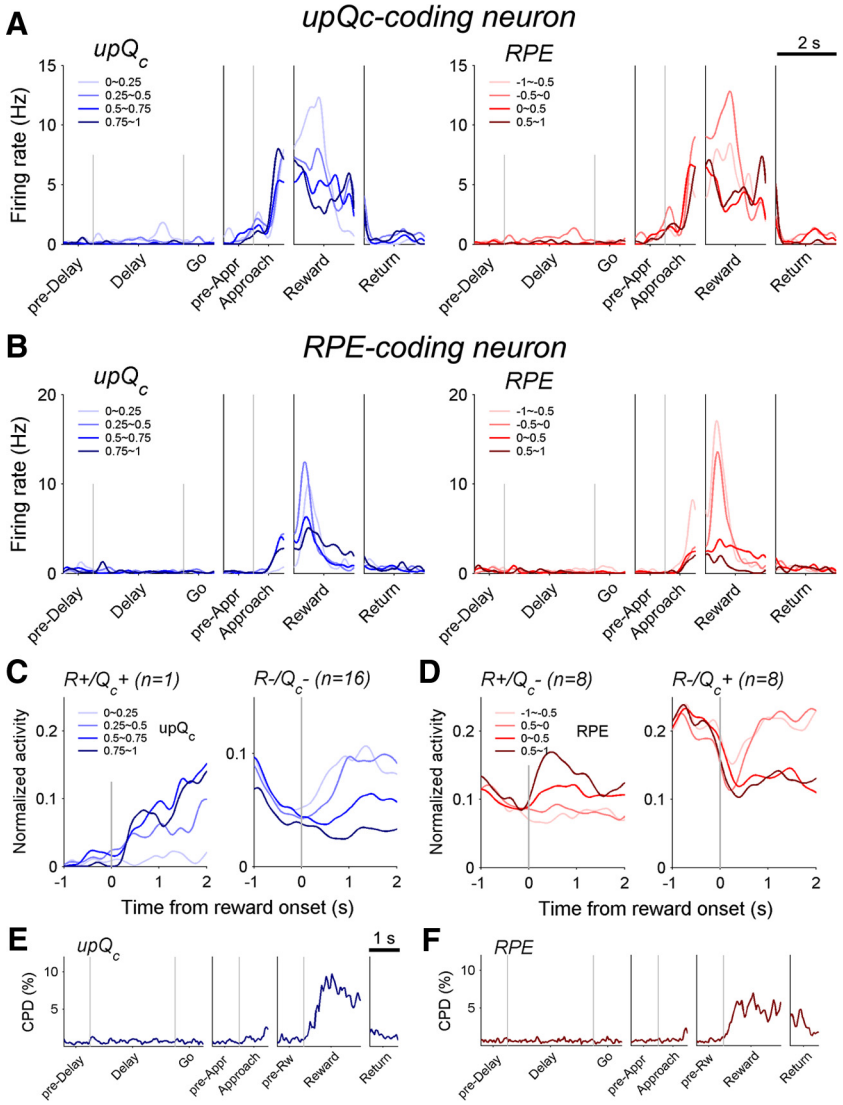

Figure 8. Neuronal activity encoding updated value or RPE. ***A***, An example VS neuron in which activity was correlated more strongly with updated chosen value [up*Qc*] than RPE (Models 8 and 9). ***B***, An example DS neuron in which activity was more correlated with RPE. For comparison, spike density functions (Gaussian kernel width = 100 ms) were estimated separately for different ranges of updated chosen value as well as RPE. ***C***, ***D***, Population-average spike density functions are shown for 1 s before and 2 s after the onset of the reward stage. Neurons that significantly modulated their activity according to both reward [$R(t)$] and chosen value [$Q_c(t)$] were divided into four groups according to the signs of their regression coefficients (the number of samples in each group is indicated in each plot). Activity of neurons with the same signs ($n = 17$) was more correlated with updated value than RPE, and therefore their spike density functions were plotted according to updated chosen value (***C***), whereas activity of neurons with opposite signs ($n = 16$) was more correlated with RPE and their spike density functions were plotted according to RPE (***D***). Activity of each neuron was normalized by each neuron's maximal response before averaging. ***E***, ***F***, The coefficient of partial determination (CPD) for RPE and updated value is shown for all behavioral stages in a moving window of 100 ms advanced in 50 ms steps. Only RPE- and updated value-coding neurons ($n = 16$ and 17, respectively) were selected in plotting the CPD for RPE and updated value, respectively.

related to chosen value and reward, we divided the neurons into two groups depending on the signs of the coefficients related to these two variables (Model 7). If a striatal neuron conveys RPE signal, then signs of the coefficients for reward [$R(t)$] and chosen value [$Q_c(t)$] should be opposite, because RPE = $R(t) - Q_c(t)$. On the contrary, the corresponding coefficients should have the same signs for a striatal neuron that conveys signals for updated chosen value because $\mathrm{up}Q_c(t) = Q_c(t) + \alpha \cdot [R(t) - Q_c(t)]$, where α is the learning rate (see Materials and Methods). This can be rearranged as $\mathrm{up}Q_c(t) = (1 - \alpha) \cdot Q_c(t) + \alpha \cdot R(t)$. Because $0 < \alpha < 1$, both coefficients for $Q_c(t)$ and $R(t)$ are positive in this equation and therefore the activity of neurons encoding updated chosen value should be modulated in the same direction by chosen value and reward. This is not merely a mathematical result, however, because whether neural activity is better correlated with RPE or updated value depends not only on the signs but also on the magnitudes of the coefficients for reward and chosen value. If neural signals for RPE or updated value were outcomes of spurious correlations, signs of the coefficients for reward and chosen value might deviate substantially from the predicted values. The analysis showed that all neurons significantly modulated by both chosen value and reward with the same signs for their regression coefficients were better explained by the regression model including the updated chosen value (Model 9) than the model including RPE (Model 8). Conversely, those neurons that were modulated significantly but oppositely by chosen value and reward were better explained by the model including RPE (Model 8) than the model including chosen value (Model 9) (Fig. 9*A*). These results suggest that neural signals related to RPE and updated value identified in the regression analyses were not an outcome of spurious correlation, further supporting the possibility that the DS and VS might convey both RPE and updated value signals during the early period of the reward stage.

Finally, we examined whether the RPE and updated value signals are processed similarly across different cell types. Seventeen striatal neurons in which activity was correlated more strongly with updated value than with RPE consisted of 14 putative MSNs (6 DS and 8 VS neurons) and 3 putative interneurons, whereas 16 neurons in which activity was more strongly correlated with RPE consisted of 7 putative MSNs (5 DS and 2 VS neurons) and 9 putative interneurons (7 DS and 2 VS neurons). Compared to their overall proportion (23.2%), putative interneurons were significantly more likely to encode RPE (Fisher's exact test, $p = 0.004$). To test whether the converse it true, we compared mean firing rates of the neurons encoding RPE or updated value. Consistent with the results from the analysis based on cell types, the mean firing rate of RPE-coding neurons (7.2 ± 1.6 Hz) was significantly higher than that of updated value-coding neurons (2.9 ± 1.4 Hz, Wilcoxon rank sum test, $p = 0.028$) (Fig. 9*B*). In these analyses, neurons encoding both chosen value and reward were identified by two independent statistical tests (Model 7), increasing the likelihood of type II (false negative) errors. Therefore, these analyses were repeated after relaxing the criterion for statistical significance (α) to 0.1 for each regression coefficient (Model 7). This resulted in 22 neurons encoding updated value, which consisted of 18 putative MSNs (7 DS and 11

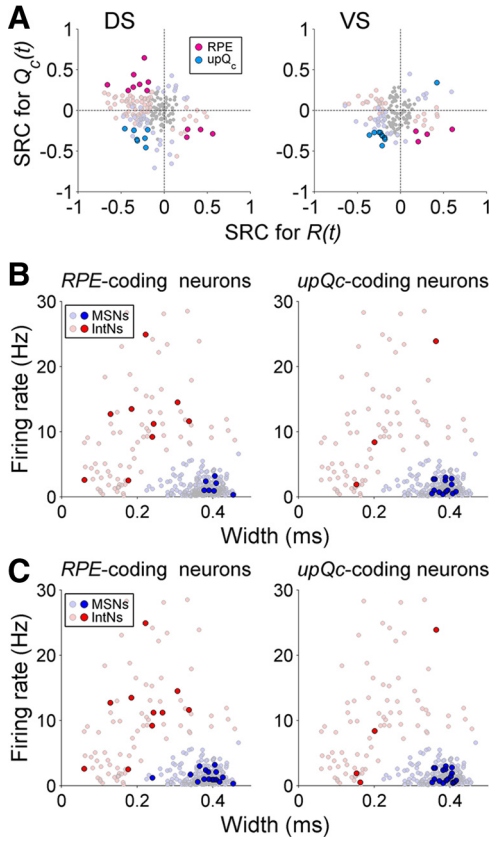

Figure 9. Characteristics of RPE- and updated value-coding neurons. ***A***, Standardized regression coefficients (SRC) related to chosen value (ordinate) and current reward (abscissa) for activity during the first 1 s of the reward stage (Model 7). Saturated colors indicate neurons encoding both reward and chosen value, whereas light colors indicate those that encoded either reward or chosen value only. The rest are indicated in gray. Red and blue indicate those neurons in which activity was more correlated with RPE- or updated chosen value, respectively (Models 8 and 9). ***B***, Scatter plots for mean firing rate and spike width. Saturated colors indicate neurons encoding RPE- (left) or updated value (right) and light colors indicate the remaining neurons. Blue and red indicate putative MSNs and interneurons, respectively. ***C***, Same as in ***B*** except that neurons were selected at $\alpha = 0.1$. DS and VS neurons are combined in ***B*** and ***C***.

VS neurons) and 4 putative interneurons (2 DS and 2 VS neurons). On the other hand, 26 RPE-coding neurons consisted of 16 putative MSNs (11 DS and 5 VS neurons) and 10 putative interneurons (8 DS and 2 VS neurons). There was a trend for the putative interneurons to preferentially encode RPE (Fisher's exact test, $p = 0.053$), and the mean firing rate of RPE-coding neurons was significantly higher than that of updated value-coding neurons (RPE-coding neurons: 5.6 ± 1.2 Hz, updated value-coding neurons: 2.4 ± 1.0 Hz, Wilcoxon rank sum test, $p = 0.014$) (Fig. 9*C*). In summary, these results show that the activity of interneurons is more likely to be correlated with RPE than updated value, suggesting different roles of MSNs and interneurons in computing RPE and updating chosen action values.

Discussion

We examined temporal profiles of neural signals for animal's choice, its outcome, and action value in the DS and VS of rats performing a free-choice task. We found that the neural signal for animal's upcoming choice was overall weak, but neural signals for action value were above chance level before animal's choice of action in both structures. Once the animal's choice was revealed, neural signals for animal's choice and chosen action value increased steeply in both structures. These neural signals persisted and were combined with the signal for reward once the outcome of the animal's choice was revealed. For some neurons, this activity was more correlated with RPE, whereas for others it was more correlated with updated value, suggesting that the value of chosen action might be updated in the striatum immediately after choice outcome is revealed. Overall, our results provide converging evidence for the role of striatum in representing and updating action value, but cast a doubt on its role in action selection.

Role of striatum in action selection

We found that upcoming choice signals were overall weak in both DS and VS, which is inconsistent with the view that the primary role of the striatum is in action selection. In the present study, significant upcoming choice signal was found in the DS, but it was well after the offset of the delay period (i.e., after movement was initiated) and only immediately (~200 ms) before the behavioral manifestation of choice. It is therefore unclear whether the upcoming choice signal represents the cause or the outcome of animal's choice. This might appear at variance with previous reports on striatal activity correlated with upcoming movement of an animal (Hikosaka et al., 1989; Alexander and Crutcher, 1990b; Apicella et al., 1992; Setlow et al., 2003; Nicola et al., 2004; Pasupathy and Miller, 2005). However, because the majority of previous studies used tasks with explicit instructions, in which the correct (or more rewarding) response was unambiguously defined, they are limited in determining whether the striatum is causally involved in action selection. It is possible that the animals in these studies made their choices covertly for future actions as soon as the correct actions were specified. In other words, choice-related striatal signals might be an outcome, rather than a cause, of animal's choice. Consistent with this possibility, a recent study has found only weak choice signals before animal's overt choice in the VS of rats engaged in a free-choice task (Ito and Doya, 2009). We have also shown previously that rat VS neurons do not convey upcoming choice signals even during a discrimination task in which the correct choice was explicitly signaled by visual signals (Kim et al., 2007).

One important issue to be resolved is the nature of the upcoming choice signal that began to arise in the DS before behavioral manifestation of animal's choice. This signal might indicate a causal role of the DS in action selection; rats may make an action selection only immediately before physically implementing its choice in the current task. Alternatively, the choice signal might be uploaded in the DS for the purpose of updating chosen value. Additional investigation is required to resolve this matter clearly. It should also be investigated whether or not strong choice signals exist before animal's choice in other parts of the striatum such as the dorsolateral striatum, which has been proposed to serve more motor-related functions than the dorsomedial striatum (Balleine et al., 2007; White, 2009).

Whereas our results do not support a direct action selection role of the striatum, they are consistent with the view that the striatum contributes to action selection by biasing animal's choice toward an action that is associated with a higher value (Samejima et al., 2005; Hikosaka et al., 2006). We found significant levels of action value signals before animal's choice of action, which is consistent with previous reports on the monkey DS (Samejima et al., 2005; Lau and Glimcher, 2008; also see Pasquereau et al., 2007). Collectively, our results are most consistent with the possibility that the striatum contributes to action selection only indirectly by conveying values associated potential choices and actual action selection takes place elsewhere.

Role of striatum in action evaluation

We obtained several lines of evidence indicating the role of striatum in evaluating the consequences of past choices. First, choice and reward signals were conjunctively coded by many striatal neurons (but see Lau and Glimcher, 2007), indicating that individual striatal neurons get access to the information on animal's choice as well as its outcome. Second, current and previous choice signals were concurrently conveyed while the animal approached the reward location. Because the animal had already made a choice before these periods, the previous choice signal might be used for the purpose of evaluating the outcome of the past choice, raising the possibility that the striatum might evaluate consequences of multiple actions. Third, during the reward stage, many neurons conveying chosen value signals also encoded the animal's choice and its outcome, indicating the convergence of choice, reward and value signals in the same neurons. Thus, the striatum is likely to be a place where the value of the chosen action estimated in the previous trial is combined with the outcome of animal's choice in the current trial. These results provide converging evidence that the DS and VS are involved in evaluating the consequences of committed actions.

Value-related signals were not static within a given trial, but underwent state-dependent changes across different behavioral stages. Our results suggest that neural signals for choice, reward, and chosen value might be integrated in the striatum to update the value of chosen action during the early reward stage. We found two types of neuronal activity that was correlated more strongly with RPE or updated chosen value. These two types of neurons displayed appropriate signs of coefficients related to reward and chosen value and consisted of different proportions of putative MSNs versus interneurons, suggesting that such distinction was not an outcome of spurious correlation. Thus, our results provide converging evidence for the existence of both types of neurons in the striatum. This result is consistent with the role of striatum in updating value, because the process of computing RPE as well as combining this signal with the existing value are required to update value. A recent study has shown that neurons in the globus pallidus (GP) projecting to the lateral habenula (LH) convey RPE signals in monkeys (Hong and Hikosaka, 2008), raising the possibility that RPE-related neural activity in the striatum contributes to RPE signals in the GP. It remains to be determined, however, how the network of subcortical structures involving the striatum, GP, LH, and midbrain dopamine neurons work together to compute RPE and update chosen value (Hikosaka et al., 2008). Although dopamine neurons have long been thought to broadcast RPE signals to widespread areas of the brain, they are limited in conveying quantitative RPE signals in the negative domain (Morris et al., 2004; Bayer and Glimcher, 2005). Thus, it is difficult to explain bidirectional RPE signals found in the present study based on dopaminergic projections. Dopaminergic inputs to the striatum might contribute only to positive RPE signals. Alternatively, RPE signals in the striatum might be computed independently from dopaminergic neuronal activity, which might be also involved in functions other than updating values, such as incentive motivation (Berridge, 2007).

Functions of dorsal versus ventral striatum

Despite distinct anatomical connectivity patterns for the DS and VS (Alexander and Crutcher, 1990a; Voorn et al., 2004), their respective functions in the control of behavior are unclear. According to a popular view (Alexander and Crutcher, 1990a; DeLong, 2000), the BG and cortex form parallel loops serving different functions, and the VS, as part of the motivational loop, is primarily involved in reward processing and motivational control of behavior (Mogenson et al., 1980; Salamone and Correa, 2002; O'Doherty, 2004). However, recent studies have found strong reward- and value-related neural activity in the DS (for review, see Daw and Doya, 2006; Hikosaka et al., 2006; Schultz, 2006), indicating that the DS is also heavily involved in reward and value processing. Our results also show that choice, reward and chosen value signals are all stronger in the DS, suggesting that the DS might play a more prominent role in updating value than the VS. Combined with recent studies (for review, see Hikosaka et al., 2006; Schultz, 2006; see also Kimchi and Laubach, 2009), our results clearly argue against the view that the DS and VS are exclusively involved in motor/cognitive functions and reward processing, respectively. It remains to be determined what specific roles the DS and VS play, and how they work together in evaluating past choices and determining the desirability of future actions. In addition, the anatomical and functional properties of cortical areas projecting to the DS and VS show substantial differences between rodents and primates (Seamans et al., 2008; Wise, 2008). Therefore, it would be important for future studies to compare the functional organization in the striatum of rodents and primates.

References

Albin RL, Young AB, Penney JB (1989) The functional anatomy of basal ganglia disorders. Trends Neurosci 12:366–375.

Alexander GE, Crutcher MD (1990a) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13:266–271.

Alexander GE, Crutcher MD (1990b) Preparation for movement: neural representations of intended direction in three motor areas of the monkey. J Neurophysiol 64:133–150.

Apicella P, Scarnati E, Ljungberg T, Schultz W (1992) Neuronal activity in monkey striatum related to the expectation of predictable environmental events. J Neurophysiol 68:945–960.

Atallah HE, Lopez-Paniagua D, Rudy JW, O'Reilly RC (2007) Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. Nat Neurosci 10:126–131.

Baeg EH, Kim YB, Jang J, Kim HT, Mook-Jung I, Jung MW (2001) Fast spiking and regular spiking neural correlates of fear conditioning in the medial prefrontal cortex of the rat. Cereb Cortex 11:441–451.

Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW (2003) Dynamics of population code for working memory in the prefrontal cortex. Neuron 40:177–188.

Baeg EH, Kim YB, Kim J, Ghim JW, Kim JJ, Jung MW (2007) Learning-induced enduring changes in functional connectivity among prefrontal cortical neurons. J Neurosci 27:909–918.

Balleine BW, Delgado MR, Hikosaka O (2007) The role of the dorsal striatum in reward and decision-making. J Neurosci 27:8161–8165.

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

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141.

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191:391–431.

Cardinal RN, Parkinson JA, Hall J, Everitt BJ (2002) Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev 26:321–352.

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

Delgado MR (2007) Reward-related responses in the human striatum. Ann N Y Acad Sci 1104:70–88.

Delgado MR, Li J, Schiller D, Phelps EA (2008) The role of the striatum in aversive learning and aversive prediction errors. Philos Trans R Soc Lond B Biol Sci 363:3787–3800.

DeLong MR (2000) The Basal Ganglia. In: Principles of neural Science, Ed 4 (Kandel ER, Schwartz JH, Jessell TM, eds), pp 853–867. New York: McGraw-Hill.

Hikosaka O (1994) Role of basal ganglia in control of innate movements,

learned behavior and cognition - a hypothesis. In: The basal ganglia IV: new ideas and data on structure and function (Percheron G, McKenzie JS, Féger J, eds), pp 589–596. New York: Plenum.
Hikosaka O, Sakamoto M, Usui S (1989) Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements. J Neurophysiol 61:780–798.
Hikosaka O, Nakamura K, Nakahara H (2006) Basal ganglia orient eyes to reward. J Neurophysiol 95:567–584.
Hikosaka O, Bromberg-Martin E, Hong S, Matsumoto M (2008) New insights on the subcortical representation of reward. Curr Opin Neurobiol 18:203–208.
Histed MH, Pasupathy A, Miller EK (2009) Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. Neuron 63:244–253.
Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309.
Hong S, Hikosaka O (2008) The globus pallidus sends reward-related signals to the lateral habenula. Neuron 60:720–729.
Huh N, Jo S, Kim H, Sul JH, Jung MW (2009) Model-based reinforcement learning under concurrent schedules of reinforcement in rodents. Learn Mem 16:315–323.
Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci 29:9861–9874.
Kawagoe R, Takikawa Y, Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. Nat Neurosci 1:411–416.
Kim S, Hwang J, Seo H, Lee D (2009) Valuation of uncertain and delayed rewards in primate prefrontal cortex. Neural Netw 22:294–304.
Kim YB, Huh N, Lee H, Baeg EH, Lee D, Jung MW (2007) Encoding of action history in the rat ventral striatum. J Neurophysiol 98:3548–3556.
Kimchi EY, Laubach M (2009) Dynamic encoding of action selection by the medial striatum. J Neurosci 29:3148–3159.
Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579.
Lau B, Glimcher PW (2007) Action and outcome encoding in the primate caudate nucleus. J Neurosci 27:14502–14514.
Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463.
Lavoie AM, Mizumori SJ (1994) Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. Brain Res 638: 157–168.
Lee D, Conroy ML, McGreevy BP, Barraclough DJ (2004) Reinforcement learning and decision making in monkeys during a competitive game. Brain Res Cogn Brain Res 22:45–58.
Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24:415–425.
Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381–425.
Mogenson GJ, Jones DL, Yim CY (1980) From motivation to action: functional interface between the limbic system and the motor system. Prog Neurobiol 14:69–97.
Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 43:133–143.
Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models, Ed 4. Boston: McGraw-Hall.
Nicola SM, Yun IA, Wakabayashi KT, Fields HL (2004) Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task. J Neurophysiol 91:1840–1865.
O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol 14:769–776.
Packard MG, Knowlton BJ (2002) Learning and memory functions of the Basal Ganglia. Annu Rev Neurosci 25:563–593.
Pasquereau B, Nadjar A, Arkadir D, Bezard E, Goillandeau M, Bioulac B, Gross CE, Boraud T (2007) Shaping of motor responses by incentive values through the basal ganglia. J Neurosci 27:1176–1183.
Pasupathy A, Miller EK (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433:873–876.
Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89:1009–1023.
Roitman MF, Wheeler RA, Carelli RM (2005) Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. Neuron 45:587–597.
Salamone JD, 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, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340.
Schmitzer-Torbert N, Redish AD (2004) Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. J Neurophysiol 91:2259–2272.
Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1–27.
Schultz W (2006) Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57:87–115.
Seamans JK, Lapish CC, Durstewitz D (2008) Comparing the prefrontal cortex of rats and primates: insights from electrophysiology. Neurotox Res 14:249–262.
Seo H, Lee D (2007) Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci 27:8366–8377.
Seo H, Lee D (2008) Cortical mechanisms for reinforcement learning in competitive games. Philos Trans R Soc Lond B Biol Sci 363:3845–3857.
Seo H, Barraclough DJ, Lee D (2009) Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J Neurosci 29:7278–7289.
Setlow B, Schoenbaum G, Gallagher M (2003) Neural encoding in ventral striatum during olfactory discrimination learning. Neuron 38:625–636.
Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA, Poldrack RA (2004) Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. Brain 127:851–859.
Sutton RS, Barto AG (1998) Reinforcement learning: An introduction. Cambridge MA: MIT.
Voorn P, Vanderschuren LJ, Groenewegen HJ, Robbins TW, Pennartz CM (2004) Putting a spin on the dorsal-ventral divide of the striatum. Trends Neurosci 27:468–474.
Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629–632.
White NM (2009) Some highlights of research on the effects of caudate nucleus lesions over the past 200 years. Behav Brain Res 199:3–23.
Wise SP (2008) Forward frontal fields: phylogeny and fundamental function. Trends Neurosci 31:599–608.