HUMAN AND ANIMAL BEHAVIOR is characterized as choosing a desirable action among many alternatives in pursuit of distant goals. Reinforcement learning theories describe an algorithm that estimates the values of actions to achieve specific goals yielding maximum future rewards (Sutton and Barto 1998). The role of the basal ganglia in reinforcement learning has been supported by a growing body of evidence. For example, neurons in the striatum represent the values of actions (Hori et al. 2009; Lau and Glimcher 2008; Lauwereyns et al. 2002; Samejima et al. 2005), which are supposed to be updated by reward prediction error signals from midbrain dopaminergic neurons (Bayer and Glimcher 2005; Satoh et al. 2003; Schultz 1998) in terms of cortico-striatal synaptic plasticity (Calabresi et al. 1996; Reynolds et al. 2001; Shen et al. 2008). Neuronal activity that reflected delay discounting of a single reward was observed in the frontal and parietal cortices (Kim et al. 2008; Louie and Glimcher 2010) and striatum (Cai et al. 2011). Human brain imaging revealed neuronal correlates of long-term reward values in the frontal and parietal cortices and basal ganglia (Kable and Glimcher 2007; McClure et al. 2004; Tanaka et al. 2004). Midbrain dopaminergic neurons were shown to encode the discounted values of a delayed single reward (Fiorillo et al. 2008; Roesch et al. 2007).

In reinforcement learning theories, it is critical to evaluate the intermediate steps of action for expected multiple future rewards and to guide the earlier steps of action toward the final step, which is known as the temporal credit assignment problem (Sleeman et al. 1982; Sutton and Barto 1998). Recently, midbrain dopaminergic neurons were shown to encode the discounted sum of multiple future rewards through a series of behavioral choices, suggesting a fundamental roles of dopaminergic neurons in guiding multistep choices to achieve specific distant goals (Enomoto et al. 2011). Nevertheless, it remains unclear whether and how striatal neurons, which are the major targets of dopaminergic neurons, represent the long-term values of multiple future rewards, despite the proposed role of the basal ganglia according to the reinforcement learning model.

We explored the neural correlates of long-term reward values in the striatum using a multistep choice task in which monkeys earned multiple rewards through a series of choices. Two monkeys searched for a reward target among three alternatives by trial-and-error and earned a reward; then they gained additional rewards by choosing the same target again. We found that the firing rate of a subset of striatal neurons was positively correlated with the long-term reward values estimated from a standard reinforcement learning algorithm, while that of another subset of neurons, in a similar proportion, was negatively correlated. The negatively correlated neuronal activity represented the values of highly discounted future rewards, while the positively correlated activity represented long-term values with low discounting. Our results provide new insights into specific and pivotal roles of the basal ganglia in decision and action selection for immediate and distant goals.

MATERIALS AND METHODS

Experimental Animals

We used two Japanese monkeys (Macaca fuscata; monkey RO, male, 9.4 kg and monkey TN, female, 6.3 kg). All animal procedures were approved by the Animal Care and Use Committee of Kyoto Prefectural University of Medicine and conformed to the United States National Institutes of Health guidelines. All surgeries were carried out under sterile conditions with the monkeys under deep pentobarbital sodium anesthesia. Anesthesia was induced with ket-
amine hydrochloride (10 mg/kg im), then with pentobarbital sodium (Nembutal, 27.5 mg/kg ip). Four head-restraining bolts and one stainless-steel recording chamber were implanted under stereotaxic guidance on the skull of each monkey. The chamber, for recording neuronal activity in the caudate nucleus and putamen, was placed laterally at a 45° angle. The center of the chamber was adjusted according to the Horseley-Clark stereotaxic coordinates (Kusama and Mabuchi 1970): lateral 10 mm (L10), anterior 18 mm (A18), and height 9 mm (H9) in the left hemisphere.

Behavioral Task

The monkeys sat on a primate chair facing a wood panel placed at 20.5 cm in front of their faces. On the panel (Fig. 1A), we embedded a small, rectangular button with a green light-emitting diode (LED) at the bottom (Start LED button, 14 × 14 mm), 3 target buttons with green LEDs in the middle row (Target LED buttons, 14 × 14 mm), and a small red LED just above the center Target LED button (GO LED, 9 mm diameter). A task trial was initiated by the illumination of the Start LED. The monkeys depressed the illuminated start button with their right hand, contralateral to the neuronal recording site. After the GO LED was illuminated for 0.7–0.9 s, the three Target LEDs were turned on. The monkeys were required to keep the start button depressed for an additional 0.6–0.7 s until the GO LED was turned off and then to release the start button and depress one of the three illuminated target buttons within 3.0 s. They were required to keep the target button depressed until a feedback beep sounded. If the reward target button was depressed, a high-tone beep (1,000 Hz for 0.2 s) sounded with a delay of 0.9–1.0 s as positive feedback, and reward water (0.032 ml/kg wt, i.e., 0.3 and 0.2 ml for monkeys RO and TN, respectively) was delivered at 0.4 s after the offset of the beep sound through a spout attached to the mouth of each monkey. If a no-reward target button was depressed, a low-tone beep (300 Hz for 0.2 s) sounded as negative feedback, and no reward was given.

The monkeys performed reward-based multistep choices using trial-and-error and repetition approaches (Fig. 1B). If the monkeys depressed a no-reward target button in the first step (S1), the second step of the search trial began with the illumination of the Start LED at 5.0 s (intertrial interval) after the monkeys had released the target button chosen in the first step. The three targets were again illuminated, and the monkeys chose once more. The monkeys were required to remember the no-reward target chosen during the first step and make another choice between the two remaining target buttons in the second step (S2 trial). If they depressed another no-reward target button, the negative feedback beep sounded, and the third step (S3 trial) started. They remembered the two previously tried no-reward targets and depressed the remaining single-target button.

Once the monkeys chose the reward target button in any search trial (S1–S3), they obtained an immediate reward, in contrast to the future rewards expected during the succeeding repeat trials (R1 and R2). The same button was used again as the reward target in the following repeat trials. The monkeys received a water reward once during the search trials (S1, S2, or S3 trials) and an additional one (monkey RO) or two (monkey TN) rewards during the repeat trials (R1 and R2 trials). Thus the monkeys received rewards twice (monkey RO) or three times (monkey TN) in a single series of choices. To instruct the monkeys that a single series of choices had ended, all four green LEDs were simultaneously flashed for 1.0 s at 2.0 s after the monkeys released the target button during the final repeat trial. The next series of trials began at 5.0 s after the flash with a new reward target at a random location. Note that the probability of choosing a reward target button was one-third in the S1 trials.

The monkeys were trained, first, on choices between two target options with search and repeat trials. It took ~3 mo before they mastered the task requirements. Then the three target options were introduced. It took an additional 3 mo of training for the monkeys to master the multistep task with three targets.

Data Recording

We recorded single-neuron discharges from the dorsal part of the caudate nucleus and putamen in the left hemisphere of the two monkeys at the rostrocaudal level between A13 and A28. We used epoxy-coated tungsten microelectrodes and a template-matching algorithm to isolate discharges from single neurons. Striatal projection neurons (Yamada et al. 2007) and tonically active neurons (cholinergic interneurons) (Inokawa et al. 2010; Yamada et al. 2004) were identified according to their background discharge rates and action potential waveforms (Kimura et al. 1990). Presumed projection neurons showed low spontaneous firing rates (<2 spikes/s) and phasic firing.
discharges in relation to one or more task events (Kimura et al. 1990). The onset times of the action potentials were recorded on a laboratory computer, together with the onset and offset times of stimuli and the events that occurred during task performance. Discharges from single neurons were recorded during one session that contained ~135 and 180 trials (45 new reward target positions for a series of choices with 2–4 and 3–5 trials) in monkeys RO and TN, respectively. The following three behavioral measures were recorded: task start time (TST: time from the illumination of the Start LED to the depression of the start button), reaction time to the GO signal (GORT: time from the GO signal to the release of the start button), and movement time (MT: time from the release of the start button to the depression of a target button). As control data for task performance, we recorded electromyographic activity using chronically implanted Teflon-coated stainless steel electrodes into the forearm and upper arm muscles. Eye movements were also monitored by measuring the corneal reflections of an infrared light beam via a video camera. Orfacial movements were detected using a strain gauge attached to the drinking spout. The analog signals of electromyographic activity, eye position, and orfacial movements were recorded at a sampling rate of 100 Hz. All of the recordings were made extensively after the monkeys had fully mastered the task performance, i.e., in 1 wk, >80% reward trials at S3 and >90% at R1 and R2, respectively.

**Data Analysis**

**Simulation of the long-term reward values by using a standard reinforcement learning model.** The expected values of discounted future rewards were estimated by using a standard reinforcement learning model, i.e., the temporal difference model (Sutton and Barto 1998), as follows:

\[
V(S_t) \leftarrow V(S_t) + \alpha[r + \gamma V(S_{t+1}) - V(S_t)]
\]  

where the reward value \(V(S_t)\) is composed of four or five different trial types, i.e., S1, S2, S3, and R1 in monkey RO, and S1, S2, S3, R1, and R2 in monkey TN (Fig. 1, B and D). \(S_t\) is the \(r^n\) trial/step in the simulation. \(r\) is the amount of reward, i.e., 0.3 and 0.2 ml in monkeys RO and TN in the reward trials, respectively, while it was zero in the no-reward trials. \(\alpha\) is the learning rate; 0 ≤ \(\gamma\) ≤ 1. \(\gamma\) is the discount rate; 0 ≤ \(\gamma\) ≤ 1. The default value of \(V(S_t)\) is zero for all trial types. \(V(S_t)\) was estimated as a function of the transition diagram for each monkey (Fig. 1, B and D) from their series of choices until they obtained multiple rewards. The value of \(V(S_t)\) in each trial type was extracted after 500 trials/steps. The average value of \(V(S_t)\) in each trial type was estimated from 1,000 simulations for each \(\gamma\) from 0–0.9 in 0.1 steps (Fig. 1, C and E). The \(\alpha\) value was set as 0.2, because the different settings of \(\alpha\) from 0.01 to 1.0 were shown to affect the results only slightly (Enomoto et al., 2011); larger values of \(\alpha\) gradually increased the variance. The long-term reward values were estimated using the same model as Enomoto et al. (2011).

**Statistical analysis of neuronal activity.** The statistical analysis of neuronal activity in the present study was the same as that used in our previous study (Yamada et al. 2011) using the same data set. Briefly, a significant increase in the discharge rate of each task period shown in Fig. 1A was determined by comparing the discharge rate during a 150-ms test window with the background rate for the 750 ms before the onset of the Start LED (Wilcoxon two-sample test, \(P < 0.05\)). In addition, we applied a criterion for significant activity as an average increase of the discharge rate of more than 2.0 spikes/s to avoid including neurons exhibiting very weak activity with almost no background activity (Kimura 1990). A subset of neurons showed discharges during the post-feedback period that were time-locked to the arm movement to release the target button rather than to the feedback beeps. In these neurons, histograms for the neuronal discharges were constructed before and after the arm movement to release the target button.

**Regression analysis: variable (or model) selection based on Akaike’s information criterion or Bayesian information criterion.** The influence of the long-term reward values on the neuronal discharge rates was examined by a variable (or model) selection method using statistical software “R” (http://www.r-project.org/). We obtained the best discount factor, \(\gamma\) (0–0.9), to estimate the neuronal discharge rates as well as the other variables described later. We compared goodness of fit based on Akaike’s Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Burnham and Anderson 2002, 2004):

\[
\text{AIC(Model)} = -2L + 2k
\]

\[
\text{BIC(Model)} = -2L + k \ln(n)
\]

where \(L\) is the maximum log-likelihood of the model, \(k\) is the number of free parameters, and \(n\) is the sample size of the model. We detected the combination of variables that showed the smallest AIC or BIC for the model. Neuronal discharge rates \(F\) were fitted by the following model:

\[
F = b_0 + \sum b_i \gamma^{y_i} = \sum b_i \gamma^{y_i}
\]

where \(b_0\) and error are the intercept and residual, respectively. All possible variables that could explain the neuronal discharge rates were included in the selection of the model as follows. \(\sum b_i \gamma^{y_i}\) are the long-term reward values estimated for \(\gamma\) values from 0 to 0.9 (Fig. 1, C and E). Search-Repeat took scalar values (1 or 0) in the search-and-repeat trials (Fig. 1, B and D), respectively; this variable was included because we demonstrated previously that striatal neuronal activity was modulated by this factor (Yamada et al. 2011). The 11 variables in the square brackets explained the neural modulation that was specific to each trial type from S1 to R2; only one of these variables was selected. Feedback took scalar values in the reward (1) and no-reward (0) trials. The Feedback term was included only during the multi-trial period. Target took scalar values (1, 0, –1) for the three target options chosen by the monkeys. These values were assigned, depending on the average discharge rates of each target. TST, GORT, and MT were also included in the model to detect the effects of behavioral variables.

We selected one combination of variables (or model) that showed the minimal AIC or BIC among all possible models. We sought the best discount rate (\(\gamma\) values from 0 to 0.9) to explain the neuronal firing rates (i.e., parameter estimation), as well as examining which decision variables, such as choices, task strategy, and so on, are represented by the neuronal discharge rates (see Supplemental Information in Padoa-Schioppa and Assad 2006 for the reason why we used the variable selection method). Our aim was to select one model from all possible combinations, as well as to obtain the best-fit discount rate. For this purpose, we used the following two-step search procedure. We first estimated the AIC or BIC in each of the 11 full models that contained 7 variables (i.e., intercept, one of the variables in the square brackets, Feedback, Target, TST, GORT, and MT), and defined which combination of variables showed the smallest AIC or BIC in each of the 11 models (last models). This allowed us to examine how smoothly the goodness of fit changed across the \(\gamma\) values (Fig. 2, D and H); hence this is almost equivalent to estimating the parameters for the discount rate (from 0 to 0.9). Then we selected one combination of variables (or model) with the smallest AIC or BIC (asterisk in Fig. 2, D and H), as well as the best-fit discount rate. Note that the selected model would include no variables, if neuronal activity was not modulated by any of these factors. Each neuronal activity was, in principle, explained by a different model as a result of model selection and composed of a different number of explanatory variables. We did not include the no-reward trials in S3, R1, and R2.
Our results from the model selection and parameter estimation used in the present study were almost consistent with the results that were generated when we used conventional statistical methods, such as multiple-regression analysis and ANOVA (see Table 2 for details). There was no significant difference in the number of modulated activity between ANOVA and BIC-based model selection for Trial Type (ANOVA, 215/528, BIC, 229/528, Fisher’s exact probability test, \( P = 0.418 \)) and target choice (ANOVA, 261/528, BIC, 278/528, \( P = 0.325 \)). In contrast, the activity of the neurons modulated by Trial Type and target choice was more frequently observed in AIC-based model selection (Trial Type, 273/528, \( P < 0.001 \), target choice, 310/528, \( P = 0.003 \)) than in ANOVA.

**Statistical analysis of behavioral performance.** Behavioral data were obtained from a total of 98,604 and 61,299 trials in 583 and 303 recording sessions in monkeys RO and TN, respectively. The reward probability in each trial type (i.e., S1, S2, S3, R1, and R2) was determined as the ratio of the number of rewarded trials to the total number of trials during each recording session. TSTs were compared among individual trial types using multiple two-sample comparisons corrected by the Bonferroni method to control the familywise error rate. Orofacial movements were monitored as the torque of the tongue on the water supply spout. The amplitude of licking was averaged as a function of trial type and compared using multiple two-sample comparisons corrected by the Bonferroni method.

**RESULTS**

**Task Performance and Estimation of Long-term Reward Values**

Two monkeys searched for a reward target among three alternatives on a trial-and-error basis (Fig. 1A) and obtained a reward; then they received additional rewards by choosing the same target again. The average probability of selecting a reward target during the search trials increased progressively from the first (S1, 33 and 32%), second (S2, 50 and 49%), and third choices (S3, 89 and 82%) (Fig. 1, A–D, monkeys RO and TN, respectively). The reward probability in the S3 trials was not 100%, because the monkeys often chose the no-reward target that was initially selected in the S1 trial. Once the monkeys found the reward target button, they obtained one (R1 in monkey RO) or two (R1 and R2 in monkey TN) additional...
rewards by choosing the same target button. Note that any search trial will lead to the immediate delivery of a reward if the monkeys chose the reward target in their series of choices.

If the animals expected only a single upcoming reward in each trial through multiple-step choices, the reward expectation simply reflects the reward probability and magnitude that the monkeys experienced in the current trial (from S1 to R2). Conversely, if they expect that they will receive multiple rewards through a series of choices, the reward expectation must reflect the multiple rewards that are expected from future trials. We estimated the long-term reward values through a series of choices in each trial type using a standard reinforcement learning model (see MATERIALS AND METHODS). The long-term reward values were estimated for each discount rate, \( \gamma \), ranging from 0 to 0.9 in 0.1 increments based on the average reward rate that the monkeys experienced throughout all of the

<table>
<thead>
<tr>
<th>Task Period</th>
<th>Long-term Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>BIC (+) 110223401 14</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 110223401 14</td>
</tr>
<tr>
<td>Pre-GO</td>
<td>BIC (+) 031222220 13</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 031222220 13</td>
</tr>
<tr>
<td>Target choice</td>
<td>BIC (+) 111122211 18</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 111122211 18</td>
</tr>
<tr>
<td>Pre-feedback</td>
<td>BIC (+) 111122211 18</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 111122211 18</td>
</tr>
<tr>
<td>Post-feedback</td>
<td>BIC (+) 111122211 18</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 111122211 18</td>
</tr>
<tr>
<td>Total</td>
<td>BIC (+) 111122211 18</td>
</tr>
<tr>
<td></td>
<td>BIC (-) 111122211 18</td>
</tr>
</tbody>
</table>

Values representing the number of neurons categorized as Long-term value, Current value, Search-Repeat, Feedback, Target, and Nonmodulated type in each task period are shown. AIC, Akaike’s Information Criterion; BIC, Bayesian Information Criterion; \( \gamma \), discount rate; (+) and (−), positive and negative correlation, respectively.

Table 2. Number of neuronal activity with distinct discharge properties defined by ANOVA and model selection based on AIC and BIC

<table>
<thead>
<tr>
<th>Task Period</th>
<th>n</th>
<th>Trial Type, no. (%)</th>
<th>Target, no. (%)</th>
<th>Interaction, no. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>124</td>
<td>ANOVA 40 (32.2)</td>
<td>24 (19.4)</td>
<td>27 (21.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIC 53 (42.7)</td>
<td>32 (25.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIC 64 (51.6)</td>
<td>41 (33.1)</td>
<td></td>
</tr>
<tr>
<td>Pre-GO</td>
<td>124</td>
<td>ANOVA 45 (36.3)</td>
<td>75 (60.5)</td>
<td>10 (8.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIC 43 (34.7)</td>
<td>76 (61.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIC 55 (44.4)</td>
<td>81 (51.6)</td>
<td></td>
</tr>
<tr>
<td>Target choice</td>
<td>157</td>
<td>ANOVA 52 (39.4)</td>
<td>100 (63.7)</td>
<td>13 (8.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIC 54 (43.4)</td>
<td>101 (64.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIC 67 (42.7)</td>
<td>111 (70.7)</td>
<td></td>
</tr>
<tr>
<td>Pre-feedback</td>
<td>123</td>
<td>ANOVA 78 (63.4)</td>
<td>62 (50.4)</td>
<td>9 (7.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIC 79 (64.2)</td>
<td>69 (56.1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIC 87 (70.7)</td>
<td>77 (62.6)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>528</td>
<td>ANOVA 215 (40.7)</td>
<td>261 (49.4)</td>
<td>59 (11.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIC 229 (43.4)</td>
<td>278 (52.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIC 273 (51.7)</td>
<td>310 (58.7)</td>
<td></td>
</tr>
</tbody>
</table>

Values representing the number of neuronal activity modulated by Trial Type (S1, S2, S3, R1, and R2), Target Choice (Left, Middle, and Right), and their interaction are shown. ANOVA and model selection based on AIC and BIC are shown. In model selection based on BIC and AIC, Trial Type indicates the neuronal activity categorized as “Long-term Value,” “Current Value,” or “Search-Repeat.” In this analysis, 128 post-feedback activities were excluded because we were unable to analyze these data with ANOVA (a singular matrix occurred in most neurons due to the absence of negative feedback in the S3, R1, or R2 trials).
experimental sessions (Fig. 1, B and D). While the number of rewards earned across choices was different between the monkeys, the estimated long-term reward values, $V(S_t)$, gradually increased from the S1 to S3 trials and then decreased in the R1 and R2 trials after the monkeys consumed those rewards (Fig. 1, C and E). If the $\gamma$ values are large (low discounting), the estimated long-term reward values will become larger compared with those with smaller $\gamma$ values. If $\gamma$ is zero (extremely high discounting), the estimated reward values will be identical to the reward values experienced in the current trial (reward probability $\times$ volume).

**Representation of the Long-Term Reward Values in the Striatum**

We recorded the activity of 292 presumed projection neurons (145 from monkey RO and 147 from monkey TN) in the caudate nucleus ($n = 118$) and putamen ($n = 174$). A significant increase of the discharge rates occurred during the start, pre-GO, target choice, pre-feedback, and post-feedback periods in 124, 124, 157, 123, and 170 neurons, respectively. This is the same data set that we used in a previous study (Yamada et al. 2011). Among the 170 neurons activated during the post-feedback period, 128 were examined quantitatively because the remaining 42 neurons, which were activated in relation to the release of the target button after feedback, were not used for the analysis presented in our previous paper.

The example neuron presented in Fig. 2A showed a phasic increase of its discharge rate before the monkey depressed the start button following the illumination of the Start LED. The activity of this neuron was highest in the S1 trial and decreased in the S2 and S3 trials as the reward probability increased (Fig. 2, B and C). However, during the repeat trials (R1), where the reward probability was the highest, its activity was higher than during the S3 trials. The activity of this neuron appeared to be inversely correlated with the long-term reward values estimated from the standard reinforcement learning model (Fig. 1, C and E). To examine whether the striatal neurons represent long-term values with the distant rewards discounted, we obtained the best-fit $\gamma$ values (from 0 to 0.9 in 0.1 increments) for the activity of each neuron. This is one of the convincing methods that have been used to find the representation of long-term values with heterogeneous discount rates (Tanaka et al. 2004), although it is a different approach to find the neural coding of subjective values of rewards estimated from the history of behavioral choices and their outcomes (Daw et al. 2006; Lau and Glimcher 2008; Samejima et al. 2005). In this example (Fig. 2, A–D), the long-term reward value estimated with $\gamma = 0.3$ showed the smallest AIC (not shown) and BIC values [Fig. 2D, $R^2 = 0.257$, the best model: “6.88 intercept $- 5.85 \text{ Value} (\gamma = 0.3) + 0.81 \text{ Target} + 0.013 \text{TST} - 0.0082 \text{ MT}””]. Note that the BIC estimated in each full model ($\gamma = 0.0$–0.9) before model selection decreased due to the decrease of the variables in the last model through selection (Fig. 2D). Another example neuron shown in Fig. 2, E–H, showed an increase in its discharge rate after the illumination of the Start LED. This activity was also negatively correlated with the estimated long-term reward value, but the best-fit $\gamma$ value

(AIC, $\gamma = 0.6$, $R^2 = 0.192$; BIC, $\gamma = 0.5$, $R^2 = 0.169$) was higher than for the neuron shown in Fig. 2A.

We also found neurons showing activity that was positively correlated with the long-term reward values. Four such neurons are shown in Fig. 3 during four task periods, i.e., preparation of target choice (Fig. 3, A and B), during target choice (Fig. 3, C and D), after choice and before outcome feedback (Fig. 3, E and F), and after feedback and during reward outcome (Fig. 3, G and H). Neurons encoding the reward values of current trials ($\gamma = 0$) were also observed (Current Value type; Fig. 4). This activity reflects the reward values expected in the current trial; thus it is a different
kind of response from that of neurons encoding rewards expected not only in the current trial, but also in multiple future trials with discounted future rewards. Activity reflecting the current value was observed during 57 task periods out of 656 (8.7%, BIC based, see Fig. 8 for details).

Approximately 25% of the neuronal activity that occurred in each task period was correlated with the estimated long-term reward values with $\gamma$ from 0.1 to 0.9 (Long-term Value type) [Fig. 5A, AIC (not shown): 27.4%, BIC: 22.1%]. There was no significant difference in the proportion of modulated neuronal activity across the task periods ($\chi^2$ test, AIC: $P = 0.196$, BIC: $P = 0.270$). The activity was either positively or negatively correlated with the long-term reward values in a similar proportion [Fig. 5B, $\chi^2$ test, AIC (not shown): $P = 0.569$, BIC: $P = 0.626$]. Thus the activity of a subset of neurons was either positively or negatively correlated with the long-term reward values during each task period.

We found that neuronal activity during 80 task periods (67 neurons) showed a positive correlation, while activity during 65 task periods (57 neurons) showed a negative correlation (Fig. 6). The distribution of the discounting functions (factors) for both positively and negatively correlated neuronal activity revealed that the positive coding neurons had larger $\gamma$ values (low discounting, $0.592 \pm 0.025$, mean and SE) than the negative coding neurons (high discounting, $0.406 \pm 0.033$, mean and SE) [Fig. 6, two-way ANOVA (correlation type and task period): $P < 0.001$ for correlation type, $P = 0.981$ for task period, $P = 0.815$ for their interaction, see Table 1 for details]. In other words, the neurons showing negative correlations expect rewards in the near future, while the neurons showing positive correlations expect rewards not only in the near future but also in the far future.

Among the 292 neurons exhibiting activation during 656 task periods, 110 (37.7%) showed activity that was either positively or negatively correlated with the long-term reward values ($\gamma 0.1–0.9$: Long-term Value type) in at least one task period. Of these 110 neurons, the majority consistently exhibited either positively or negatively correlated activity throughout the task periods (99/110, 90.0%). However, the other 10% of long-term value neurons (11/110), representing 3.8% of the total neuronal population (11/292), showed both positive and negative correlations during different task periods. Therefore, the encoding of long-term reward values with future rewards discounted at high and low rates was made mostly by the two subsets of striatal neurons as positive and negative correlations.

Location of the Neurons Encoding the Long-term Values Recorded in the Caudate Nucleus and Putamen

We examined the location of the neurons encoding the long-term values in the caudate nucleus and putamen. The neurons encoding the long-term values of multiple future rewards with $\gamma$ values from 0.1 to 0.9 did not show any predominant location on the anterior-posterior level, but were found throughout the striatum (Fig. 7A, $\chi^2$ test, $P = 0.139$).
activity encoding the current values (\( \gamma = 0 \)) and the search-repeat trials as a task strategy were observed predominantly during the pre-feedback period after the monkey had made its choice (Fig. 8A, \( \chi^2 \) test, AIC: \( P = 0.006 \), BIC: \( P = 0.001 \)). These results were almost the same as those from our previous study by using multiple-regression analysis (Yamada et al. 2011), except for one difference: neurons encoding the values and strategy during the post-feedback period were not observed frequently in this analysis. Thus the results appeared to be consistent among the different analyses.

Coding of Choices and Behavioral Measures

We also examined the influence of the target choice of the monkeys on the activity of striatal neurons, as we had examined this previously using multiple regression analysis. During preparation to choose the target and the subsequent execution of the arm movement, more than 60% of the activity of neurons encoding the long-term reward values was modulated by the chosen target (Fig. 8B). Target-dependent neurons were observed predominantly during target choice; then they gradually decreased, but were maintained until feedback appeared [\( \chi^2 \) test, AIC (not shown): \( P < 0.001 \), BIC: \( P < 0.001 \)]. The tendency of target modulation across behavioral events was similar in the entire neuronal population (Start, 25.8%, Pre-GO, 61.3%, Target Choice, 64.3%, Pre-feedback, 55.6%, Post-feedback, 29.7%). This result was consistent with our previous study using multiple regression analysis (Yamada et al. 2011).

We also assessed the effect of behavioral measures on neuronal activity, because the activity of striatal neurons is typically correlated with the behavioral parameters involved in decision making (Ito et al. 2003; Lau and Glimcher 2008). Three behavioral parameters were examined: TST, which indicates how fast the monkeys started the task trial; GORT, which indicates how fast the monkeys responded to the GO signal; and MT, which indicates the speed of the arm movement during target choice. The effects of each behavioral parameter were observed predominantly when the monkeys were performing those movements. For example, the effect of TST was observed predominantly during the start period (Fig. 8C), while the effects of GORT and MT were observed predominantly during the pre-GO and target choice periods, respectively.

DISCUSSION

In the present study, we provided direct evidence of long-term reward value coding through multistep choices for multiple rewards in the activity of striatal neurons. Temporally discounted values of single future rewards were observed in the neuronal activity of the lateral intraparietal cortex (Louie and Glimcher 2010), dorsolateral prefrontal cortex (Kim et al. 2008), orbitofrontal cortex (Roesch et al. 2006), and midbrain dopaminergic neurons (Fiorillo et al. 2008; Kobayashi and Schultz 2008). Signals for the discounted sum of multiple future rewards were also observed in the activity of dopaminergic neurons (Enomoto et al. 2011). In the monkey striatum, Lee and colleagues recently showed neurons encoding the values of a discounted single reward (Cai et al. 2011). In the present study, we found that monkey dorsal striatum neurons exhibit activity that is either positively or negatively correlated with the values of discounted multiple future rewards by adopting the same behavioral task as Enomoto et al. (2011). Our findings add a new critical piece of knowledge about how the values of future rewards are estimated in the basal ganglia.
Fig. 8. Effects of choice, behavioral parameters, and decision variables in the current trials. A: percentage of neuronal activity defined as Current Value ($\chi^2$ test, AIC: $P = 0.006$, BIC: $P < 0.001$) and by Search-Repeat type ($\chi^2$ test, AIC: $P = 0.002$, BIC: $P = 0.001$). The results shown are based on the BIC values. B: percentage of Long-term Value type activity that was modulated by target choice. The values in parentheses indicate the number of activity defined as Long-term Value in each task period. C: percentage of neuronal activity modulated by the behavioral parameters, such as task start time (TST), reaction time for the GO signal (GORT), and movement time for choices (MT). The results shown are based on the BIC values.

A reinforcement learning model of the basal ganglia (Doya 2000; Houk et al. 1995; O’Doherty et al. 2004) has been proposed in which reward values are represented in the striatum, actions are selected by comparing them, and the values are updated based on the reward prediction error signals from dopaminergic neurons. This model was supported by studies of the monkey caudate nucleus and putamen that revealed that the values of options are encoded in the activity of neurons in the striatum (Lau and Glimcher 2008; Samejima et al. 2005). Additional support was provided by studies on the activity of the rat striatum (Ito and Doya 2009; Kim et al. 2009).

In the present study, we found that the long-term values of multiple future rewards are represented in the activity of striatal neurons. This may be a neural basis for subjects selecting actions that are most likely to yield the maximum amount of total future rewards, as suggested by reinforcement learning theory (Sutton and Barto 1998). In the activity of striatal neurons, future rewards are discounted by factors ranging from 0 to 0.9. In a human brain imaging study, blood oxygen level-dependent signals showed regional differences within the striatum, indicating high discounting (small $\gamma$) in the ventro-anterior part of the striatum and low discounting (large $\gamma$) in the dorso-posterior striatum (Tanaka et al. 2004). In our study examining the activity of single neurons, in contrast, there was no apparent topographical segregation of encoding values with distinct discount rates. This may be because our recordings covered the dorsal striatum, but not the ventral striatum. However, a remarkable finding in this study was that neurons whose activity showed positive and negative correlations with long-term values had distinct properties in their discount rates. The activity of positively correlated neurons had low discounting of future rewards (large $\gamma$), whereas that of negatively correlated neurons had high discounting (small $\gamma$) (Fig. 6).

What could be the functional significance of the contrasting characteristics of the two subsets of striatal neurons? We propose one possibility in which these characteristics might reflect the distinct roles of the direct and indirect pathways of the basal ganglia circuit in action and cognition (Albin et al. 1989; Alexander and Crutcher 1990). The classical direct-and-indirect pathway model assumed that the direct pathway facilitates movement through the removal of steady inhibition, while activation of the indirect pathway inhibits movement through the enhancement of steady inhibition. Recently, this hypothesis was proved experimentally using optogenetically controlled stimulation methods. Optical stimulation of dopamine D2 receptor-expressing indirect pathway neurons in the striatum increased freezing in mice, while the stimulation of D1 receptor-expressing direct pathway neurons reduced freezing and increased locomotor behavior (Kravitz et al. 2010). Moreover, selective blockade of synaptic transmission through the direct and indirect pathways in the basal ganglia revealed their specific involvement in reward learning and the avoidance of aversive stimuli, respectively (Hikida et al. 2010). These observations imply that the neurons exhibiting activity that was positively correlated with the long-term values (i.e., increased firing at high reward values) might correspond to direct pathway neurons and play a role in action valuation with low discounting of future rewards (large $\gamma$), while the neurons exhibiting negatively correlated activity (i.e., decreased firing at high reward values) might correspond to indirect pathway neurons and play a role in action valuation with high discounting (small $\gamma$). This double dissociation may be functionally significant. For instance, for the exploration of rewards and distant goals through the direct pathway, it would be advantageous to predict and yield not only an immediately available reward, but also future rewards (low discounting). In contrast, for avoidance behavior through the indirect pathway, the evaluation of impending risk and punishment must be critical for immediate decisions and actions to escape damage without considering future events, even though they might be highly rewarding (high discounting).

The long-term values encoded in the striatal neurons are supposed to be evaluated and updated under the influence of dopaminergic signals because the striatum is the primary target of the dopaminergic innervations that convey signals for value-based learning (Bayer and Glimcher 2005; Satoh et al. 2003; Schultz 1998). It was recently shown that midbrain dopaminergic neurons encode predicted multiple future rewards (Enomoto et al. 2011). In dopaminergic neurons, the coding of long-term values was established in parallel with the process during which monkeys learn to understand and predict an
action schedule to obtain immediate and multiple future rewards, and the discount rate for future rewards was \( \sim 0.6 \) on average after the monkeys had fully learned the task. Our study revealed that the discount rates in the striatal neurons encoding long-term values as a whole were asymmetrically represented by the activity of neurons that were positively or negatively correlated with the long-term reward values. These two distinct value signals may be characteristically involved in the evaluation and choice of a variety of motor and cognitive activities via D1- and D2-like dopamine receptors through the direct and indirect pathways, respectively (Hikida et al. 2010; Nakamura and Hikosaka 2006). Thus the present study provides a new hypothesis that long-term value signals in the striatum with two distinct characteristics in light of future discounting may play vital roles in action selection through the direct and indirect pathways.

Previously, we examined the effects of other decision variables on striatal neuron activity (Yamada et al. 2011). We performed systematic regression analyses using the same data set as in the present study of the magnitude of changes in the discharge rate of all recorded neurons in relation to major factors involved in task performance, including reward probability, task strategy (search and repeat trials), feedback outcomes (positive and negative), and chosen target (left, middle, and right). We found that the activity of 8–24% of striatal neurons activated during different task epochs was specific to the task strategy adopted by the monkeys to perform individual trials (either search or repeat) (Figs. 8A and 10 in Yamada et al. 2011). A similar percentage of neurons (9–23%) were specific to the reward probability of individual trials. In the present study, we performed an additional analysis on the same data set with a special emphasis on the neuronal encoding of expected multiple future rewards and found two distinct populations of neurons: positive correlation neurons with larger discount factors, and negative correlation neurons with smaller discount factors. Therefore, these observations indicated that the two subsets of neurons encoding the expected value of multiple future rewards are different from the task strategy (search and repeat) and other major factors involved in performing our decision task.

**Effect of Decision Variables on Behavioral Measures**

During this task, a couple of decision variables affected the monkeys’ behavior. Previously, using the same behavioral task with monkeys (Muranishi et al. 2011; Yamada et al. 2011), we studied the TST of individual trials, i.e., the time from the onset of the start cue to the depression of the illuminated hold button, as conventionally used to measure the motivation to work for a reward (Lauwereyns et al. 2002; Satoh et al. 2003; Shidara et al. 1998; Watanabe et al. 2001). We found a negative correlation between the TST and the reward probability of individual choices (Fig. 3 in Muranishi et al. 2011), i.e., the higher the reward probability the shorter the TST. This relationship was monotonous, and there was no sign of a V shape in the S3 trials or an inverted V shape showing the neuronal encoding of expected multiple future rewards, as was found in the present study. Conversely, licking movements and the latency of licks after positive feedback did not change monotonously with trial type, but showed a peak in the S3 trials and decreased during the R1 and R2 trials (Fig. 12, B and D; Yamada et al. 2011). This supports the hypothesis that the monkeys’ behavior was affected by their long-term assessment of goals (discounted sum of multiple future rewards). Furthermore, using the same behavioral task with monkeys, we found that the duration of anticipatory licks preceding the feedback beeps showed a peak during the S3 trials, which then decreased during the R1 and R2 trials (Fig. 2; Enomoto et al. 2011). On the basis of these behavioral results, it is reasonable to conclude that the monkeys are making a long-term assessment of their goals during this multistep choice task.

**Representation of the Long-Term Value Signals in the Cortex and Basal Ganglia for Guiding Actions Toward Goals**

Neuronal representations of values for a single reward that was expected after a considerable delay (Kim et al. 2008; Louie and Glimcher 2010; Roesch et al. 2006) were observed during an intertemporal choice task (i.e., delayed discounting) in the cortical areas of rats and monkeys, that is, the parietal, dorsolateral prefrontal, and orbitofrontal cortices. What are the roles of these widespread temporally discounted value signals in the nervous system? It is of critical importance to understand how future rewards are discounted in the cortical circuits as well as the cortico-basal ganglia circuit. Neurons in the lateral intraparietal area represent the values for oculomotor choices (Platt and Glimcher 1999; Sugrue et al. 2005). Neurons in the orbitofrontal cortex are suggested to play a role in economic choices where values are compared within the space of goods, independent of the sensorimotor contingencies of choice (Padoa-Schioppa 2011; Padoa-Schioppa and Assad 2006). The striatum receives various kinds of cognition- and motor-related signals (Haber et al. 2006; Nambu et al. 2002; Selemon and Goldman-Rakic 1985; Takada et al. 2001) for the discounted values represented in those cortical areas. Thus the cortico-basal ganglia loops may play pivotal roles in action selection, goods selection, and cognitive decisions in concert with the cortico-cortical and other cortico-subcortical circuits.

Temporal discounting of neuronal value coding may be especially useful for the valuation and economic choice of the reward events expected with short (e.g., within a few seconds) and variable delays (Fiorillo et al. 2008). Striatal coding for the sum of expected multiple future rewards may serve as an important neural mechanism that is involved in the pursuit of unseen distant goals by the assignment of values to individual actions and thus the solving of temporal credit assignment problems in reinforcement learning theories (Montague et al. 1996; Sutton and Barto 1998). In this case, striatal coding may play crucial roles in the valuation of a chain of reward events that is expected over a longer time scale, such as several tens of minutes and hours (e.g., during a tennis match), several days and months (e.g., to predict stock price changes), and even years (e.g., life planning).

The neuronal signals represented in the striatum are diverse and heterogeneous in each neuron. For example, different types of value signals are represented in different subpopulations of neurons, such as “action value,” coding of the value of one specific action irrespective of choosing this action (Samejima et al. 2005), and “chosen value” (values for chosen actions) (Lau and Glimcher 2008; Pasquevau et al. 2007). Striatal neurons show selective activation to a particular action sequence (Kermadi and Joseph 1995; Miyachi et al. 2002; Ueda...
and Kimura 2003) when the actions are guided toward the reward outcomes. Subsets of neurons encode different aspects of decision signals, such as values, task strategy, choice of target, and their outcomes (Yamada et al. 2011). All of these signals appear to be derived from the selection of specific cortical inputs through learning in which dopamine signals play a central role (Balleine et al. 2007; Doya 2000). While the signals for behavioral valuation in the striatum changed dynamically through the value-based multistep choice task, the representation of long-term values was encoded persistently throughout the task periods (Fig. 5), suggesting that striatal activity might guide animals toward their goals. In support of this view, activation of the human striatum is correlated with the rapid generation of the best next move when playing the board game shogi (Wan et al. 2011). In contrast, coding of current values and task strategy occurred in striatal neurons after choices in each trial, suggesting that these neurons might be specifically involved in evaluating completed actions on a trial-by-trial basis for the next choice toward the rewards (Yamada et al. 2011). The evaluative signals in the striatum may meet for the choice and selection of actions by downstream circuits.

ACKNOWLEDGMENTS

We thank K. Samejima, T. Minamimoto, Y. Hori, K. Yamanaka, K. Louie, R. B. Rutledge, and Y. Sakai for valuable comments, and R. Sakane and H. Harada for technical assistance.

Present addresses: M. Kimura and K. Enomoto: Brain Science Institute, Tamagawa University, 6-1-1, Tamagawa-Gakuen, Machida, Tokyo 194-8610, Japan; N. Matsumoto: Department of Food and Health Sciences, Faculty of Environmental and Symbiotic Sciences, Prefectural University of Kumamoto, 3-1-100 Tsukide, Kumamoto 862-8502, Japan.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: H.Y., H.I., and M.K. conception and design of research; H.Y. and H.I. performed experiments; H.Y. analyzed data; H.Y., K.E., and M.K. interpreted results of experiments; H.Y. prepared figures; H.Y. and H.I. performed experiments; H.Y. analyzed data; H.Y., N.M., Y.U., K.E., and M.K. approved final version of manuscript.

REFERENCES


Sellem LD, Goldman-Rakic PS. Longitudinal topography and interdigitata


