Behavioral/Systems/Cognitive

Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia

Makoto Ito[1] **and Kenji Doya**[1,2]

[1]Neural Computation Unit, Okinawa Institute of Science and Technology, Okinawa 904-2234, Japan, and [2]Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute International, Kyoto 619-0288, Japan

Reinforcement learning theory plays a key role in understanding the behavioral and neural mechanisms of choice behavior in animals and humans. Especially, intermediate variables of learning models estimated from behavioral data, such as the expectation of reward for each candidate choice (action value), have been used in searches for the neural correlates of computational elements in learning and decision making. The aims of the present study are as follows: (1) to test which computational model best captures the choice learning process in animals and (2) to elucidate how action values are represented in different parts of the corticobasal ganglia circuit. We compared different behavioral learning algorithms to predict the choice sequences generated by rats during a free-choice task and analyzed associated neural activity in the nucleus accumbens (NAc) and ventral pallidum (VP). The major findings of this study were as follows: (1) modified versions of an action–value learning model captured a variety of choice strategies of rats, including win-stay–lose-switch and persevering behavior, and predicted rats' choice sequences better than the best multistep Markov model; and (2) information about action values and future actions was coded in both the NAc and VP, but was less dominant than information about trial types, selected actions, and reward outcome. The results of our model-based analysis suggest that the primary role of the NAc and VP is to monitor information important for updating choice behaviors. Information represented in the NAc and VP might contribute to a choice mechanism that is situated elsewhere.

Introduction

The theory of reinforcement learning (Sutton and Barto, 1998) plays a key role in understanding the choice behavior of animals and humans, and the function of the basal ganglia (for review, see Daw and Doya, 2006; Corrado and Doya, 2007; O'Doherty et al., 2007). According to the most basic reinforcement learning algorithm, the Q-learning model, the subject learns a reward-maximizing behavior by repeating three steps: (1) predicting the future rewards obtained by taking each candidate action, referred to as an "action value"; (2) selecting an action stochastically so that actions with higher action values are selected with higher probability; and (3) updating the action value according to the difference between the predicted and actually obtained reward. This difference is termed the "temporal difference (TD) error." The firing activity of midbrain dopamine neurons (Schultz et al., 1997) and the blood oxygen level-dependent signal from the striatum (O'Doherty et al., 2003; Tanaka et al., 2004; Daw et al., 2006; Hampton et al., 2006) both show activity patterns that are analogous to TD errors. Since the striatum is the major target of dopaminergic projection, these findings suggest that the striatal dopamine system is the most likely neural substrate of TD error-based learning (step 3) (Houk et al., 1995; Montague et al., 1996; Doya, 1999, 2000). The neural substrates of action values (step 1) and action selection (step 2), in contrast, are less clear. Recent neural recording studies in monkeys have found action value-like neuronal activity in the dorsal striatum (Samejima et al., 2005; Lau and Glimcher, 2007, 2008; Pasquereau et al., 2007). However, the findings of a large number of lesion studies in rats suggest that the nucleus accumbens (NAc) plays a major role in choice learning, especially from stochastic and delayed rewards (Cardinal and Cheung, 2005; Cardinal and Howes, 2005; Cardinal, 2006). Action value-like firing has also been found in the globus pallidus (Pasquereau et al., 2007) and other cortical areas (Platt and Glimcher, 1999; Dorris and Glimcher, 2004). One difficulty in identifying the neural loci of action valuation and selection arises from different researchers using different learning algorithms to analyze neural activity. These differing algorithms include the Q-learning model (Samejima et al., 2005), a modified version of the Q-learning model (Barraclough et al., 2004), and the local matching law (Sugrue et al., 2004).

The aims of this study are as follows: (1) to test which computational model best captures the choice learning process in animals; and (2) to elucidate how action values and action selection are realized in different parts of the corticobasal ganglia circuit. To this end, we compared different behavioral learning algorithms to predict choice sequences of rats during a free-choice task. Moreover, we analyzed associated neural activity in the NAc and a downstream structure, the ventral pallidum (VP). The major findings of this study were that (1) modified versions of the Q-learning model assuming variable parameters can capture a variety of choice strategies, including win-stay–lose-switch and

Received Dec. 25, 2008; revised May 13, 2009; accepted June 15, 2009.

Correspondence should be addressed to either Makoto Ito or Kenji Doya, Neural Computation Unit, Okinawa Institute of Science and Technology Promotion Corporation, Initial Research Project, 12-22 Suzaki, Uruma Okinawa 904-2234, Japan. E-mail: ito@oist.jp or doya@oist.jp.

DOI:10.1523/JNEUROSCI.6157-08.2009

 0270-6474/09/299861-14$15.00/0

persevering actions, and (2) there are remarkably few neurons coding action values in the NAc and VP.

Materials and Methods

Subjects

Male Long–Evans rats ($n = 6$ rats; 250–350 g body weight) were housed individually under a reversed light/dark cycle (lights on at 8:00 P.M.; off at 8:00 A.M.). Experiments were performed during the dark phase. Food was provided after training and recording sessions so that body weights dipped no lower than 90% of the initial level. Water was supplied *ad libitum*. The Okinawa Institute of Science and Technology Animal Research Committee approved the study.

Apparatus

All training and recording procedures were conducted in a 35 × 35 × 35 cm experimental chamber placed in a sound-attenuating box (Ohara). The chamber was equipped with three nose poke holes on a wall and a pellet dish on the opposite wall (Fig. 1*A*). Each nose poke hole was equipped with an infrared sensor to detect head entry, and the pellet dish was equipped with an infrared sensor to detect the presence of a sucrose pellet (25 mg) delivered by a pellet dispenser. The chamber top was open to allow connections between electrodes mounted on the rat's head and an amplifier. House lights, a video camera, and a speaker were placed above the chamber. A computer program written by LabVIEW (National Instrument) was used to control the speaker and the dispenser, and to monitor the state of the infrared sensors.

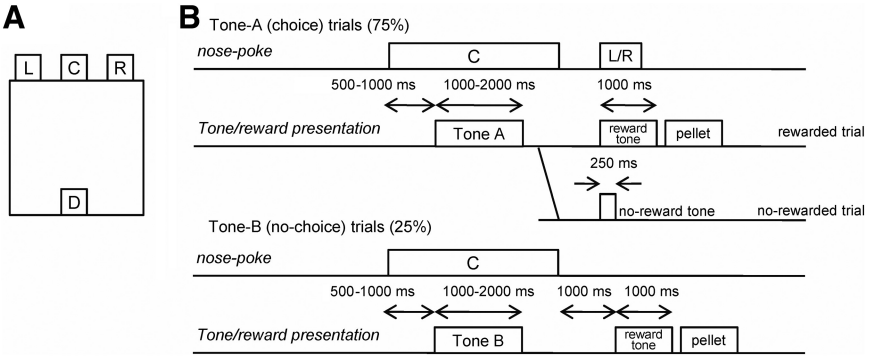

Figure 1. ***A***, Schematic illustration of the experimental chamber. The chamber was equipped with three holes for nose poking (L, left hole; C, center hole; R, right hole) and a pellet dish (D) on the opposite wall. ***B***, Schematic representation of conditional free-choice task. After a rat maintained a nose poke in the center hole for 500–1000 ms, one of two discriminative stimuli, tone A or tone B, was stochastically chosen and presented. For tone A presentation, the rat was required to perform a left or right nose poke (choice trial). After the left or right nose poke, a sucrose pellet was delivered stochastically with a certain probability depending on the rat's choice (for example, 90% reward probability for the left choice and 50% reward probability for the right choice). Reward availability was informed by different tone signals, which were presented immediately after the left or right nose poke. The reward probability in choice trials was fixed in a block, and the block was changed to the next block with a different reward probability when the average of the last 20 choices reached 80% optimal. One of four types of reward probability [(left, right), (90, 50%), (50, 90%), (50, 10%), and (10, 50%)] was used for each block. For tone B presentation, a pellet was delivered deterministically 1000 ms after the exit from the center hole (no-choice trial).

Behavioral task

The animals were trained to perform a conditional free-choice task, a combination of a tone discrimination task and a reward-based free-choice task (Fig. 1) using nose poke responses. Each trial started with a tone presentation (start tone; 2300 Hz; 1000 ms). When the rat performed a nose poke in the center hole for 500–1000 ms, a cue stimulus, tone A (4700 Hz; 1000–2000 ms) or tone B (2000 Hz; 1000–2000 ms), was presented. These tones were presented randomly but with 75% probability for tone A and 25% probability for tone B. The rat had to maintain the nose poke in the center hole during the presentation of the cue tone, or the trial was ended as an error trial after the presentation of an error tone (9500 Hz; 1000 ms).

In choice trials, after the offset of tone A, the appropriate response was for the rat to perform a nose poke in either the left or right hole within 1 min after exiting the center hole. Otherwise, the trial was ended as an error trial after the error tone. After a left or right nose poke was performed, either a reward tone (500 Hz; 1000 ms) or a no-reward tone (500 Hz; 250 ms) was presented stochastically depending on the rat's choice, according to the current left–right probability block (as described below). The reward tone was followed by the delivery of a sucrose pellet to the food dish. In no-choice trials, after the offset of tone B, the appropriate response was for rats not to perform a nose poke in either the left or right hole. After 1 s of delay after the exit of the center hole, the reward tone was presented deterministically and a sucrose pellet was delivered. Otherwise, namely, if the rat performs a nose poke in the left or right hole within 1 s after the exit of the center hole, the trial was ended as an error trial after the error tone. A successful trial ended when the rat picked the pellet on the dish or after the presentation of the no-reward tone. A new trial started after an intertrial interval (2–4 s).

The purpose of no-choice trials (tone B) was to determine the timing of the rats' decision in choice trials (tone A). With a possibility of tone B, with which a nose poke to either hole caused an error, rats had to decide to go to one of the holes only after discrimination of tone A.

The reward probabilities for left and right nose pokes in choice trials were selected from four pairs [(left, right), (90, 50%), (50, 90%), (50, 10%), and (10, 50%)]. The probability pair was fixed during a block. The same block was held until at least 20 choice trials were completed. Subsequently, the reward probability setting was changed when the choice frequency of the more advantageous side during the last 20 choice trials reached 80%. The sequence of the reward probability pairs was given in a pseudorandom order so that all four pairs were used in every four blocks, and the same pair was never given consecutively. Each rat performed at least four blocks per day.

Surgery

After rats mastered the conditional free-choice task, they were anesthetized with pentobarbital sodium (50 mg/kg, i.p.) and placed in a stereotaxic frame. The skull was exposed and holes were drilled in the skull over the recording site for anchoring screws. Two drivable electrode bundles were implanted into NAc in the right hemisphere [anteroposterior (AP), +1.7 mm; mediolateral (ML), 0.8 mm and 1.7 mm relative to the bregma; dorsoventral (DV), −6.2 mm relative to a flat skull surface], and one (two rats) or two (four rats) bundles were implanted into the VP in the right hemisphere (AP, −0.40 mm; ML, 2.5 mm, or 2.1 and 2.6 mm relative to the bregma; DV, −6.9 mm relative to a flat skull surface). The electrode bundle was composed of eight Formvar-insulated, 25 μm bare diameter nichrome wires (A-M Systems), and was inserted into a stainless-steel guide cannula (0.3 mm outer diameter; Unique Medical). The tips of the microwires were cut with sharp surgical scissors so that ~1.5 mm of the tips protruded from the cannula. Each tip was electroplated with gold to obtain an impedance of 100–200 kΩ at 1 kHz. The electrode bundles were advanced by 125 μm per recording session to acquire activity from new neurons.

Electrophysiological recording

Recordings were performed while the rats performed the conditional free-choice task. Neuronal signals were passed through a head amplifier at the head stage and then fed into the main amplifier through a shielded cable. Signals passed through a bandpass filter (50–3000 Hz) and were led to a data acquisition system (Power1401; CED), by which all waveforms that exceeded an amplitude threshold were time-stamped and saved at a sampling rate of 20 kHz. The amplitude threshold for each channel was adjusted so that action potential-like waveforms were not missed while minimizing triggers by noise. After the recording session, the following off-line spike sorting was performed using Spike2 (Spike2; CED): the recorded waveforms were classified into several groups based on their shapes and a template

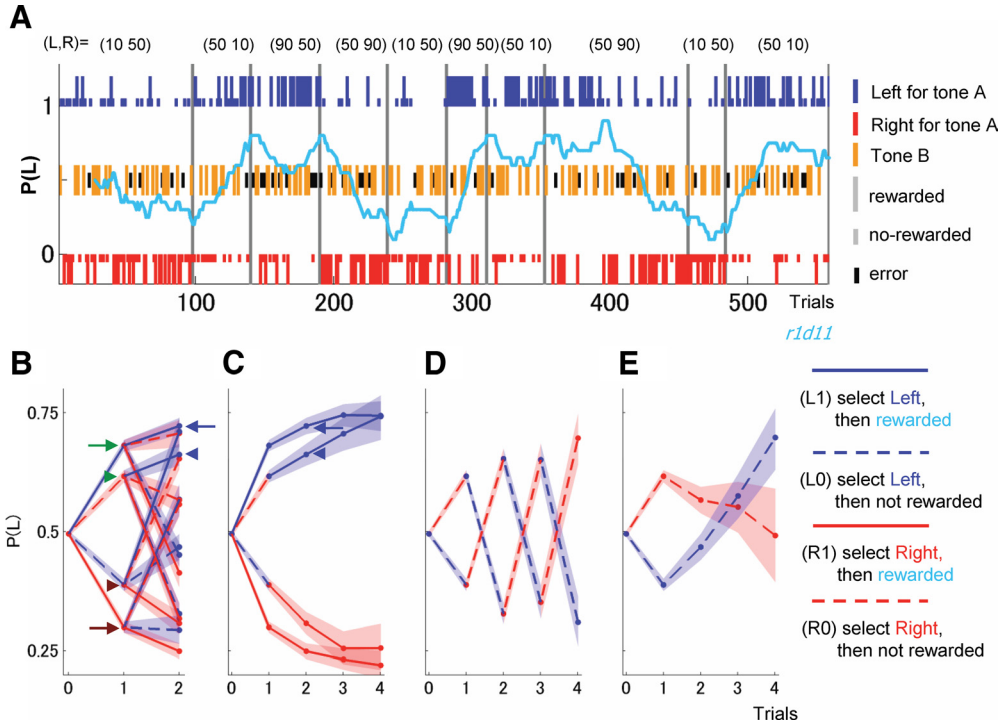

Figure 2. ***A***, Representative example of a rat's performance during one session of the conditional free-choice task. The blue and red vertical lines indicate individual choices in choice trials. The orange and black vertical lines indicate no-choice trials and error trials, respectively. The long lines and short lines represent rewarded and no-reward trials, respectively. The light blue trace in the middle indicates the probability of a left choice in choice trials (average of the last 20 choice trials). ***B–E***, The rat's strategy in choice trials, represented by left choice probabilities after different experiences with 99% confidence intervals (shaded bands). ***B***, The left choice probability for all possible experiences in one and two previous trials. Four types of experiences in one trial [left or right times rewarded (1) or no reward (0)] are represented by different colors and types of line. For instance, left probability after R0 is indicated by the right edge of a red broken line (the green arrowhead), and left probability after R0 L1 (R0 and then L1) is indicated by the right edge of a blue solid line connecting to the red broken line (blue arrowhead). ***C***, Left choice probabilities for frequently occurring sequences of four experiences. These patterns indicate rewarded experiences that gradually reinforce the selected action. A blue arrowhead and a blue arrow represent the same data indicated by the blue arrow and the arrowhead in ***B***. ***D***, Left choice probabilities for sequences of four no-reward experiences. No-reward experiences tended to switch the rat's choices. ***E***, Left choice probabilities for persevering behavior. An increase in the probability of a selected action after a no-reward outcome suggests that rats tend to continue selecting the same choice regardless of a no-reward outcome.

waveform for each group was computed by averaging. The groups of waveforms that generated templates that appeared to be action potentials were accepted, and others were discarded. We tested whether the accepted waveforms were recorded from multiple neurons or single neurons using principal component analysis.

Histology

After all experiments were completed, the rats were subjected to the same anesthetization described in the surgery section, and a 10 μA positive current was passed for 30 s through one or two recording electrodes of each bundle to mark the final recording positions. Rats were perfused with 10% formalin containing 3% potassium hexacyanoferrate (II), and the brain was carefully removed so that the microwires would not cause tissue damage. Sections were cut at 60 μm on an electrofreeze microtome and stained with cresyl violet. The final positions of the bundles of electrodes were confirmed using dots of Prussian blue. The position of each recorded neuron was estimated from the final position and the moved distance of the bundle of electrodes. If the position was outside the NAc or VP, the recording data were discarded. Tracks of accepted electrode bundles are shown in Figure 6.

Model-free behavioral analysis

In the behavioral analysis, no-choice trials and error trials were removed and the remaining sequences of choice trials were used. We denote the action in the tth choice trial as $a(t) \in \{L,R\}$, the reward as $r(t) \in \{0,1\}$, and the experience as $e(t) = (a(t), r(t)) \in \{L1,L0,R1,R0\}$. The conditional probability of making a left choice given the preceding sequence of experiences is estimated by the following:

$$\hat{P}(a(t) = L | e(t-1), \ldots, e(t-d)) = \frac{N_L(e(t-1), \ldots, e(t-d))}{N_L(e(t-1), \ldots, e(t-d)) + N_R(e(t-1), \ldots, e(t-d))}, \quad (1)$$

where $N_L(e(t-1), \ldots, e(t-d))$ and $N_R(e(t-1), \ldots, e(t-d))$ are the numbers of occurrence of the left (L) and right (R) actions, respectively, after $e(t-d), \ldots, e(t-1)$ during a block of choice trials. The 99% confidence interval of the estimation was given by Bayes' inference with a constant prior. Assuming that $N_L(e(t-1), \ldots, e(t-d))$ and $N_R(e(t-1), \ldots, e(t-d))$ were sampled from a binomial distribution with no prior knowledge about $\hat{P}_L$ (constant prior), the posterior distribution of $\hat{P}_L$ is a β distribution with parameters $N_L(e(t-1), \ldots, e(t-d))$ and $N_R(e(t-1), \ldots, e(t-d))$. The 99% confidence interval of $\hat{P}_L$ was calculated by a function in the MATLAB Statistics ToolBox, "betainv([0.005 0.995], $N_L(e(t-1), \ldots, e(t-d))$,$N_R(e(t-1), \ldots, e(t-d))$)."

Behavioral models

We considered five different models of action choice: (1) the Markov model, (2) the local matching law (Sugrue et al., 2004), (3) the standard Q-learning model (Watkins and Dayan, 1992; Sutton and Barto, 1998), (4) a forgetting Q-learning model (Barraclough et al., 2004), and (5) a differential forgetting Q-learning model.

Markov model. A dth-order Markov model is a purely descriptive model of the sequence of action choice and reward outcome. Given past experiences $e(1{:}t-1)$, which is a shorthand notation for $e(1),e(2), \ldots, e(t-1)$, the prediction of the dth-order Markov model was given by the following:

$$P_L(t) = \frac{N_L(e(t-d{:}t-1)) + 1}{N_L(e(t-d{:}t-1)) + N_R(e(t-d{:}t-1)) + 2}, \quad (2)$$

where $N_a(e(t-d{:}\,t-1))$ is the number of a (L or R) after $e(t-d{:}\,t-1)$ in the training data. The reason for adding one and two to the numerator and the denominator, respectively, was to define the prediction even if the pattern of experiences did not appear in the training data. This definition of the prediction corresponds to maximum a posteriori probability estimation for the posterior distribution of the left choice probability based on Bayes' inference with a β distribution $B(\alpha = 2, \beta = 2)$ prior, which has a peak at 0.5.

A dth-order Markov model has 4^d free parameters because there are four types of possible experiences in a single trial. The Markov model can represent any sequential dynamics with a large d. Therefore, high-order Markov models were expected to provide a good approximation of the upper bound of the prediction accuracy if the amount of training data was sufficient.

Q-learning models. To model rats' choice learning processes, we took an extension of the Q-learning model, a standard reinforcement-learning algorithm (Watkins and Dayan, 1992; Sutton and Barto, 1998) as follows. The action value $Q_i(t)$, which is the estimate of the reward from taking an action $i \in \{L,R\}$, is updated by the following:

$$Q_i(t) = \begin{cases} (1-\alpha_1)Q_i(t-1) + \alpha_1\kappa_1 & \text{if} \quad a(t-1) = i,\ r(t-1) = 1 \\ (1-\alpha_1)Q_i(t-1) - \alpha_1\kappa_2 & \text{if} \quad a(t-1) = i,\ r(t-1) = 0, \\ (1-\alpha_2)Q_i(t-1) & \text{if} \quad a(t-1) \neq i,\ r(t-1) = 1 \\ (1-\alpha_2)Q_i(t-1) & \text{if} \quad a(t-1) \neq i,\ r(t-1) = 0 \end{cases} \quad (3)$$

where $a(t)$ and $r(t)$ are the action and reward at the tth trial. The parameter α_1 is the learning rate for the selected action, α_2 is the forgetting rate for the action not chosen, κ_1 represents the strength of reinforcement by reward, and κ_2 represents the strength of the aversion resulting from the no-reward outcome. This can be reduced to standard Q-learning by setting $\alpha_2 = 0$ (no forgetting for actions not chosen) and $\kappa_2 = 0$ (no aversion from a lack of reward). This rule can be made equivalent to the version of Q-learning with forgetting of the values of actions not chosen (Barraclough et al., 2004) by setting $\alpha_1 = \alpha_2$. For convenience, in this paper, we refer to the three-parameter model with $\alpha_1 = \alpha_2$ as "Q-learning with forgetting (F-Q-learning)," and the full four-parameter model as "Q-learning with differential forgetting (DF-Q-learning)."

Using the action values, the prediction of the choice at trial t was given by the following:

$$P(a(t) = L) = \frac{1}{1 + \exp\{-(Q_L(t) - Q_R(t))\}}. \quad (4)$$

Using the choice probability of the ratio of action values

$$P(a(t) = L) = \frac{Q_L(t)}{Q_L(t) + Q_R(t)} \quad (5)$$

and setting the parameters to $\alpha_1 = \alpha_2$, $\kappa_1 = 1$, and $\kappa_2 = 0$, the model becomes equivalent to the local matching law (Sugrue et al., 2004).

F-Q- and DF-Q-learning models can represent "win-stay–lose-switch" behavior with large positive κ_1 and κ_2. Furthermore, they can represent "persevering" behavior when there is a large positive κ_1 and a large negative κ_2. A negative value of κ_2 means that even a no-food outcome has a reinforcing effect and the rat is regarded as maximizing the effective reward defined by the weighted sum of food and no-food outcomes by κ_1 and κ_2. Forgetting of the action value for the action not chosen by nonzero setting of α_2 can be regarded as a regularization term based on prior knowledge of possible changes in the reward setting.

Parameter fitting and model evaluation. To fit the parameters to the rats' choice data and evaluate the models, we used the likelihood criterion, the probability that the observed data (a sequence of selected actions) was sampled from the model. For a single trial, the likelihood $z_s(t)$ for the model prediction $P_L(t) = P(a(t) = L \mid e(1{:}\,t-1))$ is given by the following:

$$z_s(t) = \begin{cases} P_L(t) & \text{if} \quad a(t) = L\,. \\ 1 - P_L(t) & \text{if} \quad a(t) = R \end{cases} \quad (6)$$

For the data from N sessions with T_s trials in the session s, the normalized likelihood is given by the following:

$$Z = \left[\prod_{s=1}^{N} \left[\prod_{t=1}^{T_s} z_s(t) \right]^{-T_s} \right]^{-N}. \quad (7)$$

The normalized likelihood takes a maximum value of 1 when all predictions are deterministic [$P_L(t) = 0$ or 1 for all t] and all of the rat's choices are exactly matched, and takes the value of 0.5 when predictions are made with chance-level accuracy [$P_L(t) = 0.5$ for all t]. If the animal's choice is probabilistic with $P_L(t) = p$, then the normalized likelihood of the best model that exactly predicts the probability is limited by $Z_{\max} = p^p (1-p)^{(1-p)}$, which is lower than p for $0.5 < p < 1$, because of choices of lower probability actions. This causes an apparently low normalized likelihood; for example, $Z_{\max} = 0.606$, for an animal's probabilistic choice, with $p = 0.8$ and $Z_{\max} = 0.543$ for $p = 0.7$.

A total of 70 sessions of behavioral data recording, consisting of sequences of selected actions and reward outcomes in choice trials, was divided into training (35 sessions) and test data (35 sessions). Free parameters of each model were determined so that the (normalized) likelihood of the training data was maximized. The normalized likelihood of the test data was then calculated with the determined parameters. This was regarded as the prediction accuracy of the model. Evaluation of the models using prediction accuracy implicitly takes into account the penalty of the number of free parameters.

For the Q-learning models, we considered the cases of fixed parameters and time-varying parameters. Table 2 summarizes the parameters and variables of the models used in this study.

For the fixed parameter models, a set of parameters α_1, α_2, κ_1, and κ_2 were assumed to be constant for all the sessions in the test data. For the experience data $e(1{:}T)$ of each session, the sequence action values $Q_i(1{:}T)$ were computed by the update Equation 3 with the initial values of $Q_i(1) = 0$; the sequence of choice probability $P(a(t) = L)$ was obtained by Equation 4 using $Q_i(t)$. The set of parameters that maximized the normalized likelihood of the training data was used for evaluating the test data.

For the time-varying parameter models, parameters α_1, α_2, κ_1, and κ_2 were assumed to vary according to the following:

$$\alpha_j(t+1) = \alpha_j(t) + \zeta_j \quad \text{for} \quad j \in \{1, 2\}$$
$$\kappa_j(t+1) = \kappa_j(t) + \xi_j \quad \text{for} \quad j \in \{1, 2\}, \quad (8)$$

where ζ_j and ξ_j are noise terms, respectively drawn independently from the Gaussian distributions $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\kappa^2)$. The predictive distribution $P(h(t) \mid e(1{:}\,t-1))$ of parameters $h = [Q_L, Q_R, \alpha_1, \alpha_2, \kappa_1, \kappa_2]$ given past experiences $e(1{:}\,t-1)$ was estimated using the particle filter (see supplemental Methods, available at www.jneurosci.org as supplemental material) (Samejima et al., 2004). In this estimation, the initial distributions of $Q_L(1)$, $Q_R(1)$, $\alpha_1(1)$, and $\alpha_2(1)$ were uniform in [0,1] and the initial distributions of $\kappa_1(1)$ and $\kappa_2(1)$ were uniform in [0,4].

Table 1. Summary of the free parameters determined by a least-square fitting method

	No. of parameters	α_1	α_2	κ_1	κ_2	Q_0	Averaged residual error
Standard Q (const)	3	0.37	0.00[a]	2.37	0.00[a]	1.00	0.0073
F-Q (const)	4	0.96		0.99	0.64	0.00	0.0010
DF-Q (const)	5	0.52	0.88	1.18	1.44	0.71	0.0002
Local matching law	2	0.05		1.00[a]	0[a]	0.07	0.0102

const, Constant.

[a]Indicates fixed parameters.

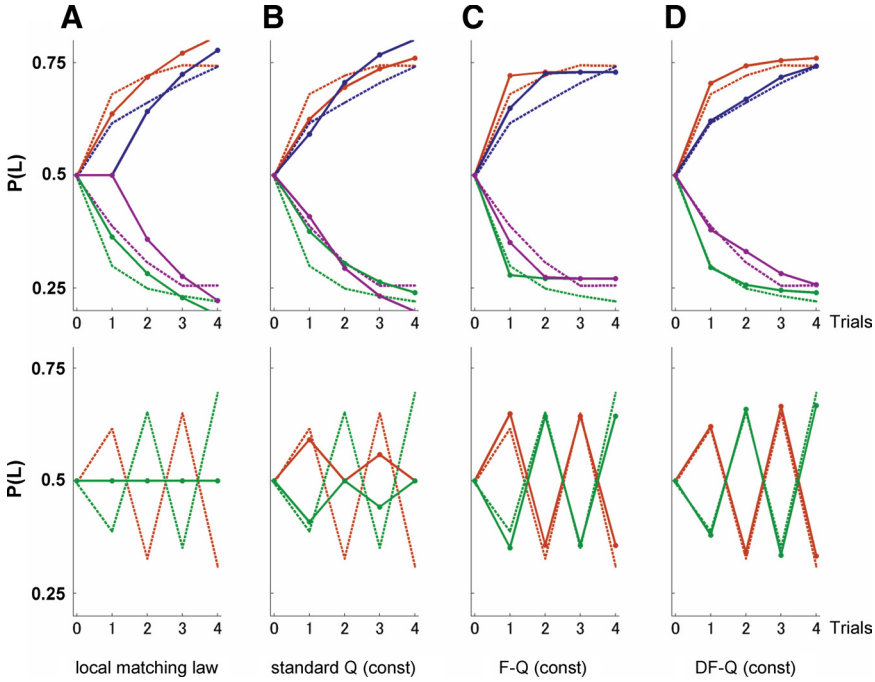

Figure 3. Model fitting to the rats' strategy by a least-square method. The choice probabilities predicted by the local matching law (***A***), the standard Q-learning model (***B***), F-Q-learning model (***C***), and DF-Q-learning model (***D***) are shown for repeated rewarded choices, the sequences after one unrewarded choice (top panel) and repeated unrewarded choices (bottom panel). The broken lines indicate the choice probabilities of the rats (same as in Fig. 2*C*,*D*), and the solid lines indicate the choice probabilities predicted by each model from the choice and reward sequences of the corresponding color. In the lower panel of ***A***, the green solid line completely overlaps with the orange solid line. The free parameters of each model were determined so that squared errors of choice probabilities between the model and rats was minimized (Table 1). The numbers of free parameters including the initial action value [$Q_0 = Q_L(1) = Q_R(1)$] in local matching law, the standard Q, F-Q, and DF-Q are 2, 3, 4, and 5, respectively.

The prediction of the model at trial t was defined by the mean value of $P_L(t)$ with respect to a joint distribution of $Q_L(t)$ and $Q_R(t)$, so that

$$P_L(t) = E[1/(1 + \exp\{-(Q_L(t) - Q_R(t))\})]. \quad (9)$$

Free parameters of Q-learning models with a time-varying parameter were σ_α and σ_κ, and were determined by maximizing the likelihood of the training data.

Neural analysis

Striatal neurons have often been classified into phasically active neurons (PANs) and tonically active neurons (TANs) in recording studies, particularly experiments performed in monkeys. We attempted to classify our NAc neurons into PANs and TANs, but could not find clear criteria for classification of firing properties. Thus, the following analyses were performed without discriminating between PANs and TANs.

Comparison of perievent time histograms. To test whether neuronal activity in the NAc and VP was modulated by differences in discriminative tone, selected action, reward availability, and/or reward probability, a Mann–Whitney U test at the 1% significance level was performed for the 1 s time bin defined in Figure 8*A*. In testing for the modulation of the discriminative tone, the number of spikes during the 1 s after the onset of the discriminative tone was compared between choice trials and no-choice trials (i.e., state-coding neurons). In testing for the modulation of action selection, neuronal activity was examined in two time bins: during the 1 s before the exit from the center hole [delay period before the choice (i.e., action command-coding neurons)], and during the 1 s after the exit from the center hole [choice period (action-coding neuron)]. For these time bins, the number of spikes was compared between left-selected and right-selected choice trials. In testing for the modulation of reward, the number of spikes during the 1 s after the tone presentation, informing of the reward availability, was compared between rewarded choice trials and no-reward choice trials (i.e., reward-coding neurons).

In testing for reward probability, neuronal activity in only the last 20 choice trials in each block was tested, when the rats were assumed to have learned the reward probability. The left action value-coding neurons (Q_L-coding neurons) were defined as the neurons in which spike counts were significantly different between blocks with a reward probability of left, right = (90, 50%) and blocks with a probability of (10, 50%), but not significantly different between the probabilities (50, 90%) and (50, 10%). In the same way, the right action value-coding neurons (Q_R-coding neurons) were defined as the neurons in which spike counts were significantly different between blocks with probabilities of (50, 90%) and (50, 10%), but not between blocks with probabilities of (90, 50%) and (10, 50%).

Regression analysis. We used multiple linear regression analysis to test neuronal correlations with action values, and other reward-related values, including the sum of the action values (i.e., "state value"), and the difference of the action values (i.e., "policy").

To detect neurons coding the action value, the neuronal activity in only the last 20 choice trials in each block was applied using the following regression model:

$$y(t) = \beta_0 + \beta_1 Q_L(t) + \beta_2 Q_R(t), \quad (10)$$

where $y(t)$ was the number of spikes at trial t in the 1 s bins defined in Figure 8*A*: before nose poking at the center hole (phase 1), after the onset of the cue tone (phase 2), and before initiation of action (phase 3). β_i is the regression coefficient, and $Q_L(t)$ and $Q_R(t)$ are the reward probabilities assigned for a left or right choice, respectively, in each trial (i.e., a reinforcement learning model was not used). Left action value-coding neurons and right action value-coding neurons (Q_L- and Q_R-coding neurons) were defined as neurons that had significant regression coefficients to Q_L but not to Q_R, and to Q_R but not to Q_L, respectively (t test, $p < 0.01$).

To detect neurons coding the sum of the action values (state value-coding neurons) and the difference between the action values (policy-coding neurons), an alternative regression model

$$y(t) = \beta_0 + \beta_1[Q_L(t) + Q_R(t)] + \beta_2[Q_L(t) - Q_R(t)] \quad (11)$$

was used (Seo and Lee, 2007). State value-coding neurons and policy-coding neurons were defined as neurons that, respectively, exhibited a significant regression coefficient to $[Q_L + Q_R]$ but not to $[Q_L - Q_R]$, and to $[Q_L - Q_R]$ but not to $[Q_L + Q_R]$ (t test, $p < 0.01$).

Information analysis. To elucidate when and how much information of each event [such as state (discriminative tone), action, reward, and action values] was coded in the NAc and VP, the mutual information shared between firing and each event was calculated using a sliding time window (duration 500 ms). A measure of "mutual information" quantifies the mutual dependence of two variables.

For instance, to evaluate the mutual information shared between the firing of a neuron and selected actions, we used data from the choice trials in the session in which the activity of the neuron was recorded. The number of spikes in a certain time window in a trial was defined as a random variable F, taking the value f_i, which represented a class of the number of spikes n, where i is the index of the class satisfying $N_i \leq n < N_{i+1} (i = 1 \ldots N_F)$. N_F is the number of the classes, and we chose $N_F = 4$ in this analysis. N_i is the border between the classes, which was determined for each neuron so that the variance of the members in the classes was minimized, namely,

$$\{N_1^*, \ldots, N_5^*\} = \arg\min_{\{N_1, \ldots, N_5\}} \sum_{i=1}^{5} (L_i - \langle L \rangle)^2, \quad (12)$$

where L_i is the number of f_j, $\langle L \rangle$ is the average of L_i over i, and $\{N_1^*, \ldots, N_5^*\}$ are the optimized $\{N_1, \ldots, N_5\}$, which were used for calculating the mutual information. Practically, $\{N_1^*, \ldots, N_5^*\}$ were obtained by testing all possible combinations of the values. The reason for this binning was to reduce the bias of estimated mutual information. The action was defined as a random variable X taking the value x_j ($j = 1 \ldots N_X$), where N_X was 2, and x_1 and x_2 correspond to left and right, respectively.

The mutual information between F and X was defined by the following:

$$I(F, X) = \sum_{i=1}^{N_F} \sum_{j=1}^{N_X} p(f_i, x_j) \log \frac{p(f_i, x_j)}{p(f_i)p(x_j)}, \quad (13)$$

where $p(f_i, x_j)$ was the joint probability distribution function of F and X, and $p(f_i)$ and $p(x_i)$ were the marginal probability distributions of F and X, respectively. Although we could not determine these probability distributions exactly, the mutual information could be approximated by the following:

$$\hat{I}(F, X) = \sum_{i=1}^{N_F} \sum_{j=1}^{N_X} \frac{M_{ij}}{M} \log \frac{M_{ij}M}{M_{i\cdot}M_{\cdot j}} - \frac{1}{2M \log 2}(U_{FX} - U_F - U_X + 1), \quad (14)$$

where M_{ij} is the number of the pairs of f_i and x_j in the session. $M_{\cdot j}$ and $M_{i\cdot}$ are the sums over i and j of M_{ij}, respectively. M is the sum over both i and j, corresponding to the total number of trials. U_{FX} is the number of nonzero M_{ij} for all i and j, U_F is the number of nonzero $M_{\cdot j}$ for all i, and U_X is the number of nonzero $M_{\cdot j}$ for all j. The first term represents the direct approximation of $I(F, X)$. This term is, however, a biased estimator. To correct this bias, the second term, the first-order approximation of the bias, was subtracted from the direct approximation (Panzeri and Treves, 1996).

To calculate the mutual information between neuronal firing (f_i) and action (x_1 = left; x_2 = right) or reward (x_1 = rewarded; x_2 = no reward), we used the data from choice trials in the session in which the activity of the neuron was recorded. For calculating the mutual information between neuronal firing and the discriminative tone (x_1 = tone A; x_2 = tone B), data from both choice and no-choice trials were used. For calculating the mutual information of action values, first, Q_L and Q_R in choice trials estimated by an F-Q-learning model with a time-varying parameter were binarized: $Q_i(t)$ values larger than the median of Q_i were set to x_1 and the rest were set to x_2. The binarized action values were then used to calculate mutual information.

Mutual information per second for each neuron was calculated using a sliding time window (duration, 500 ms; step size, 100 ms), which was set based on the onset time of tone A or B, or the exit time of the center hole. Finally, the averages of the mutual information over all neurons in the NAc and VP were obtained.

To test whether the obtained mutual information was significant, the threshold indicating significant information ($p < 0.01$) was obtained in the following way: a binary event, x_1 or x_2, was generated randomly for each trial, and the averaged mutual information between this random

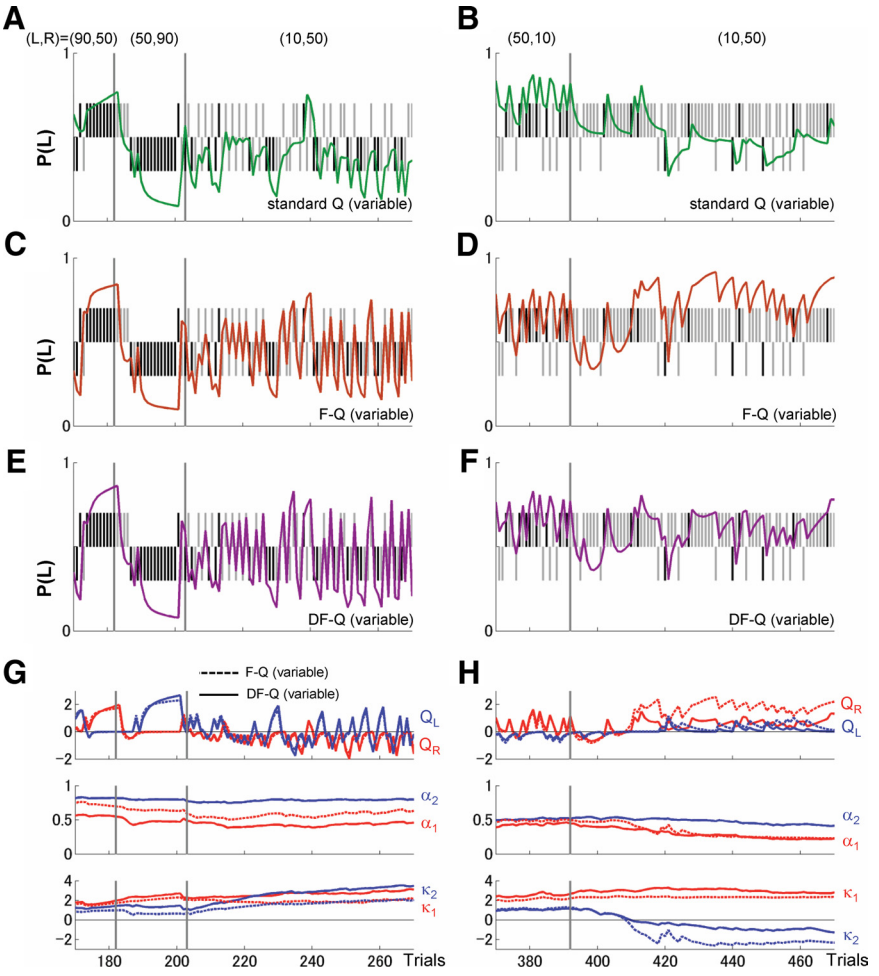

Figure 4. Example of trial-by-trial predictions of rats' choices based on reinforcement learning algorithms. ***A–F***, Representative examples of trial-by-trial predictions using the standard Q-learning model (***A***, ***B***), F-Q-learning model (***C***, ***D***), and DF-Q-learning model (***E***, ***F***). In all models, parameters were assumed to be variable (see Materials and Methods). In the panels of the left side and the right sides, different choice data were applied. The probability that a rat would select left at trial t was estimated from the rat's past experiences $e(1), \ldots e(t-1)$ and plotted at trial t. The actual rat's choice at each trial is represented by a vertical line. The top lines and bottom lines indicate left and right choices, respectively. The black and gray colors indicate rewarded and no-rewarded trials, respectively. ***G***, ***H***, Estimated model parameters of F-Q- (broken lines) and DF-Q-learning model (solid lines) during the predictions.

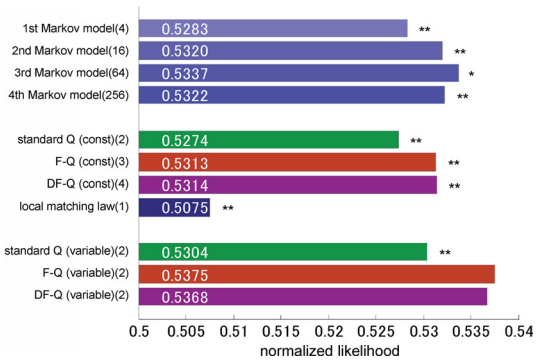

Figure 5. Accuracy of each model in trial-by-trial prediction of rats' choice. The prediction accuracy was defined by the normalized likelihood of test data. The free parameters of each model were determined by the maximization of the likelihood of training data. Numbers followed by the name of models indicate the numbers of free parameters of each model. "const" means that the parameters of the model, such as the learning rate, was assumed to be constant for all sessions, and "variable" means that the parameters were assumed to be variable. The double and single asterisks indicate a significant difference from the prediction accuracy of F-Q-learning model (variable); $p < 0.01$ and $p < 0.05$ in paired-sample Wilcoxon's signed rank tests, respectively.

Table 2. Summary of the free parameters used in each model to predict rats' choices

	No. of parameters	α_1	α_2	κ_1	κ_2	σ_α	σ_κ
Standard Q (const)	2	0.45	0.00[a]	1.25	0.00[a]		
F-Q (const)	3	0.55		1.25	0.55		
DF-Q (const)	4	0.60	0.45	1.20	0.45		
Local matching law	1	0.01		1.00[a]	0[a]		
Standard Q (variable)	2					0.005	0.10
F-Q (variable)	2					0.005	0.07
DF-Q (variable)	2					0.010	0.07

The parameters were determined so that the likelihood of the training data was maximized. const, Constant.
[a]Indicates fixed parameters.

Table 3. Summary of the average ± SD of the parameters estimated by the Q-learning models across all data (both the training data and test data)

	α_1	α_2	κ_1	κ_2
Standard Q (variable)	0.45 ± 0.23	0.00[a]	1.86 ± 0.68	0.00[a]
F-Q (variable)	0.47 ± 0.25		1.61 ± 0.77	1.10 ± 0.87
DF-Q (variable)	0.45 ± 0.19	0.46 ± 0.19	1.69 ± 0.77	1.24 ± 0.92

[a]Indicates fixed parameters.

event and spikes were obtained for the NAc and VP using the same method described above. This calculation was repeated 100 times with new random events. The second largest mutual information for each time window was then regarded as the threshold indicating whether or not information was significant.

Results

Behavioral data

In the conditional free-choice task, a rat chose an action of nose poking through the left hole [$a(t) = \mathrm{L}$] or the right hole [$a(t) = \mathrm{R}$] after tone A presentation. A food reward was given stochastically depending on the choice. The reward probabilities for the left and right choices were chosen from four left–right probability pairs [(90, 50%), (50, 90%), (50, 10%), and (10, 50%)] and were fixed during a block of trials until the choice probability of the more advantageous action in the last 20 choice trials reached 80%. Figure 2*A* shows a representative example of a rat's performance during 10 consecutive blocks of trials. The choice probability of the rat changed according to the experience of rewarded and nonrewarded actions. The choice trials were interspersed with no-choice trials in 25% of the block, so as not to allow the rat to prepare an action before presentation of the choice tone. On average, 105.6 trials were required in one block, and 8.1 blocks (from 5 to 12 blocks) were performed in one session. Here, we report the results of all 60,221 trials in 70 recording sessions performed by six rats, consisting of 39,175 choice trials (65.1%), 12,707 no-choice trials (21.1%), and 8329 error trials (13.8%).

Behavioral analysis

Reaction time and tone discrimination

In the choice trials, the appropriate response was for rats to perform a nose poke to left or right hole within 1 min after exiting the center hole. In no-choice trials, the appropriate response was for rats not to perform a nose poke in either the left or right hole to obtain a food pellet 1 s after exiting the center hole.

The reaction time in the choice trials, the time period from the exit of the center hole to the entry of the left or right hole, was <1 s in 72% of the trials and had a median of 0.75 s. The errors in the choice trials, in which rats did not perform any nose pokes within 1 min for tone A presentation, were only 0.39% of the choice trials. The errors in no-choice trials, in which rats performed a nose poke at the left or right hole within 1 s for tone B presentation, were 0.73% of the no-choice trials. These data show that the rats discriminated the tones well and responded differently to each.

History dependence of action choice

In the following behavioral analysis, no-choice trials and error trials were removed and the remaining sequences of choice trials were used. We first analyzed the conditional probability of making a left choice given preceding experiences of actions and rewards (Fig. 2*B–E*). There were four possible types of experiences in each trial: L1, L0, R1, and R0, where L or R denotes left or right choice, respectively, and 1 or 0 denotes rewarded or nonrewarded trials, respectively. Both the experience L1 and R0 in the previous trial increased the probability of a left choice in the current trial [Fig. 2*B*, green arrow, $P(\mathrm{L}|\mathrm{L1}) = 0.68$; green arrowhead, $P(\mathrm{L}|\mathrm{R0}) = 0.62$]. The left choice probability after L1 was significantly higher than after R0 (Mann–Whitney *U* test, $p < 0.0001$). Equally, both the experience of R1 and L0, which are symmetrically opposite experiences of L1 and R0, decreased the probability of a left choice in the current trial (that is, increased the probability of a right choice). The left probability after R1 was significantly lower than after L0 [Fig. 2*B*, brown arrow, $P(\mathrm{L}|\mathrm{R1}) = 0.30$; brown arrowhead, $P(\mathrm{L}|\mathrm{L0}) = 0.39$] (Mann–Whitney *U* test, $p < 0.0001$). These results indicate that the choice probability was modulated by experience in previous trials and that this modulation was stronger after rewarded experience than after nonrewarded experience.

Experiences before the previous trial were also found to affect current choices. Figure 2*B* shows the probabilities of making a left choice in the current trial after all combinations of types of experience in one and two previous trials. For instance, the conditional probability $P(\mathrm{L}|\mathrm{R0\ L1}) = 0.66$ (Fig. 2*B*,*C*, blue arrowhead), which is the probability of a left choice (at trial t) after R0 (at trial $t - 2$) and L1 (at trial $t - 1$), significantly differs from $P(\mathrm{L}|\mathrm{L1\ L1}) = 0.72$ (Fig. 2*B*,*C*, blue arrow), the probability of a left choice after L1 and then L1, even though the experience in one of the previous trials was the same in both conditions (Mann–Whitney *U* test, $p < 0.0001$). This result suggests that the different effects of the experiences R0 and L1 before the previous trial still affected the current choice. We examined the effects of the experiences of up to four previous trials (Fig. 2*C–E*). Among $4^4/2 = 128$ combinations of four consecutive experiences in which left–right sequences were symmetrical, the sequence L1 L1 L1 L1 and R1 R1 R1 R1 was observed most frequently (6.87% of all observed sequences of four trials). The sequence R0 L1 L1 L1 and L0 R1 R1 R1 was observed second most frequently (2.87%). Figure 2*C* shows the left choice probabilities after the subsequences of these cases. For instance, for the pattern L1 L1 L1 L1, the left choice probabilities, namely $P(\mathrm{L}|\mathrm{L1})$, $P(\mathrm{L}|\mathrm{L1\ L1})$, $P(\mathrm{L}|\mathrm{L1\ L1\ L1})$, and $P(\mathrm{L}|\mathrm{L1\ L1\ L1\ L1})$, increased with the number of L1 experiences and converged around 0.75. A comparison of the left choice probabilities for L1 L1 L1 L1 and R0 L1 L1 L1 results indicated that the effects of L1 and R0 were gradually reduced by following the sequence of L1. There was no significant difference between $P(\mathrm{L}|\mathrm{L1\ L1\ L1\ L1}) = 0.74$ and $P(\mathrm{L}|\mathrm{R0\ L1\ L1\ L1}) = 0.74$ (Mann–Whitney *U* test, $p = 0.94$). These features were also observed in the symmetrically opposite patterns R1 R1 R1 R1 and L0 R1 R1 R1.

The sequences R0 L0 R0 L0 and L0 R0 L0 R0 were observed third most frequently (2.55%). Figure 2*D* shows the probabilities of left choices after these subsequences. These results illustrate that there was a "switching" behavior in which rats tended to choose the other action after one action was not rewarded. The responses to certain sequences were surprising. Figure 2*E* shows

the left choice probability for the subsequences L0 L0 L0 L0 and R0 R0 R0 R0, which consisted of 1.28% of all observed sequences across the four trials. The probability of a left choice gradually increased after two or more left choices with no reward. In particular, P(L|L0 L0 L0) and P(L|L0 L0 L0 L0) were both >0.5. This indicates that rats sometimes persevered with one choice even after repeated nonrewarded trials. This behavior would be expected to be maladaptive for adapting to a new setting, but adaptive if the best choice produced only sparse rewards.

In summary, these results demonstrated the following: (1) rewarded experiences gradually reinforced the selected action, like the update rule in reinforcement learning algorithms; (2) nonrewarded experiences led rats to switch their choice; and (3) rats sometimes persevered with one choice even after repeated nonrewarding outcomes.

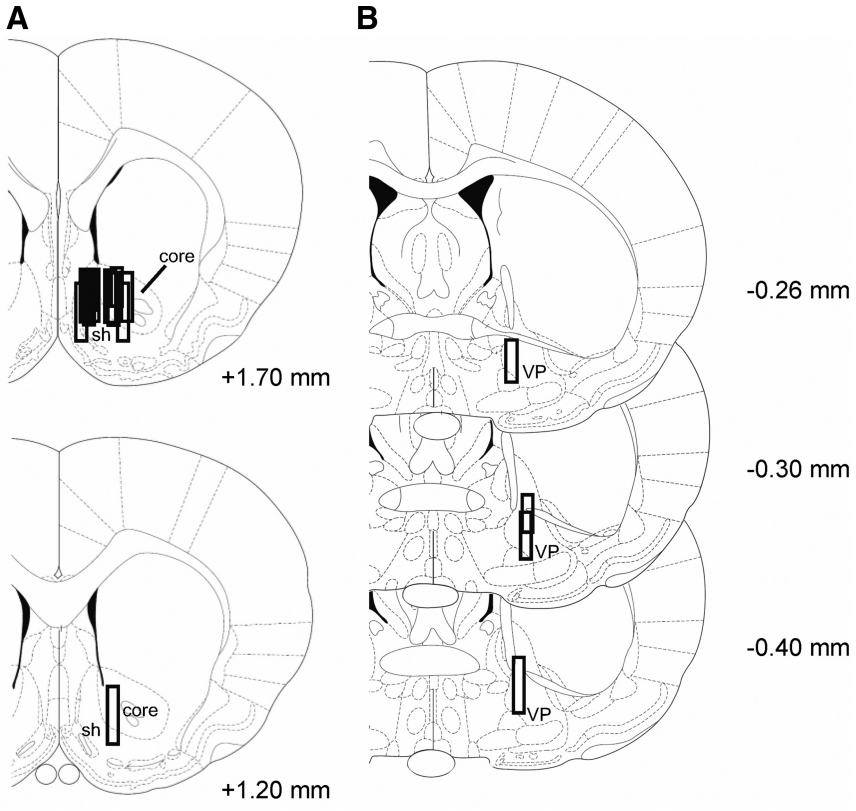

Figure 6. Tracks of accepted electrode bundles for all rats are illustrated by rectangles. Each diagram represents a coronal section referenced to the bregma (Paxinos and Watson, 1998). Data recording from the sites in ***A*** and ***B*** were treated as neuronal activity in the NAc and VP, respectively. core, Nucleus accumbens core; sh, nucleus accumbens shell; VP, ventral pallidum.

Modeling rats' strategies with a reinforcement learning model

To examine the extent to which different learning models can represent rats' strategies, four different models were fitted to the choice sequence data shown in Figure 2, *C* and *D*, by a least-squares method.

Q-learning models update two action values, $Q_L(t)$ and $Q_R(t)$, according to action and reward experiences. The probability of an action choice is given by a sigmoid function of the difference of the two action values, namely, the following:

$$P(a(t) = L|e(1), \ldots, e(t-1)) = \frac{1}{1 + \exp\{-(Q_L(t) - Q_R(t))\}}, \quad (15)$$

where $e(t)$ can take one of four values {L0, L1, R0, R1}. The most generalized model, Q-learning with differential forgetting (DF-Q), had four parameters: α_1, the learning rate for the chosen action; α_2, the forgetting rate for the actions not chosen; κ_1, the reinforcing strength of a reward; and κ_2, the aversive strength of a no-reward outcome. By fixing $\alpha_2 = \kappa_2 = 0$, this rule becomes the standard Q-learning model. By fixing $\alpha_1 = \alpha_2$, it becomes the Q-learning with forgetting (F-Q) model. By fixing $\alpha_1 = \alpha_2$, $\kappa_1 = 1$, and $\kappa_2 = 0$, and using the choice probability by the ratio of action values

$$P(a(t) = L|e(1), \ldots, e(t-1)) = \frac{Q_L(t)}{Q_L(t) + Q_R(t)}, \quad (16)$$

the DF-Q model is equivalent to the local matching law (Sugrue et al., 2004).

Choice probabilities were updated using the local matching law, and the standard Q-, F-Q-, and DF-Q-learning models, according to the six sequences of four experiences shown in Figure 2, *C* and *D*. The free parameters of each model were determined so that squared errors of choice probabilities between the model and rats was minimized (Table 1). Since the initial action value [$Q_0 = Q_L(1) = Q_R(1)$] was also treated as a free parameter, the numbers of free parameters in the local matching law, the standard Q, F-Q, and DF-Q were 2, 3, 4, and 5, respectively. Figure 3 shows the predicted action choice probabilities of the four models after optimization of the free parameters. The results showed that the local matching law (Fig. 3*A*) and the standard Q-learning model (Fig. 3*B*) were not able to represent switching behavior, whereas the F-Q- and DF-Q-learning models were (Fig. 3*C*,*D*). The averaged residual errors of the local matching law, the standard Q, F-Q, and DF-Q models were 0.0102, 0.0073, 0.0010, and 0.0002, respectively.

Although the characteristics of each model were clarified by fitting these models, this comparison may not be equivalent because the number of free parameters was not considered. Generally, a model with a larger number of free parameters shows a better fit to data, although this sometimes results in overfitting, which suggests that the model can no longer fit to new data. Therefore, a strict comparison is necessary to measure the accuracy of the models in predicting new data.

Prediction of choices by models

We then examined which computational model most accurately captured the rats' choice behaviors on the basis of the accuracy of prediction of rats' choices. This examination compared the Markov model, the standard Q-learning model, the F-Q-learning model, and the DF-Q-learning model with constant or time-varying parameters.

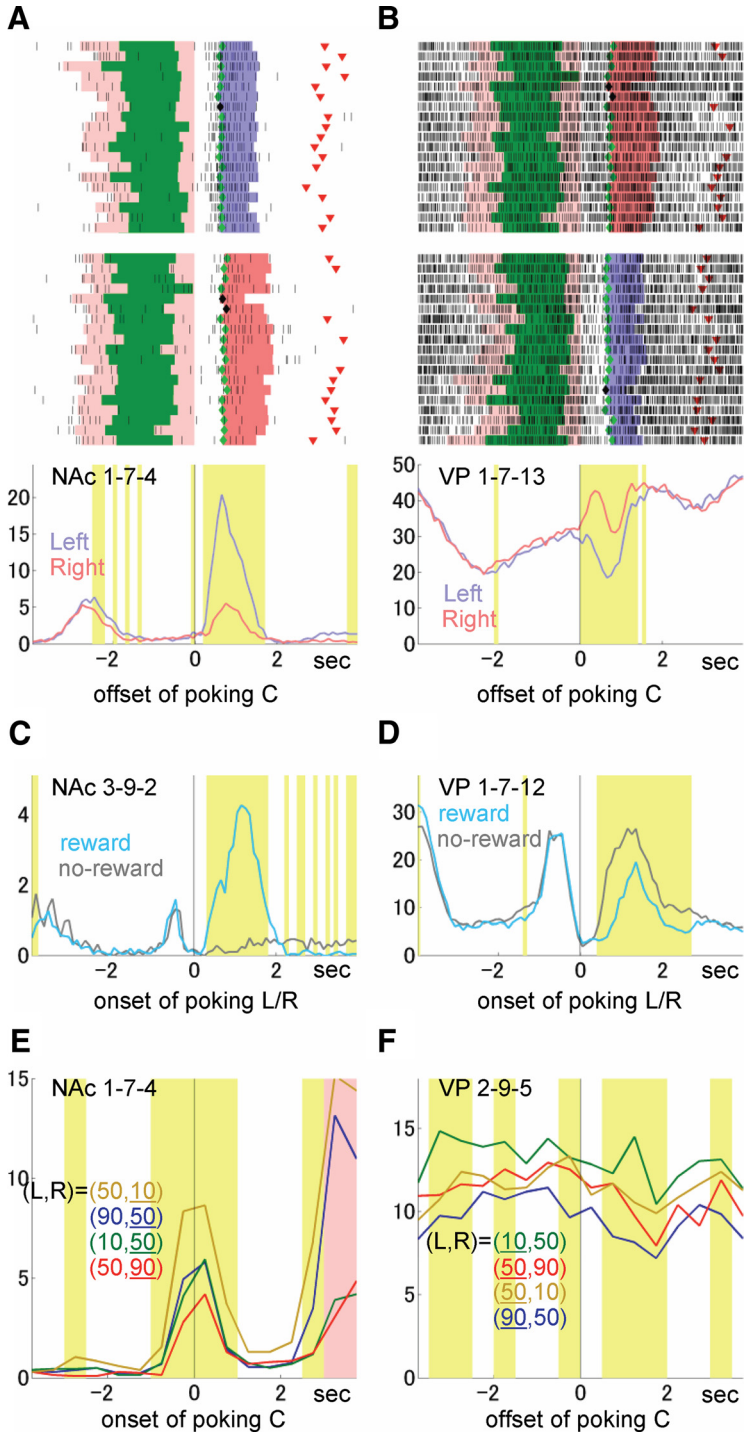

Figure 7. Examples of neuronal activity in the NAc (***A***, ***C***, ***E***) and VP (***B***, ***D***, ***F***) modulated by various task events. ***A***, ***B***, ***D***, and ***E*** were neuronal responses recorded in the same session. Of these, ***A*** and ***E*** are data from the same neuron. ***A***, ***B***, Examples of neuronal activity modulated by the selected action (action-coding neurons). Top (bottom) rasters show spikes and events on choice trials in which a left (right) nose poke was selected. The perievent time histograms in the bottom panels are aligned with the exit from the center hole. ***C***, ***D***, Examples of neuronal activity modulated by the availability of reward (reward-coding neurons). The perievent time histograms are aligned with the onset of a reward tone or no-reward tone. ***E***, ***F***, Examples of neuronal activity coding reward probability for one of two actions (action value-coding neurons). The perievent time histograms for last 20 choice trials in four different blocks are shown by different colors. ***E***, The histograms were aligned with entry to the center hole. There is a significant difference in the activity between block (50, 10) and block (50, 90), but no difference between block (90, 50) and block (10, 50) around the entry (yellow bins, $p < 0.01$, from −1 to 1 s). This suggests that the activity codes the reward probability for right action. Because this neuron also coded the selected action (as shown in ***A***), the firing rates were significantly different between (90, 50) and (10, 50), and between (50, 90) and (50, 10), 3 s after the entry to the center hole (pink bins). ***F***, The histogram were aligned with the exit from the center hole. This VP neuron coded the reward probability for left action. The yellow bins indicated a significant difference in the firing rate between block (10, 50) and block (90, 50) ($p < 0.01$) and no difference between blocks (50, 90) and (50, 10). The green bands in the rasters show the time of presentation of tone A. The pink bands behind green bands represent the time periods of center nose pokes. The blue and red bands represent left and right nose pokes, respectively. The green and black diamonds indicate the onset of reward and no-reward tones, respectively. The red triangles indicate the time of a rat's picking up a sucrose pellet from the pellet dish. Each perievent time histograms were constructed for 100 ms bins (***A–D***) or 500 ms bins (***E***, ***F***). The yellow bins in the histograms show significant differences in firing rate (Mann–Whitney *U* test, $p < 0.01$).

The Markov model estimates the conditional probability of an action choice given the action and reward experiences in the last d trials, namely, the following:

$$P(a(t) = L | e(t-d), \ldots, e(t-1)). \quad (17)$$

The conditional probabilities are estimated from the occurrence frequency of action and reward sequences.

The parameters of the Q-learning models, such as α_1, α_2, κ_1, and κ_2, were assumed to be either fixed during a session, or to slowly vary with two drift-rate parameters σ_α and σ_κ. The time course of the action values and the time-varying parameters were estimated from the experimental data using the method of particle filtering (Samejima et al., 2004, 2005) (see Materials and Methods; supplemental Methods, available at www.jneurosci.org as supplemental material).

A total of 70 sessions of behavioral data was divided into training data (35 sessions; 19,986 choice trials) and test data (35 sessions; 19,186 choice trials). Free parameters of each model were determined so that the likelihood of the predictions by the models of the training data was maximized. The normalized likelihood of the predictions by the models of the test data was used as the measure of model performance (see Materials and Methods).

Figure 4*A–F* shows representative examples of trial-by-trial prediction of the standard Q (Fig. 4*A*,*B*), F-Q (Fig. 4*C*,*D*), and DF-Q (Fig. 4*E*,*F*) models with variable parameters. Estimated action values Q_L, Q_R, and the parameters of F-Q and DF-Q are shown in Figure 4, *G* and *H*. In Figure 4, *A*, *C*, and *E*, the predictions by these three models were almost the same during the (50, 90%) reward probability block, when the rats predominantly chose the right hole. However, after the reward setting was changed to (10, 50%), the rats started to adopt a win-stay–lose-switch-like strategy; selecting the same action after a rewarded trial, and switching the

action after a no-rewarded trial. This behavior was captured by the F-Q- and DF-Q-learning models, with both models showing an increase in the nonreward aversion parameter κ_2 close to the reward reinforcement parameter κ_1, which caused nearly symmetric ups and downs in the Q_L and Q_R (Fig. 4*G*). This change, in turn, caused rapid swings in choice probability (Fig. 4*C*,*E*). However, the standard Q-learning model, with fixed parameters $\alpha_2 = \kappa_2 = 0$, continued to predict higher choice probability for the right-hand action (Fig. 4*A*). Thus, the F-Q- and DF-Q-learning models predicted the switching of choice better than the standard Q-learning model (Fig. 4*A*).

Figure 4, *B*, *D*, and *F*, shows an example for persevering behavior, when the rat continued to choose the less advantageous left action in a (10, 50%) block. The F-Q and DF-Q models captured this behavior by exhibiting a negative value for the aversion parameter κ_2 (Fig. 4*H*), with which both reward and no-reward outcomes are positively reinforced. The standard Q model, however, could not predict left action choice.

These results illustrate that the F-Q- and DF-Q-learning models are able to capture the actual rats' behaviors, including win-stay–lose-switching and persevering behaviors, which the standard Q-learning model could not. Moreover, the performance of the F-Q and DF-Q models were remarkably similar.

Figure 5 shows a comparison of the prediction performance of different models using the normalized likelihood of the model for 35 sessions of test data. The free parameters of each model were given by the maximal likelihood estimate from 35 sessions of training data (listed in Table 2). The normalized likelihood of the first-order Markov model was 0.528, which can be regarded as the baseline performance for a prediction model. The normalized likelihood for the second- and third-order Markov models increased to 0.532 and 0.534, respectively. Note that, when an animal's choice is probabilistic, the normalized likelihood for even the best model becomes close to 0.5 (see Materials and Methods). The normalized likelihood of 0.534 for the third-order Markov model is equivalent to that of the optimal model for action choice probability ($p = 0.68$). The performance of the fourth-order model was 0.532, lower than that of the third-order model, presumably because of overfitting of its 256 free parameters.

Predictions by the standard Q-, F-Q-, and DF-Q-learning models with constant parameters (α and κ) were much better (0.527, 0.531, and 0.531, respectively) than the predictions of the local matching law (0.508), but poorer than that of the first- or second-order Markov models. One of the reasons for the lower prediction accuracies of the local matching law and the standard Q model with constant parameters is that neither model could represent the switching behavior that rats often exhibited (Fig. 3).

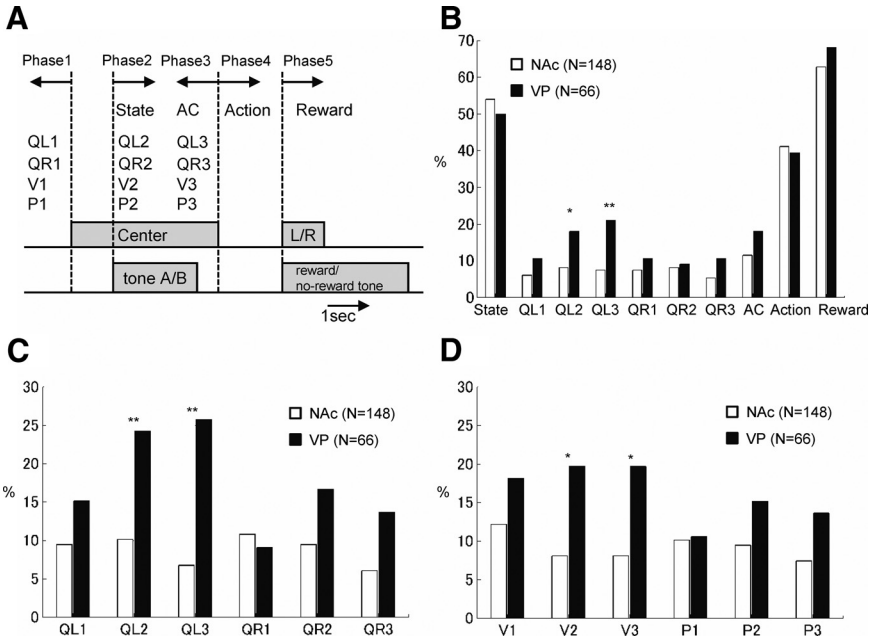

Figure 8. Information coded in the NAc and VP. ***A***, Time bin neuronal activity was examined as follows: for 1 s before the onset of the a nose poke at the center hole (phase 1), after the onset of the cue tone (phase 2), before initiation of action (phase 3), after the action onset (phase 4), after the onset of the reward or no-reward tone (phase 5). ***B***, The population of neurons that showed significant selectivity (Mann–Whitney *U* test, $p < 0.01$) for each event. State-coding neurons are defined as neurons that showed a significantly different firing rate in choice and no-choice trials for 1 s after the onset of the cue tone (phase 2). The neurons coding action values for left or right choices (Fig. 7*E*,*F*) were detected for three different time bins, phases 1–3. QLn and QRn indicate the action values for left and right during phase *n*, respectively. Note that these action value-coding neurons were detected by simple comparisons of firing rate in different blocks, not using computational models. Action command (AC)-coding neurons are defined as neurons that showed an action selectively during the 1 s before initiation of action (phase 3). Action-coding neurons are the neurons showing action selectivity during 1 s after the action onset (phase 4) (Fig. 7*A*,*B*). Reward-coding neurons are the neurons that showed different firing rate between rewarded trials and no-reward trials during 1 s after the onset of the reward or no-reward tone (phase 5) (Fig. 7*C*,*D*). ***C***, The population of neurons coding the action values detected by a linear regression analysis. The reward probabilities for left and right were used as regressors (a model-free analysis). The neurons with a significant coefficient for the reward probability for either left or right were defined as the action value-coding neurons for left and right, respectively. QLn and QRn indicate the action values for left and right during phase *n*, respectively. ***D***, The population of neurons coding the state value and the policy detected by a linear regression analysis. The sum of the reward probabilities for both actions and the difference of them were used as regressors (a model-free analysis). The neurons with a significant coefficient for ether the sum or the difference were defined as the state value and the policy-coding neurons, respectively. Vn indicates the state value during phase *n*, and Pn the policy during phase *n*. All populations were significantly larger than the chance level (binomial test, $p < 0.01$). The single and double asterisks indicate significant differences in the percentages of coding neurons between the NAc and VP; $p < 0.05$ and $p < 0.01$, respectively, in Mann–Whitney *U* test.

With time-varying parameters, the performance of the standard Q, F-Q, and DF-Q models improved (0.5304, 0.5375, and 0.5368, respectively). The F-Q-learning model with time-varying parameters produced the best prediction performance. The averages and SDs of the estimated parameters are summarized in Table 3. The values of the normalized likelihood of the models may appear low, but this is an expected consequence if rats themselves make stochastic action choices (see Materials and Methods). If we take binary predictions from the choice probability [left for $P_L(t) > 0.5$, right for $P_L(t) < 0.5$], the average percentage of correct predictions over 35 test sessions is 68%. Furthermore, the meaningful performances of the standard Q, F-Q, and DF-Q models were also conformed by plotting the actual left-choice probability for the difference between Q_L and Q_R (supplemental Data 1, available at www.jneurosci.org as supplemental material).

A statistical test for the likelihood of each session in the test data showed that the prediction accuracies of both the F-Q- and DF-Q-learning models with time-varying parameters were significantly higher than those of all other models (paired-sample Wilcoxon's signed rank test, $p < 0.05$ for third Markov model,

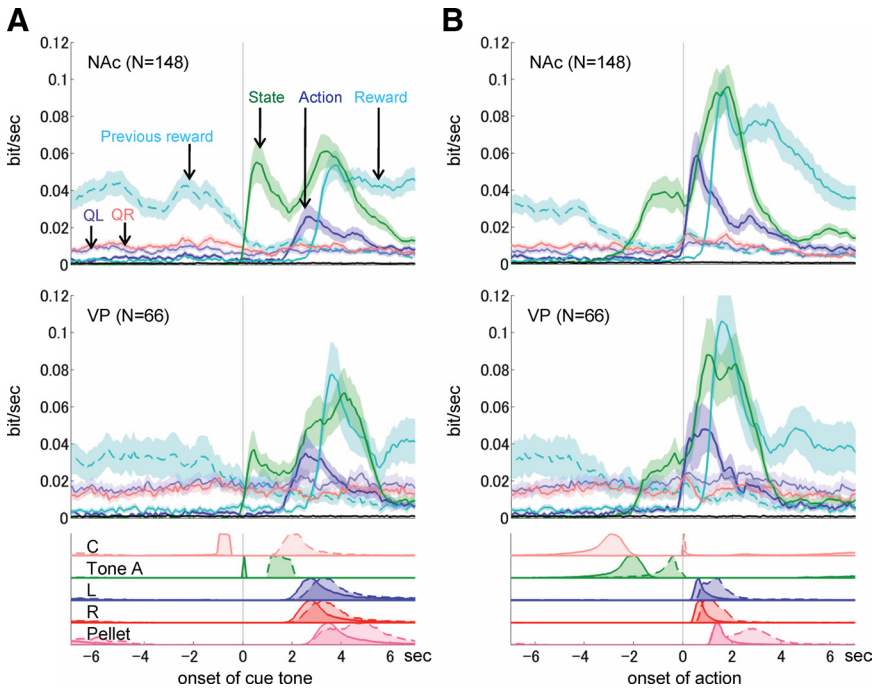

Figure 9. Information coded in the NAc and VP. The mutual information per 1 s between firing and each event was calculated using a sliding time window (duration, 500 ms; step size, 100 ms) and averaged across all neurons recorded in the NAc and VP. The mutual information on action values (QL and QR) was calculated using the estimated action values based on F-Q-learning model with time-varying parameters. ***A*** and ***B*** are aligned with the onset of the discriminative tone and the initiation of action (exit time from the center hole), respectively. The black lines close to the horizontal axes show a threshold indicating significant information ($p < 0.01$). The bottom panels show the normalized distribution of onset time (solid lines) and offset time (broken lines) for nose pokes at C, L, and R, presentation of tone A, and the sensor for detecting a pellet on the dish.

$p < 0.01$ for other models). There was no significant difference between these two models (paired-sample Wilcoxon's signed rank test, $p = 0.067$). Considering the number of the internal parameters of the F-Q model (α_1, κ_1, and κ_2) is less than that of the DF-Q model (α_1, α_2, κ_1, and κ_2), the F-Q-learning model can be regarded as the better model. These results show that the F-Q- and DF-Q-learning models performed better than the best Markov model, a descriptive model generated purely from the rats' choice behaviors. Note that high-order Markov models were expected to provide a good approximation of the upper bound of the prediction accuracy (see Materials and Methods).

Neuronal recording data

Neuronal activity in the NAc and VP was recorded using three or four drivable bundles of eight microwires each. These bundles were advanced between recording sessions so that data from new neurons were acquired in each session. Tracks of electrode bundles are shown in Figure 6. Subsequent analysis was performed for a total of 148 NAc neurons and 66 VP neurons from six rats during 70 sessions.

Neuronal activity in the NAc and VP

VP neurons showed a higher firing rate than NAc neurons overall. The average and SD of the firing rate of VP neurons and NAc neurons during the task were 10.2±9.1 and 2.6±2.6 Hz, respectively.

In both the NAc and VP, a large number of neurons changed their activity depending on the type of cue tone (supplemental Data 2, available at www.jneurosci.org as supplemental material), the action choice, the reward outcome, and/or the reward probability for one of two actions. Figure 7 (rasters for perievent histograms in *C–F* are shown in supplemental Data 2, available at www.jneurosci.org as supplemental material) shows representative examples of these neurons. The colors in the raster plots show the timing of different task events, and the yellow bands in the perievent histogram plot show the time bins in which the firing rates showed significant differences (Mann–Whitney *U* test, $p < 0.01$). The activity of the neurons illustrated in Figure 7, *A* and *B*, changed with the rat's choice during movement to perform nose pokes at the left or right holes, indicating that they were action-coding neurons. The neurons illustrated in Figure 7, *C* and *D*, showed different firing patterns between rewarded and nonrewarded trials after the onset of the reward or no-reward tone, informing rats of reward availability, indicating that they were reward-coding neurons.

Figure 7, *E* and *F*, shows examples of neurons that appeared to represent action values, namely, the reward probability for one of two action candidates (Samejima et al., 2005). The activity of these neurons was compared in different blocks. For the NAc neuron shown in Figure 7*E*, the firing rate around the time of nose poking at the center hole was significantly different between the (50, 10%) and (50, 90%) blocks (Mann–Whitney *U* test, $p < 0.01$), where the reward probability of only the right action changed. However, the firing rate was not significantly different between the (90, 50%) and (10, 50%) blocks, where only the left reward probability was different. Therefore, this neuron can be considered to code the reward probability for the choice of right hole (i.e., the right action value). The VP neuron shown in Figure 7*F* can be considered to code the reward probability for a left action, because there was a significant difference in the firing rate between (10, 50%) and (90, 50%) blocks (Mann–Whitney *U* test, $p < 0.01$) and no difference between (50, 90%) and (50, 10%) blocks.

Figure 8*B* shows the percentage of neurons that encoded different variables at different timings within a trial, as shown in Figure 8*A*. Before the onset of the nose poke action at the center hole (phase 1), the percentage of Q_L-coding neurons in the NAc was 6% (9 of 148), and 11% (7 of 66) in the VP. The activity of these neurons was correlated with the left but not with right reward probability. The percentage of Q_R-coding neurons in the NAc was 7% (11 of 148), and 11% (7 of 66) in the VP. After the onset of the cue tone (phase 2), 54% (80 of 148) of NAc neurons and 50% (33 of 66) of VP neurons showed different firing patterns in response to the choice tone and the no-choice tone, suggesting that they were state-coding neurons. The percentage of Q_L-coding neurons in the NAc was 8% (12 of 148) and 18% (12 of 66) in the VP. The percentage of Q_R-coding neurons in the NAc was 8% (12 of 148) and 9% (6 of 66) in the VP. Before the initiation of action (phase 3), Q_L-coding neurons represented 7% (11 of 148) of the neurons in the NAc and 21% (14 of 66) of those in the VP. Q_R-coding neurons represented 5% (8 of 148) of the neurons in the NAc and 11% (7 of 66) of those in the VP. Neurons encoding the action to be chosen before execution rep-

resented 11% (17 of 148) of the neurons in the NAc and 18% (12 of 66) of those in the VP (action command-coding neurons). After the action onset (phase 4), the proportion of action-coding neurons increased to 41% (61 of 148) in the NAc and 39% (26 of 66) in the VP. After the onset of the reward tone or no-reward tone (phase 5), 63% (93 of 148) of NAc neurons and 68% (45 of 66) of VP neurons coded reward availability.

Comparison of the percentages of coding neurons in the NAc and VP revealed no differences with trial type, action command, chosen action, and reward outcome (Mann–Whitney U test, $p = 0.59$, $p = 0.19$, $p = 0.80$, and $p = 0.45$, respectively). Surprisingly, the percentages of Q_L-coding neurons both after cue onset (phase 2) and before action initiation (phase 3) were significantly higher in the VP than in the NAc (Mann–Whitney U test, $p = 0.03 < 0.05$, $p = 0.004 < 0.01$, respectively), whereas those of Q_R-coding neurons were not significantly different between areas at the two different times (Mann–Whitney U test, $p = 0.81$, and 0.17).

Of the state-coding neurons, 48% (38 of 80) in the NAc and 36% (12 of 33) in the VP showed a greater response to tone A than to tone B. Of the action command-coding neurons, 53% (9 of 17) in the NAc and 50% (6 of 12) in the VP showed greater activity in left-selected trials than in right-selected trials. In the action-coding neurons, 57% (35 of 61) in the NAc and 50% (13 of 26) in the VP showed greater activity in left-selected than in right-selected trials. All of these populations were not significantly different from 50% (Wilcoxon's signed rank test, $p > 0.05$). However, 73% (68 of 93) of the reward-coding neurons in the NAc showed a larger response to the no-reward than to the reward tone. This percentage of cells was significantly larger than the 50% predicted by chance (Wilcoxon's signed rank test, $p < 0.0001$). The percentage of reward-coding neurons in the VP (56%; 25 of 45), however, was not significantly different from 50% (Wilcoxon's signed rank test, $p > 0.05$).

In addition, to examine whether neurons in the core or shell of the NAc had different properties, we separated NAc neurons into tentative "core" and "shell" groups based on the positions of the electrodes. The same analysis described above found no significant differences between core and shell neurons (supplemental Data 3, available at www.jneurosci.org as supplemental material).

In addition to the action value-coding neurons, differential action value-coding neurons (policy-coding neurons) and action-independent value-coding neurons (state value-coding neurons) were also found in the NAc and VP using regression analysis (see Materials and Methods). Action value-coding neurons had already been detected by a comparison of perievent time histograms (Fig. 8*B*) and were also detected in a regression analysis conducted to compare the policy-coding neurons and value-coding neurons as described below (Fig. 8*C*). Q_L-coding neurons, which were found to have a significant regression coefficient to Q_L (t test, $p < 0.01$) but not to Q_R, were present in percentages of 9, 10, and 7% in the NAc, and 15, 24, and 26% in the VP for different time bins (Fig. 8*A*, phases 1–3). In phases 2 and 3, there were significant differences in the size of the population of Q_L-coding neurons in the NAc, and in the VP ($p = 0.007$ and $0.0001 < 0.01$, Mann–Whitney U test). Q_R-coding neurons, which were found to have a significant regression coefficient to Q_R (t test, $p < 0.01$) but not to Q_L, were present in percentages of 11, 9, and 6% in the NAc, and 9, 17, and 14% in the VP. These regression analysis results were consistent with those obtained by comparing the perievent time histograms (Fig. 8*B*). However, the number of the detected action value-coding neurons was slightly larger in Figure 8*C* than in Figure 8*B*. State value-coding neurons, which were found to have a significant regression coefficient to $[Q_L(t) + Q_R(t)]$ but not to $[Q_L(t) - Q_R(t)]$, were present in percentages of 12, 8, and 8% in the NAc, and 18, 20, and 20% in the VP (Fig. 8*D*). In phases 2 and 3, these differences between the sizes of the populations in the NAc and VP were significant ($p = 0.02$ and $0.02 < 0.05$, respectively; Mann–Whitney U test). Policy-coding neurons, which were found to have a significant regression coefficient to $[Q_L(t) - Q_R(t)]$ but not to $[Q_L(t) + Q_R(t)]$, were present in percentages of 10, 9, and 7% in the NAc, and 11, 15, and 14% in the VP (Fig. 8*D*). All percentages shown in Figure 8 were significantly larger than those predicted by chance ($p < 0.01$, binomial test).

To elucidate when and how much information about each event was represented in the NAc and VP, the amount of mutual information shared between neuronal firing and each event was calculated using a sliding time window (duration, 500 ms; step size, 100 ms). Figure 9 shows the average amount of mutual information for all neurons in the NAc and VP aligned at the time of cue tone onset and action onset, respectively. The information coding state in the NAc and VP (green lines) increased immediately after the onset of the cue tone, and again after exiting from the center hole. Whereas the second activation peak during action may reflect the different movements (heading to the left or right holes, or heading directly to the food dish), the first peak may reflect the different strategy the rat adopts after the tone offset. Information related to action choice was very weak in both the NAc and VP before the onset of the choice (before the exit of the center hole), but increased rapidly after the choice onset, and showed a peak around the time of entry to the left or right hole. Information related to the reward started to increase at the presentation of reward/no-reward tone, which was presented at the time of entry to the left or right hole. The reward information gradually decreased, but was sustained until the subsequent trial (previous reward shown in Fig. 8*C*,*D*).

Q_L and Q_R values estimated by the F-Q-learning model were used to calculate information about action values. Information regarding action values Q_L and Q_R in the NAc and VP were found to be larger than the threshold values, indicating a significant level of information ($p < 0.01$) (see Materials and Methods). However, this level of information was still relatively low and without clear peaks, whereas information about state, action, and reward was robustly represented in the NAc and VP.

Discussion

In this study, we analyzed rats' choice strategy and neural coding in the basal ganglia during a free-choice task with stochastic reward. In a model-free statistical analysis of the choice and reward sequence, we verified that a rats' choice probability of an action increases with a reward outcome and decreases, albeit to a lesser degree, with a nonreward outcome (Fig. 2*B*). Moreover, we found that this effect persists until at least three trial steps in the future (Fig. 2*C*). However, the results showed that rats sometimes adopt a win-stay–lose-switch strategy (Fig. 2*C*,*D*) or a perseverating behavior in which a nonrewarding action is continued persistently (Fig. 2*E*). Using model-based prediction of rats' choice from the preceding sequence of actions and rewards, we showed that the F-Q-learning model (Barraclough et al., 2004) and the DF-Q-learning model with forgetting rate and no-reward aversion parameters were able to predict the rats' choice sequence better than the best multistep Markov model (Fig. 5). Considering that the number of the parameters of F-Q-learning model (α_1, κ_1, and κ_2) is less than that of the DF-Q-learning model (α_1, α_2, κ_1, and κ_2), the F-Q model is regarded as superior to the

DF-Q model. The F-Q- and DF-Q-learning models with time-varying parameters not only served as the best descriptive models of rats' choice strategies, but also serve as normative models that explain "why" rats adopt such strategies (Corrado and Doya, 2007).

The results of the current study suggested that neurons in the NAc and VP code different types of information during different phases of behavior. We found evidence that the trial type was coded after the cue tone onset, action values for the choices were coded before and after action initiation, the action was coded after action initiation, and the reward was coded from the reward-tone onset to the start of the next trial. The coding of action values was less dominant, but a significant amount of information regarding action values was observed throughout the trial (Fig. 9).

Modeling rats' choice behavior

A major goal of this study was to formulate a dynamic choice learning model that could describe actual animals' choice behaviors, and also have the ability to normatively explain the aim or goal of the processes described previously (Corrado and Doya, 2007). We started from statistically analyzing rats' choice and reward sequences. This analysis clarified that rats' strategies often change, for example to a win-stay–lose-switch strategy (Fig. 2*C*,*D*) or a persevering strategy (Fig. 2*E*).

The DF-Q-learning model we propose here has the basic goal of maximizing the acquired reward, but is general enough to accommodate the different ways that the values of nonchosen actions and nonrewarded actions are reinforced. Introduction of two additional parameters to the typical Q-learning model, the forgetting rate (α_2) and the impact of a no-reward event (κ_2), enabled the model to represent win-stay–lose-switch behavior when there is a large positive κ_2 (Fig. 3), and to represent persevering behavior when there is a large negative κ_2.

The estimation of time-varying parameters during sessions (supplemental Methods, available at www.jneurosci.org as supplemental material) (Samejima et al., 2004, 2005) enabled the model to predict choices when strategies are changeable. We compared the trial-by-trial choice prediction performance of this model with versions of the Markov model with different orders. The results showed that the DF-Q-learning model with time-varying parameters predicts rats' choices more accurately than the best descriptive model for choice and reward sequences, namely, the third-order Markov model. Surprisingly, introducing the constraint that the forgetting rate α_2 was equal to the learning rate α_1 (F-Q) did not degrade the prediction accuracy, suggesting that rats' strategies are changeable, but within a subspace of strategies that the F-Q model can represent.

Information coding in NAc and VP

The activity of NAc neurons modulated by types of discriminative tones (state-coding neurons) may code either the difference in expected reward (i.e., one pellet for the no-choice tone, and less than one pellet for the choice tone) or the difference in behaviors after the tone. The activity of these state-coding neurons is consistent with previous reports of a subset of NAc neurons that fire selectively to discriminative stimuli in discrimination tasks (Setlow et al., 2003; Nicola et al., 2004a; Wilson and Bowman, 2005). The activity of NAc neurons modulated by different actions during the execution of the poking (action-coding neurons) might code either differences in the physical movements or differences in the spatial position of rats. This notion is consistent with previous reports showing that the responses of a subset of NAc neurons changed with different choices of actions in discrimination tasks and a spatial-delayed matching-to-sample task (Chang et al., 2002; Kim et al., 2007; Taha et al., 2007). In the current study, we found evidence that information related to action lasted beyond the timing of reward delivery after the choice. This finding is consistent with the finding of long-lasting representations of past choices in the ventral striatum, including in the NAc (Kim et al., 2007). Information related to reward in the NAc showed a peak immediately after presentation of the tone associated with reward availability. This activity gradually decreased but lasted until the next trial. This finding is consistent with previous studies reporting that neurons in the NAc and the amygdala, which has a projection to the NAc, respond to reward-predictive stimuli and also consummatory behavior itself (Schoenbaum et al., 1999; Setlow et al., 2003; Nicola et al., 2004b; Wilson and Bowman, 2004, 2005; Wan and Peoples, 2006).

There are few reports of VP recording during learning behaviors compared with the large number of NAc studies. However, a study by Tindell et al. (2004) showed that VP neurons responded to both a conditioned stimulus and a reward-unconditioned stimulus in a pavlovian conditioning paradigm. In the present study, we found that VP neurons coded state information and action information as well as reward information. Notably, the populations of VP neurons coding state, action, and reward information were almost same as those of NAc neurons. This suggests that the NAc and VP might work together, rather than having clearly disparate roles.

Information regarding state, action, and reward is important for updating a behavioral choice. The various representations in the NAc and VP found in this study could be summarized as information necessary for updating a choice behavior. In particular, we showed that a representation of information regarding action in the NAc and VP lasted beyond the peak of information regarding reward (Fig. 9). This simultaneous representation of action and reward information might be necessary to modulate neuronal circuits related to action selection.

The neuronal representation of action value, state value, and policy as well as action and reward have been previously found in the dorsal striatum (Samejima et al., 2005; Lau and Glimcher, 2007, 2008; Pasquereau et al., 2007), the globus pallidus (Pasquereau et al., 2007), and other cortical areas (Platt and Glimcher, 1999; Dorris and Glimcher, 2004). In this study, we found similar value coding in the NAc and VP of rats. Interestingly, the population of left action value-coding neurons and state value-coding neurons in the VP was significantly larger than that in the NAc (Fig. 8). In our experiments, all VP and NAc neurons were recorded from the right hemisphere. Therefore, VP neurons coding action values for contralateral poking behavior were predominant. However, it is still unclear why the representation of action values for the left was not predominant in the NAc.

Although action value-coding neurons exist in the NAc and VP, their population size and information content were relatively small compared with those of the state-, action-, and reward-coding neurons. In addition, the time course of information regarding action value was almost flat (Fig. 9). This finding conflicts with the results of a recent recording study in the caudate nucleus of monkeys, which found a gradual increase in the population size and averaged firing rate of action value-coding neurons toward the onset of action execution (Lau and Glimcher, 2008). Furthermore, information regarding action commands was relatively small in both areas. Similarly, recent rat studies reported that there were few action command-coding neurons in the NAc in discrimination tasks (Kim et al., 2007; Taha et al.,

2007). Considering these relatively minor representations of the action value and the action command, the action evaluation and action selection might not be the primary role of NAc and VP. In any case, these areas might represent state, action, and reward information that can be useful for action evaluation and learning.

In previous studies, action value-coding neurons have been found in the dorsal striatum of monkeys (Samejima et al., 2005; Pasquereau et al., 2007; Lau and Glimcher, 2008). The representation of action values might be dominant in the dorsal rather than the ventral striatum. This would be consistent with the results of a lesion study in rats, which suggested that the dorsomedial striatum is involved in action–outcome learning (Yin et al., 2005). Moreover, a functional magnetic resonance imaging study in humans also suggested that the dorsal striatum contributes to action selection, whereas the ventral striatum contributes to reward expectation (O'Doherty et al., 2004).

References

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

Cardinal RN (2006) Neural systems implicated in delayed and probabilistic reinforcement. Neural Netw 19:1277–1301.

Cardinal RN, Cheung TH (2005) Nucleus accumbens core lesions retard instrumental learning and performance with delayed reinforcement in the rat. BMC Neurosci 6:9.

Cardinal RN, Howes NJ (2005) Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. BMC Neurosci 6:37.

Chang JY, Chen L, Luo F, Shi LH, Woodward DJ (2002) Neuronal responses in the frontal cortico-basal ganglia system during delayed matching-to-sample task: ensemble recording in freely moving rats. Exp Brain Res 142:67–80.

Corrado G, Doya K (2007) Understanding neural coding through the model-based analysis of decision making. J Neurosci 27:8178–8180.

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

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Dorris MC, Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron 44:365–378.

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

Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. Curr Opin Neurobiol 10:732–739.

Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26:8360–8367.

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds). Cambridge, MA: MIT.

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.

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.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Nicola SM, Yun IA, Wakabayashi KT, Fields HL (2004a) Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task. J Neurophysiol 91:1840–1865.

Nicola SM, Yun IA, Wakabayashi KT, Fields HL (2004b) Firing of nucleus accumbens neurons during the consummatory phase of a discriminative stimulus task depends on previous reward predictive cues. J Neurophysiol 91:1866–1882.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

O'Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35–53.

Panzeri S, Treves A (1996) Analytical estimates of limited sampling biases in different information measures. Netw Comput Neural Syst 7:87–107.

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.

Paxinos G, Watson C (1998) The rat brain in stereotaxic coordinates, Ed 4. San Diego: Academic.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400:233–238.

Samejima K, Doya K, Ueda Y, Kimura M (2004) Estimating internal variables and parameters of a learning agent by a particle filter. In: Advances in neural information processing systems (Thrun S, Saul LK, Sholkopf B, eds). Cambridge, MA: MIT.

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

Schoenbaum G, Chiba AA, Gallagher M (1999) Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. J Neurosci 19:1876–1884.

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

Seo H, Lee D (2007) Temporal filtering of reward signals in the dorsal anterior cingulated cortex during a mixed-strategy game. J Neurosci 27:8366–8377.

Setlow B, Schoenbaum G, Gallagher M (2003) Neural encoding in ventral striatum during olfactory discrimination learning. Neuron 38:625–636.

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782–1787.

Sutton RS, Barto AG (1998) Reinforcement learning. Cambridge, MA: MIT.

Taha SA, Nicola SM, Fields HL (2007) Cue-evoked encoding of movement planning and execution in the rat nucleus accumbens. J Physiol 584:801–818.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887–893.

Tindell AJ, Berridge KC, Aldridge JW (2004) Ventral pallidal representation of pavlovian cues and reward: population and rate codes. J Neurosci 24:1058–1069.

Wan X, Peoples LL (2006) Firing patterns of accumbal neurons during a pavlovian-conditioned approach task. J Neurophysiol 96:652–660.

Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8:279–292.

Wilson DI, Bowman EM (2004) Nucleus accumbens neurons in the rat exhibit differential activity to conditioned reinforcers and primary reinforcers within a second-order schedule of saccharin reinforcement. Eur J Neurosci 20:2777–2788.

Wilson DI, Bowman EM (2005) Rat nucleus accumbens neurons predominantly respond to the outcome-related properties of conditioned stimuli rather than their behavioral-switching properties. J Neurophysiol 94:49–61.

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22: 513–523.