Neuron, Vol. 41, 269–280, January 22, 2004, Copyright ©2004 by Cell Press

Dopamine Neurons Can Represent Context-Dependent Prediction Error

Hiroyuki Nakahara,[1,5,]* Hideaki Itoh,[2,5]
Reiko Kawagoe,[3,5] Yoriko Takikawa,[3,5]
and Okihide Hikosaka[4]
[1]Lab for Mathematical Neuroscience
RIKEN Brain Science Institute
2-1 Hirosawa
Wako, Saitama 351-0198
Japan
[2]Department of Computational Intelligence
and Systems Science
Interdisciplinary Graduate School of Science
& Engineering
Tokyo Institute of Technology
Yokohama 226-8502
Japan
[3]Department of Physiology
Juntendo University
School of Medicine
Tokyo 113-8421
Japan
[4]Lab of Sensorimotor Research
National Eye Institute
National Institute of Health
Bethesda, Maryland 20892

Summary

Midbrain dopamine (DA) neurons are thought to encode reward prediction error. Reward prediction can be improved if any relevant context is taken into account. We found that monkey DA neurons can encode a context-dependent prediction error. In the first noncontextual task, a light stimulus was randomly followed by reward, with a fixed equal probability. The response of DA neurons was positively correlated with the number of preceding unrewarded trials and could be simulated by a conventional temporal difference (TD) model. In the second contextual task, a reward-indicating light stimulus was presented with the probability that, while fixed overall, was incremented as a function of the number of preceding unrewarded trials. The DA neuronal response then was negatively correlated with this number. This history effect corresponded to the prediction error based on the conditional probability of reward and could be simulated only by implementing the relevant context into the TD model.

Introduction

Midbrain dopamine (DA) neurons are thought to play a key role in reinforcement learning. Schultz and colleagues showed that DA neurons respond to unexpected reward with a burst of spikes and respond to a reward omission with a pause of spikes. They further showed that DA neurons start to respond to a predictive stimulus rather than to the primary reward itself in a task where the reward always follows the stimulus (Ljungberg et al., 1992; Schultz, 1998). The data led to the hypothesis that DA neurons encode a reward expectation error (Barto, 1995; Houk et al., 1995; Montague et al., 1996; Schultz et al., 1995, 1997). Such an error signal can be used for learning to obtain reward (Graybiel et al., 1994; Rescorla and Wagner, 1972; Reynolds et al., 2001; Sutton and Barto, 1981).

More specifically, the theory of temporal difference (TD) learning has been very successful in relating the activity of DA neurons to reinforcement learning (Barto, 1995; Houk et al., 1995; Montague et al., 1996; Schultz, 1998; Schultz et al., 1997). The idea that the basal ganglia circuit realizes TD learning is attractive, because a family of TD learning allows one to learn the sequence of optimal decisions (Sutton and Barto, 1998). A computer program of TD learning can learn to play backgammon with a world-class human expert (Tesauro, 1994). In TD learning, the expected reward is computed by summarizing the time-delayed rewards; this is acquired by a function, called the value function, based on past experiences. For this acquisition, the value function uses a reinforcement signal, called the TD error. The TD error is a specific form of reward prediction error, i.e., the difference between the reward and the reward expectation with some adjustments (Equation 1, below). The strength of TD learning lies in the way the TD error is used, i.e., the backward propagation of the TD error over the sequence of events. Owing to this characteristic, once learning is established, a system can learn optimal decisions (even ones in earlier events) to obtain reward that may come after a number of events. DA responses mentioned above appear to represent this backward-propagated TD error (Fiorillo et al., 2003; Satoh et al., 2003). Thus, if realized in the basal ganglia circuit, TD learning would allow animals to learn and make optimal decisions to reach reward (Arbib and Dominey, 1995; Montague et al., 1995; Nakahara et al., 2001; Suri and Schultz, 1998).

However, the exact correspondence between DA response and the TD error remains to be established. What kind of TD error do DA neurons represent? What information do DA neurons take into account to provide reward prediction error? So far, most physiological studies on DA neurons of primates have been done under conditions in which the reward probability was determined by the sensory information given in a trial (of an experimental block), but was not influenced by preceding trials (Fiorillo et al., 2003; Hollerman and Schultz, 1998; Mirenowicz and Schultz, 1994; Romo and Schultz, 1990; Waelti et al., 2001). Accordingly, theoretical studies on DA responses have assumed that the TD error is computed based only on the sensory information given in a trial (Berns and Sejnowski, 1998; Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Suri, 2001; Suri and Schultz, 2001).

In contrast, there are many real-world situations in which the reward probability is different depending on

*Correspondence: hiro@brain.riken.go.jp
[5]These authors contributed equally to this work.

contexts. In brief, the context is meant to be the information that is contingent on preceding events beyond one trial. Reward prediction can be improved by taking the relevant context into account. For example, suppose that you participate in a lottery to draw a red ball from a bag with one red ball and nine blue balls. Obviously, the probability of winning the lottery is 1/10. The probability becomes 1, however, if you know and remember that there is only one red ball left in the bag when nine people have already drawn the blue balls. To improve reward prediction, the context must be relevant. Remembering which hand other people used to draw balls never helps improve reward prediction. Later, we further clarify what the context refers to in the present paper.

We now ask the following: what happens in DA responses if the context has something to say about the likelihood of reward delivery? As we indicated above, the reward expectation would be completely different depending on whether the context is considered or not. Can DA neurons cope with such a situation? If so, does TD learning need to be modified?

To answer these questions, we had monkeys perform a task in which reward probability changed in relation to the preceding trials, whether rewarded or unrewarded. This task is called the contextual task. We also had monkeys perform a task in which reward probability was fixed. This is called the noncontextual task. The noncontextual task is used as a control for investigation of the contextual task. A major objective of our study was to examine whether DA neurons can adapt their activity by taking into account a relevant context in the contextual task. We found that the characteristics of DA responses in the noncontextual task matched with the TD model used in previous studies (called the conventional TD model). In the contextual task, we found that after extensive experience with the task, DA responses represented a reward prediction error better than that predicted by the conventional TD model. We propose the contextual TD model that uses a relevant context and show that DA neurons exhibit a context-dependent prediction error similar to the contextual TD model.

Results

Dopamine Response and Reward Prediction Error in Noncontextual Task

We first investigated the activity of DA neurons using a classical conditioning task (Figure 1A; Experimental Procedures). A visual cue (a spot of light) was followed by a reward (a drop of water) with 50% probability (at random). The probability was independent of whether the preceding trials were rewarded, and thus this task is called "noncontextual task." DA neurons responded phasically to both the cue and the reward. A typical example is shown in Figure 1A. When the reward was given, the DA neuron increased its activity (red-shaded period in Figure 1A); when the reward was not given, the neuron decreased its activity (blue-shaded period). We recorded from 21 DA neurons in one monkey (G), finding that a majority of them differentiated between the presence and absence of reward (between the shaded periods; 16/21, 76% by t test, $p < 0.05$). This observation is consistent with a previous study (Fiorillo et al., 2003).

This observation can be explained, roughly, by the notion that DA neurons carry a reward prediction error. Simply speaking, the cue predicts a 50% reward on the average, since the reward was given randomly with 50% chance. If the reward is then given (100% reward), the reward prediction error is $+50\%$. If no reward is given (0% reward), the reward prediction error is -50%. The DA response to the presence and absence of reward followed this pattern, although the magnitude of the actual DA response does not seem to precisely follow $\pm 50\%$, possibly due to the floor effect for the suppressive DA response to the reward undelivery.

From the viewpoint of TD learning, the DA response corresponds to the TD error, given by

$$\mathrm{TD} = \gamma V(s') + r - V(s), \qquad (1)$$

where the sensory state changes from s to s'; $V(s)$ and $V(s')$ are the value functions in each sensory state; and r is the reward in this state transition (Experimental Procedures). Value function is the function that accepts a state as input and returns the expected reward as output (Experimental Procedures). The discount factor γ ($0 \leq \gamma \leq 1$) determines by how much delayed rewards are discounted. This form of the TD error is due to the construction of the value function ("TD" can be regarded as referring to this fact; see Experimental Procedures for more details). Our simple explanation above, featuring the $\pm 50\%$ error, corresponds to the case when $\gamma = 0$, because the above explanation is based on $r - V(s)$.

We further noted that the magnitude of the DA neuronal responses changed depending on the history of reward delivery. In Figure 1B, the mean response magnitude is plotted against the number of trials since the last rewarded trial (postreward trial number, PRN). The magnitude of the excitatory response to reward increased with PRN, while that of the inhibitory response to no-reward decreased with PRN. Each of the excitatory and inhibitory responses formed a statistically significant positive slope (by F test, $p < 0.01$ to examine whether a nonzero slope exists).

Why should there be a positive slope even though the reward prediction error is always either $+50\%$ or -50%? Indeed, this history effect can be expected if we assume that DA neurons adjust their response, i.e., reward prediction error, gradually trial after trial. Let us explain and quantify this viewpoint in terms of the TD model. We will first provide an intuitive explanation, followed by the simulation result. In this task, the probability of reward is 50% *on average*, and therefore, the TD error is $\pm 50\%$ *on average* (in case of $\gamma = 0$). However, note that the probability of reward would fluctuate "locally" over trials. Among five recent trials, for example, there may be five reward trials or no reward trials. In TD learning (Sutton, 1991; Sutton and Barto, 1998), the value function is modified, using TD error, as

$$V(s) \rightarrow V(s) + \alpha \mathrm{TD}, \qquad (2)$$

where α is a small learning constant. The value function adjusts its output at every state transition, albeit only gradually (due to the small learning constant α). The TD error changes as the value function changes (Equation

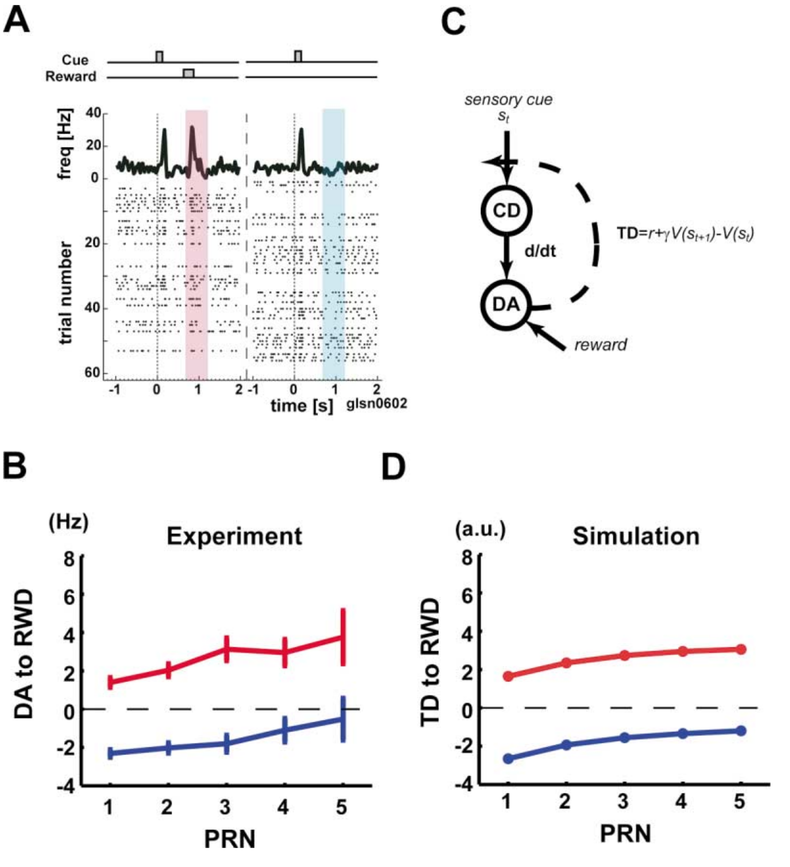

Figure 1. Dopamine Neuronal Responses in Noncontextual Task: Classical Conditioning Task with 50% Probability of Reward

Dopamine (DA) responses to reward delivery and omission match the response of the conventional TD model.

(A) Task procedure and an example of DA response. A central spot of light (duration, 150 ms) was presented in random intervals, and a reward (drop of water) was delivered 500 ms later with a 50% probability (Experimental Procedures). In the raster, trials are shown in chronological order from top to bottom but are separated for rewarded trials (left) and unrewarded trials (right). Spike activity of a DA neuron is aligned with cue onset. Red- and blue-shaded regions indicate the time window (700–1200 ms after the cue onset) to compute the reward responses to the delivery and omission of reward in (B). Spike histogram was created by using 20 ms bins, where the average in each bin (say, j-th bin) was taken by using 0.25 m(j − 1) + 0.5 m(j) + 0.25 m(j + 1), and m(j) is the spike count at the j-th bin.

(B) Dependency of DA neuronal responses on postreward trial number, called PRN. Population average of DA response (n = 21) to the reward delivery (red) and the reward omission (blue) is shown with respect to PRN. DA responses are shown after subtracting the average firing rate of all trials. Error bars indicate the standard errors. Although PRN could be larger, it is shown up to five because the number of samples for larger PRNs decreased significantly.

(C) Schematic diagram of conventional TD model.

(D) TD error response to the reward delivery (red) and nondelivery (blue) with respect to PRN. Error bars are not shown since they are negligibly small.

1). Therefore, the TD error is influenced by the local fluctuation of recent trials, being ±50% only on average.

Suppose that no reward trials are repeated. By the local nature of TD learning (Equation 2), the value function $V(s)$ is modified more toward interpreting that the visual cue indicates no reward. With this lowered reward expectation, a reward delivery in the next trial is more "surprising;" it signifies a higher positive reward prediction error. Consequently, DA neurons give a stronger excitatory response (i.e., DA responses to reward for the larger PRN in Figure 1D). On the other hand, the absence of reward is less surprising when no reward trials are repeated, signifying a lower negative reward expectation error. Consequently, DA neurons give a weaker inhibitory response (i.e., responses to nonreward for the larger PRN in Figure 1D). The story is opposite if reward trials are repeated: the value function $V(s)$ is modified more toward interpreting that the visual cue indicates reward. With this heightened reward expectation, the delivery of reward signifies a lower positive prediction error so that DA neurons give a lower excitatory response (i.e., responses to reward for PRN = 1 in Figure 1D). On the other hand, the absence of reward signifies a higher negative prediction error, so that DA neurons give a stronger inhibitory response (i.e., responses to nonreward for PRN in Figure 1D).

Before moving to the simulation to examine the above intuition, let us briefly explain the hypothesized scheme of TD model in Figure 1C, based on previous studies (Houk et al., 1995; Schultz et al., 1997). DA neurons are considered to receive reward information (r) and the outputs of the value function. For DA neurons to emit TD error (e.g., at time t), outputs of the value function at two consecutive times (e.g., t and $t + 1$) should be subtracted from each other (with the multiplication of the discount factor; $\gamma V(s') - V(s)$ or equivalently $\gamma V(s_{t+1}) - V(s_t)$). This operation is indicated by *d/dt* in the figure (Schultz et al., 1997). With reward information (r), DA response (TD error) is given by TD $= r + \gamma V(s') - V(s)$.

Caudate (CD) neurons receive sensory information (s) through the projection of sensory-related cortical areas and are supposed to emit the output of the value function (Houk et al., 1995). It is known that GABAergic CD neurons project to the substantia nigra and connect to DA neurons directly and indirectly (Parent and Hazrati, 1994). The direct and indirect pathways are considered to realize the operation of $\gamma V(s') - V(s)$ (Houk et al., 1995). Although their exact relations to TD learning is still under debate (Dayan, 2002; Doya, 2002), we take the hypothesized scheme of Figure 1C as a basis for our simulations.

When running the simulation of the noncontextual task with TD model, we used a discrete time and each trial consisted of two states, where one state started with the cue appearance, transiting to the next state with reward delivery (or nondelivery) (Experimental Procedures). The simulated TD error (Figure 1D) followed the same pattern as DA neurons (Figure 1B), i.e., the positive slope. In order to plot Figure 1D, we had to determine the two free parameters of the model, namely the discount factor γ and the learning constant α. As an exhaustive search of using the minimal square loss, we chose the best values that gave the maximal correlation between the TD errors and DA responses over different PRNs (Experimental Procedures). The chosen values were $\alpha = 0.3$ and $\gamma = 0.9$. We emphasize that the tendency for a positive slope was observed across the almost entire parameter range.

To summarize, the characteristics of DA neuronal responses in the noncontextual task matched the characteristics of the TD model. The positive slope of DA response dependency on PRN (Figure 1B) is in accord with the local nature of TD learning. This phenomenon could have been inferred from the TD hypothesis, but has never been shown as an experimental result.

Dopamine Response in Contextual Task: Postreward Number Effect Is Reversed

In the noncontextual task, the reward probability was always 50% in each trial. However, the reward probability may be different if a relevant context is taken into account in a more general setting. One example of such a context is the sequential order of preceding trial types; this factor was implemented in the task below (therefore called contextual task).

The contextual task was basically a memory-guided saccade task with four possible target positions, but reward was given for a correct saccade to only one of these positions (Figure 2A; Kawagoe et al., 1998). A visual cue stimulus indicated not only the saccade goal but also whether a reward could be obtained after the upcoming saccade. Within a block of 60 trials, one out of four directions was associated with reward, while the other three directions were not rewarded. No indication was given as to which direction was presently rewarded, except for the actual reward. For each DA neuron, the task was performed in at least four blocks with four different reward directions (Figure 2B).

We used a pseudorandom schedule to choose target direction in each trial: within each subblock of four trials, each of all four directions was chosen randomly but only once (Figure 3E). The start or end of each subblock was not indicated to the monkey. This schedule let the reward probability be 25% and at the same time induced a specific structure of the reward probability in relation to the preceding trials. For example, a rewarded trial always came if and after there were six consecutive unrewarded trials, because the number of six trials was the maximal number of consecutive unrewarded trials within two consecutive subblocks (Figure 3E).

Figure 2B illustrates a typical example of DA neuronal activity. In the block when the right-up (RU) direction was rewarded, the DA neuron responded to the RU cue with a phasic excitation, whereas its activity decreased in response to the other cues that indicated no reward. When the reward direction was changed in another block, the DA neuron changed its activity completely but still followed the same principle: an excitatory response to the reward-indicating cue and an inhibitory response to the nonreward-indicating cue (an example with raster is shown in Figure 3A). This differentiated response was true for almost all DA neurons (by t test, $p < 0.05$, 31/32 in monkey G; 16/16 in monkey H, which are from "the late stage;" see below; Kawagoe et al., 2003).

The TD model can approximately explain the response pattern. Since the reward was given with 25% probability on average, the reward prediction before the cue presentation would be 25% on average. If the cue indicates reward (100% reward), then the reward prediction error is +75%, while if the cue indicates no reward (0% reward), the reward prediction error is −25%. DA neurons showed no response to reward itself except for the first couple of trials in a block of the experiment. This is probably because the presence or absence of reward had already been indicated by the cue, and as a result there was no reward prediction error at the time of reward delivery. Moreover, the monkeys did not know which direction was rewarded at the beginning of the block.

However, the DA neuronal responses were different between the noncontextual and contextual tasks when we considered the history of reward delivery (PRN effect). In monkeys with sufficient experience of the task (late stage of learning), the PRN effect showed a statistically significant negative slope. This was true for both monkey G (Figure 3C) and monkey H (Figure 3D) in both reward-indicating (red line) and nonreward-indicating cues (blue) (F test, $p < 0.01$). The results were opposite to the positive slopes obtained in the noncontextual task. Interestingly, there was no clear PRN effect when one monkey (G) was examined when he was less experienced (early stage) (F test, $p > 0.05$; Figure 3B; Y. Takikawa et al., 1999, Soc. Neurosci., abstract). These results suggest that DA neurons acquired the PRN effect with a negative slope through experience.

The early-stage data were collected after the monkey (G) had already experienced 60 blocks but before 300 blocks, roughly corresponding to 2–6 weeks' experience. The late stage refers to the data collected after the monkey (G) experienced around 600 blocks, roughly corresponding to more than 5 months' experience (Experimental Procedures). Note that this early stage does not represent the first encounters with the task, when the monkey might not have fully acquired the task procedure. When we compared the error rate between two stages, they were not so different (0.084 and 0.081; Experimental Procedures), showing that the monkeys could perform the task well even in this early stage.

Probability of Reward Conditional to Postreward Number and Dopamine Response

One critical difference between the two tasks lies in the probabilistic structure over the trials. In the noncontextual task, the probability of a reward trial was always the same (50%), regardless of what happened in the preceding trials. In contrast, in the contextual task, the probability of a reward trial was 25% but varied in relation to the preceding trials. This is induced by the pseu-

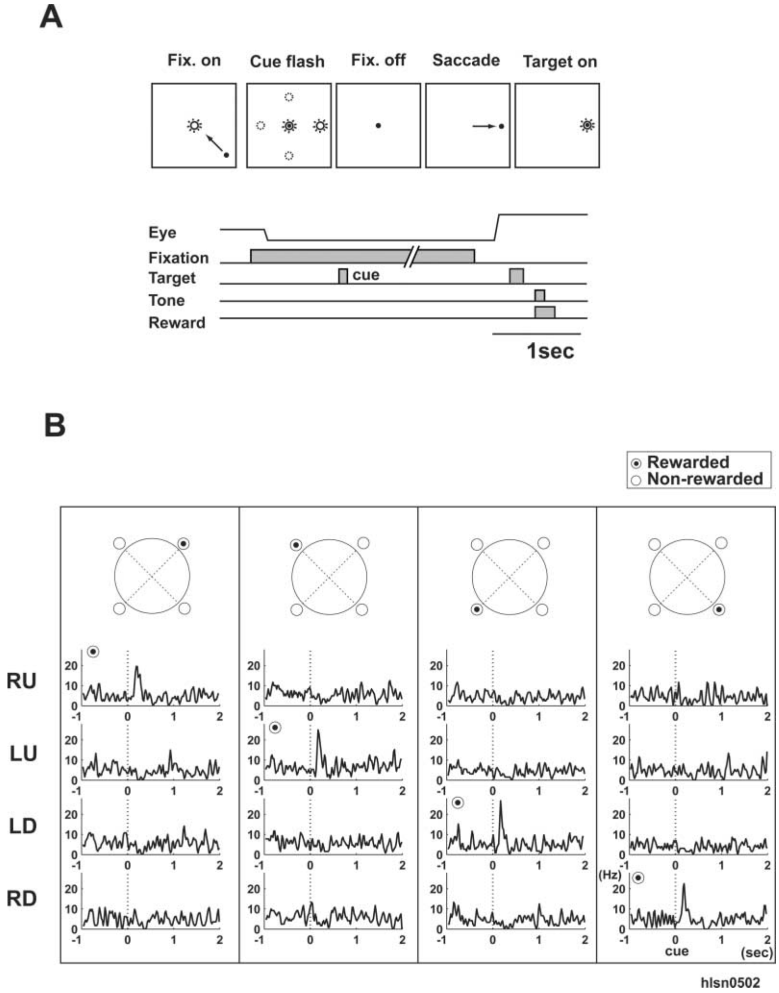

Figure 2. DA Neuronal Responses in Contextual Task

(A) Task was a one-direction-rewarded version of the memory-guided saccade task (1DR). A peripheral cue stimulus, which was presented in one of four directions, indicated the position of the saccade to be made later after the fixation point went off. Only one direction was rewarded throughout a block of experiments.

(B) Responses of a DA neuron to the cue stimuli. Four different directions were rewarded in four different blocks, and spike histograms are shown in four columns. The spike histograms, which are aligned with cue onset, are shown separately for different cue directions (RU, right-up; LU, left-up; LD, left-down; RD, right-down). Rewarded direction is indicated by a bull's eye mark. Target eccentricity was 20°.

dorandom schedule for target selection mentioned above. Mathematically, this type of probability is called conditional probability. While there can be various types of conditional probability, the relevant type here is the probability of reward conditional to PRN, i.e., Pr[reward|PRN]. In the contextual task, this conditional probability of reward changed in relation to the number of PRN, whereas the probability of reward was the same as Pr[reward] = 0.25. In Figure 4B, we plotted Pr[reward|PRN] based on the pseudorandom schedule (Experimental Procedures). The conditional probability of reward increased with PRN. It was the lowest (0.0625) if the preceding trial was rewarded (PRN: 1), while it was the highest (1.0) if six preceding trials were not rewarded (PRN: 7). The actual probabilities during recording of DA neurons (dashed line in Figure 4B) closely matched the theoretical probabilities (solid line). Therefore, in the contextual task, the conditional probability Pr[reward|PRN] should give a better prediction of an upcoming reward than the probability Pr[reward], which remains at 0.25 in any trial.

DA neurons appeared to use the conditional probability in responding to the cue. Consider a trial following a rewarded trial, where the trial has the lowest conditional reward probability (PRN = 1 in Figure 4B). A reward-indicating cue in this trial then signifies a higher positive reward expectation error and hence DA neurons showed stronger excitatory responses (Figures 3C and 3D). On the other hand, a nonreward-indicating cue at PRN 1 signifies a lower negative reward expectation and hence DA neurons showed weaker inhibitory responses (Figures 3C and 3D). In contrast, the conditional reward probability becomes higher as the current trial is preceded by more no-reward trials (higher PRN in Figure 4B): a reward-indicating cue induces weaker DA excitatory responses (Figures 3C and 3D), while a nonreward-indicating cue induces stronger DA inhibitory response (Figures 3C and 3D).

For the results in the noncontextual task, we discussed the local nature of TD learning, so that it is natural to expect that similar effects also exists in DA responses in the contextual task. Our intuitive explanation above, however, is given by ignoring this effect. Below, we quantify the above intuition by simulations of TD model.

We previously showed that the saccade velocity in the contextual task exhibits the PRN dependency (Takikawa et al., 2002b). As PRN increases, the saccade velocity for unrewarded trials increased and the percentage of error trials for unrewarded trials decreased. The PRN dependency was unclear for rewarded trials, most likely due to ceiling and flooring effects. These results may be interpreted as if the monkeys knew the condi-

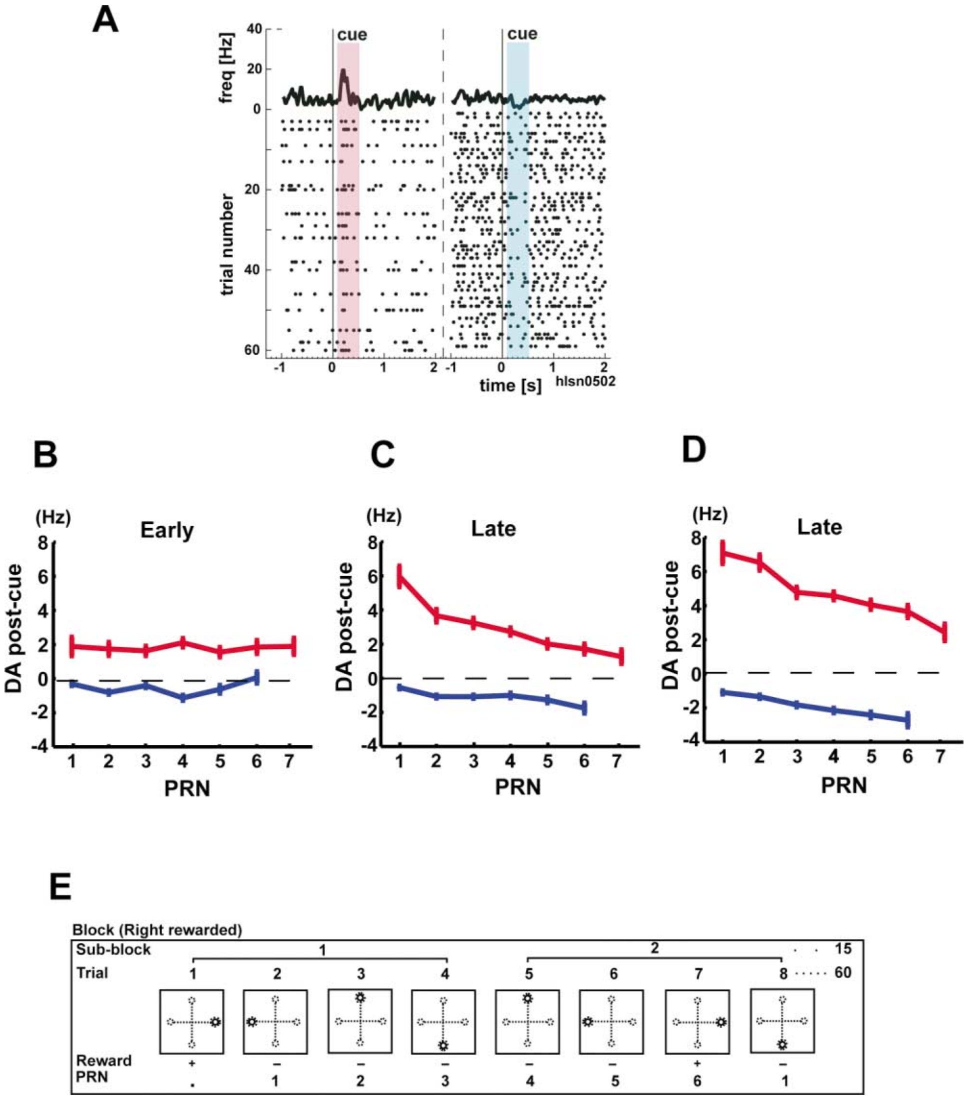

Figure 3. Context Dependency of DA Neuronal Responses

(A) Cue response of a DA neuron in a block of 1DR (corresponding to the block in the most left column in Figure 2B), shown as a raster display and histogram. Trials are shown in chronological order from top to bottom but are separated for rewarded trials (left) and unrewarded trials (right). Data are aligned on cue onset. Red- and blue-shaded regions indicate the time window (100–500 ms after the cue onset) to compute the cue responses in (B)–(D).

(B–D) Population averages of DA responses to the reward-indicating cue (red) and to the nonreward-indicating cue (blue) are shown with respect to PRN. Due to the method of pseudorandomization, PRN ranged from 1 to 7 for the reward-indicating cue and from 1 to 6 for the nonreward-indicating cue. Data are shown for monkey G in the early stage (B, n = 21) and the late stage (C, n = 32), and for monkey H in the late stage (D, n = 16). The responses are shown after subtracting the average firing rate of all trials.

(E) Example to indicate the pseudorandomization that determined the sequence of trials in the contextual task.

tional probability of reward: given a higher PRN, the monkeys are more willing to perform the current trial with a higher reward expectation for the next trial.

It was possible, at least in theory, that the monkeys could predict reward more accurately than by knowing only PRN. Consider trial #1 in Figure 3E, which is rewarded. If the monkey understands the pseudorandom schedule and knows that it is the first trial in a subblock, the monkey can predict that the reward probability is zero in the next three trials. Thus, knowing the position of the current trial in a subblock makes reward prediction more accurate. However, we think that this is unlikely, because it was extremely difficult to keep track of the boundary between subblocks in this task: the start or end of each subblock was not indicated to the monkey, and furthermore, the first trial of an experimental block could start at any element of a subblock (Experimental Procedures). Nonetheless, we examined whether DA response or saccade velocity could differentiate, depending on the position of the current trial in a subblock. Results indicated that neither DA response nor saccade velocity differentiated, implying that the monkeys could

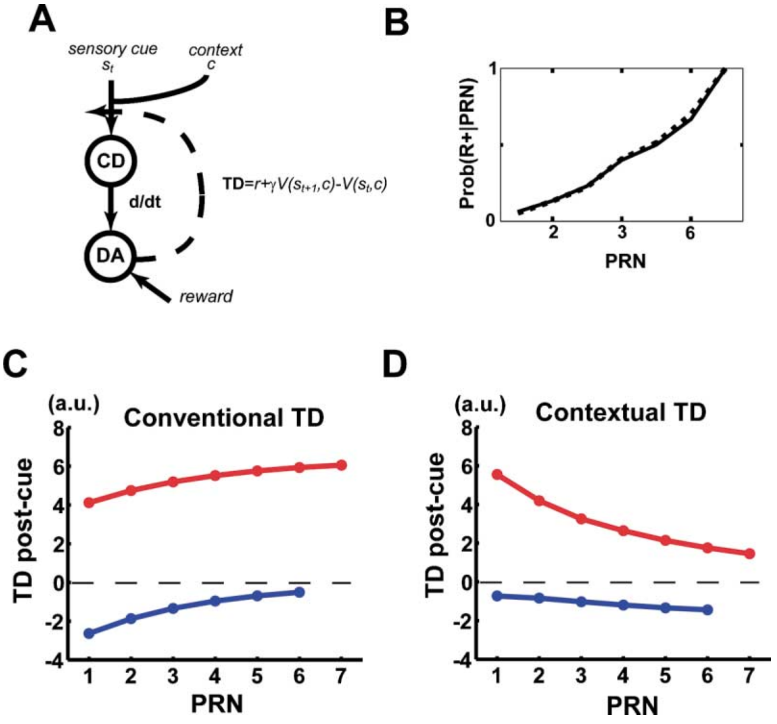

Figure 4. Performance of TD Models

(A) Schematic diagram of the contextual TD model that uses PRN as a context input.

(B) Probability of reward, conditional to PRN. Solid line indicates theoretical values, while dashed line indicates empirical values from the experiment.

(C) TD error of the conventional TD model to the reward cue (red) and the nonreward cue (blue).

(D) TD error of the contextual TD model.

not keep track of the subblock boundary (see Supplemental Figure S2 and Table S1 at http://www.neuron.org/cgi/content/full/41/2/269/DC1).

Contextual Temporal Difference Model

Can the TD model predict the sequential order dependency of the DA neuronal response in the contextual task? According to the TD model in its current form (hereafter called "the conventional TD model"), the magnitude of the excitatory response to the reward-indicating cue should increase with PRN, while that of the inhibitory response to the nonreward-indicating cue should decrease with PRN. Because the conventional TD model does not have access to PRN, it can only learn the probability of reward, not the conditional probability. This is true even when values for the discount factor γ and the learning constant α are examined by an exhaustive search. This pattern, however, was clearly dissimilar to the DA neuronal response when the monkey was highly experienced in this task (Figures 3C and 3D).

We thus revised the TD model by implementing the memory of PRN (Figure 4A) and call it the contextual TD model (H. Itoh et al., 2002, Soc. Neurosci., abstract). In the contextual TD model, the value function *V*(*s*, *c*) is now a function of the current sensory input *s* and the context *c*, i.e., PRN. The TD error of the contextual TD model is then given by

$$\mathrm{TD} = \gamma V(s', c) + r - V(s, c), \tag{3}$$

and the learning of the value function occurs, using this TD error.

The only difference between the conventional and contextual TD models is that the contextual TD model has the context input *c* (with "counting errors;" see below). We used the same parameter values that were chosen in the noncontextual task ($\alpha = 0.3$, $\gamma = 0.9$) for both models in the contextual task simulations. Each trial consisted of three states, namely precue (a state before cue flash), postcue (a state after cue flash), and fixation-off states (a state after fixation-off). At the transition from one state to another, the model could make an action, namely, a saccade in one of the four directions or no movement. Reward was given to the model after it correctly made a saccade in the fixation-off state in the rewarded condition; otherwise no reward was given. The order of trials in a block of simulation was determined by the same pseudorandom schedule as in the experiments.

In Figure 4D is shown the simulation results using the contextual TD model. Here, the TD error shows the PRN effect with a negative slope for both rewarded and nonrewarded trials. The results were very similar to what we observed as DA neuronal responses in the contextual task (Figures 3C and 3D). In contrast, the simulation using the conventional TD model (Figure 4C) was unsuitable since it produced positive slopes.

In the result of Figure 4D, we allowed the contextual TD model to make an error in counting the number of preceding unrewarded trials (i.e., PRN), which we call the counting error, and set its probability at 40%. We emphasize that the wide range of counting errors (roughly 0%–60%) leads to the same qualitative result (Experimental Procedures; see Supplemental Figure S1 at http://www.neuron.org/cgi/content/full/41/2/269/DC1; Meck, 1996; Nieder and Miller, 2003; reference therein). Notably, as the counting error further increases, the slope of the TD-PRN curve changes from negative to positive. This observation leads us to one interesting speculation: the dependency of DA response on PRN

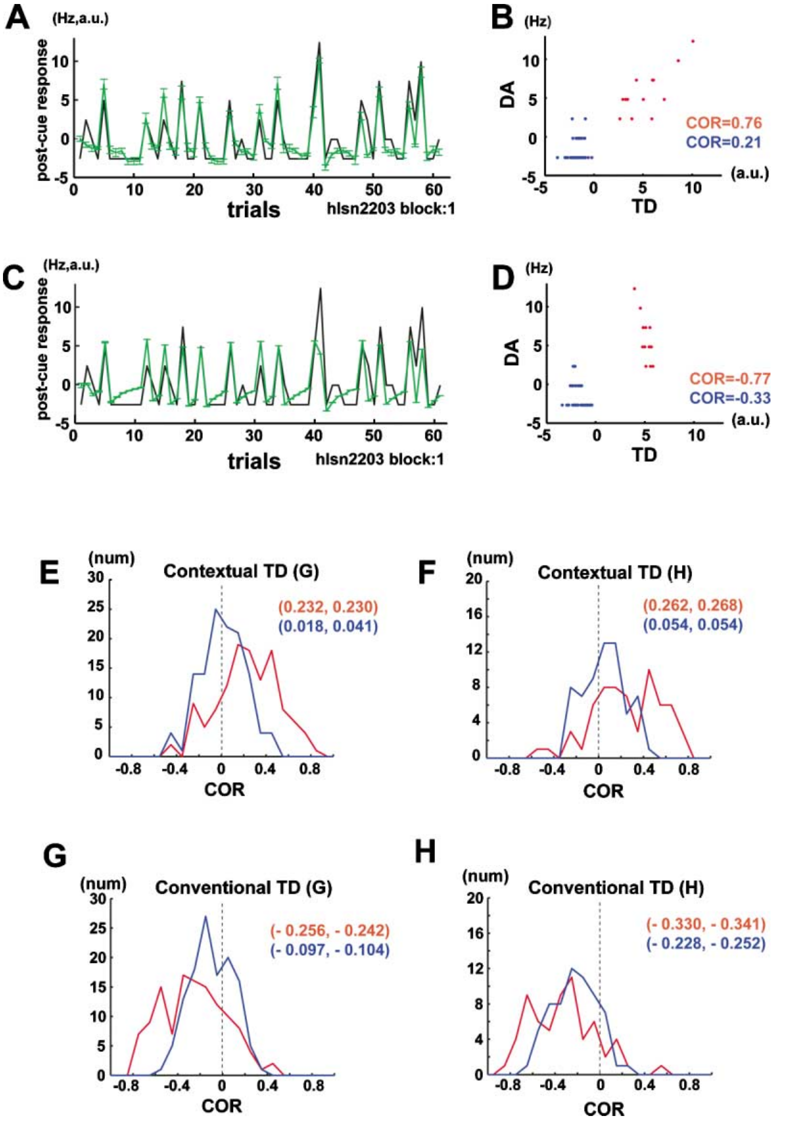

Figure 5. Trial-by-Trial Comparison between the Actual and Simulated DA Responses

(A) For one block of 1DR are plotted the actual responses of a DA neuron to the cue (black) and the corresponding simulated response (TD errors) (green) based on the contextual TD model.

(B) Data in (A) are replotted as a scatter plot with the actual DA responses (ordinate) and the simulated DA responses (abscissa). We did not include the first eight trials of a block (Experimental Procedures; this also applies to D). Red and blue points indicate the trials with the reward-indicating and nonreward-indicating cues, respectively. The correlation coefficient (COR) was calculated separately for the two cue types.

(C) TD errors of the conventional TD model are superimposed with DA responses in the same experiment block as in (A).

(D) Corresponding scatter plot.

(E–H) Histograms of CORs between the DA responses and TD errors. (E) and (F) show the histogram for the COR of the contextual TD errors with DA responses of the monkey G and the monkey H, respectively. (G) and (H) show the COR histogram for the conventional TD errors with DA responses of the monkey G and the monkey H, respectively. CORs are shown separately for reward trials (red) and nonreward trials (blue), while the two numbers in each color represent mean and median of the corresponding histogram distribution. The number of samples corresponds to the number of neurons multiplied by the number of blocks.

in the early stage (Figure 3B) was rather low, possibly because the monkey was making a larger counting error. This issue, however, remains to be examined further and is beyond the scope of the present work.

So far, we have shown that the PRN effect of DA responses can be simulated by the contextual TD model, not by the conventional TD model. This effect is examined with DA response averaged with respect to each PRN. Can the contextual TD errors match DA responses in individual trials? Figure 5A shows the actual and simulated DA responses for one block of trials. The black line shows the cue responses of a DA neuron in individual trials of the block. Superimposed are the simulated cue responses of a DA neuron (i.e., TD error) predicted by the contextual TD model (green line). For comparison, the simulated response by the conventional TD model is shown in Figure 5C for the same block of trials. It appears that DA responses are better predicted by the contextual TD model than the conventional TD model.

To quantify this observation, we replotted the DA responses and the TD errors and computed the correlation coefficient (COR) between them (Figures 5B and 5D). To avoid pseudocorrelation, the COR was computed separately for the rewarded and unrewarded trials. The contextual TD model had higher COR values in both rewarded and unrewarded trials than did the conventional TD model. The results for all DA neurons are summarized in the histograms, separately for the two monkeys (Figures 5E–5H). COR values tended to be positive with the contextual TD model, but negative with the conventional TD model. We performed two statistical tests: (1) the mean was examined against the null hypothesis of being zero by t test with $p < 0.05$; and (2) the median was examined against the same null hypothesis by a nonparametric binomial test with $p < 0.01$. We found that the mean and median COR values were statistically significantly positive in monkey H for both rewarded and unrewarded trials. In monkey G, the COR values were significantly positive for rewarded trials, but not for unrewarded trials. In contrast, the conventional TD model yielded significantly negative COR values in both monkeys for both rewarded and unrewarded trials.

These results together suggest that in order to provide a prediction error, DA neurons take into account the probability of reward conditional to PRN, in the same way the contextual TD model does. One may wonder how this contextual TD model behaves in the noncontextual task. There is no benefit in considering the conditional probability for reward prediction in the noncontextual task because Pr[reward] = Pr[reward|PRN]. Then, the contextual TD model behaves similarly to the conventional TD model in the noncontextual task.

Discussion

Our results suggest that midbrain DA neurons change their activity depending on the presence and absence of a context of reward delivery. In particular, we showed that DA neurons could represent a context-dependent reward prediction error. In the noncontextual task, reward was delivered probabilistically in each trial regardless of the preceding trials, and therefore there was no context that would have improved the reward prediction. In this case, DA responses corresponded to a reward prediction error that was based on the (unconditional) reward probability, i.e., Pr[reward]. Specifically, they behaved similarly to the TD error given by the conventional TD model that only takes into account the current sensory input for reward prediction. The positive slope of DA response appeared in relation to the number of trials since the last rewarded trial (postreward trial number, PRN), as shown in Figure 1B. This may reflect the local nature of TD learning.

On the other hand, reward prediction can be improved if a relevant context exists and is taken into account. In the contextual task, the probability of reward was the same but the probability of reward conditional to the preceding trials changed. After sufficient experience in conducting the task, DA responses represented prediction error better than that predicted by the (unconditional) reward probability. This feature was represented as the negative slope of DA responses in relation to PRN (Figures 3C and 3D). Based on these findings, we proposed a contextual TD model that uses both sensory information and context information to improve reward prediction. We have shown that the contextual TD model, not the conventional TD model, successfully simulated DA responses in the contextual task. This suggests that DA neurons represented reward prediction error based on the conditional reward probability, i.e., Pr[reward|PRN], as the contextual TD model did.

It is worth noting that PRN was not explicitly indicated as a relevant context for reward prediction anywhere in the contextual task; it was a hidden variable and needed to be discovered. It was not a local context that could be maintained only within a trial, such as the internal states (Berns and Sejnowski, 1998; Montague et al., 1996; Schultz et al., 1997; Suri, 2001; Suri and Schultz, 2001), but a global context because it had to be counted and maintained over trials. Furthermore, PRN had to be maintained "dynamically" over trials; it had to be reset after each rewarded trial. Importantly, the negative slope of the PRN effect was obtained from monkeys that had been trained extensively on the task (Figures 3C and 3D). The same monkey showed the PRN effect with no clear slope when it was less well trained (Figure 3B). These results suggest that the context was acquired over the course of the monkey's experience in this particular task. Studies in literature on machine learning (Sutton, 1991; Sutton and Barto, 1998) suggested that the internal model of the environment and the hidden variables accelerates the learning speed in reinforcement learning and allows the mechanism to deal flexibly with changes in the environment. Our results provide experimental support for this notion; the basal ganglia circuit may use information given by an internal model of the task, or the context, to predict reward and thus produce reward prediction error.

The contextual TD model is a specific modification of the conventional TD model that directly used a specific context (PRN) as input. It is possible that other modifications, or different form of contexts, may produce the same prediction error. In other words, it remains to be investigated how DA neurons become able to produce a context-dependent prediction error in the contextual task. From a broader perspective, it remains to be investigated what form of context information DA neurons can use in what kind of contextual task.

We based our simulation on the hypothesis that the caudate (CD) neurons work as value function. An intriguing question is whether the CD neurons behave as predicted by the contextual TD model or by the conventional TD model in the contextual task. Our previous studies (Lauwereyns et al., 2002a, 2002b; Takikawa et al., 2002a) showed that many CD neurons increased their activity before a cue came on. Our preliminary results suggest that these CD neurons behave as the value function of the contextual TD model (H. Itoh et al., 2002, Soc. Neurosci., abstract). Contextual information may originate from brain areas outside the basal ganglia, especially the frontal and parietal cortical areas (Coe et al., 2002; Matsumoto et al., 2003; Schall et al., 2002; Tanji, 2001). For example, the dorsolateral frontal cortex (Kobayashi et al., 2002; Leon and Shadlen, 1999; Watanabe, 1996), the presupplementary motor area (Nakamura et al., 1998; Shima et al., 1996), the supplementary eye field (Lu et al., 2002; Schlag-Rey et al., 1997; Stuphorn et al., 2000), the anterior cingulate cortex (Procyk et al., 2000; Shidara and Richmond, 2002; Shima and Tanji, 1998), and the parietal area LIP (Platt and Glimcher, 1999) contain neurons that appear to encode reward-related contexts or sensorimotor contexts. It will be very important to study how these areas work synergistically to create a memory of reward context.

Experimental Procedures

Experiments

We used two male Japanese monkeys (*Macaca fuscata*; monkeys G and H). The monkeys were kept in individual primate cages in an air-conditioned room where food was always available. At the beginning of each experimental session, they were moved to the experiment room in a primate chair. The monkeys were given restricted amounts of fluid during periods of training and recording. Their body weight and appetite were checked daily. Supplementary water and fruit were provided daily. All surgical and experimental protocols were approved by the Juntendo University Animal Care and Use Committee and were in accordance with the NIH Guide for the Care and Use of Animals.

Noncontextual Task: Classical Conditioning Task

In each trial a spot of light (duration, 150 ms) was presented at the center of the screen. After 500 ms, either a reward (drop of water) together with a tone was delivered (rewarded trial) or only the tone was delivered (nonrewarded trial). The rewarded and nonrewarded trials were chosen randomly at each trial with 50% probability. The intertrial interval was randomized (3.5–10 s). The monkey was not required to fixate or make eye movements. Data were obtained from monkey G.

Contextual Task: One-Direction Rewarded Task, i.e., 1DR Task

In each trial a spot of light was presented at the center of the screen. The monkey was required to keep fixating the spot. After 1 s, another spot of light (cue stimulus) was presented at one of four positions in four cardinal directions (eccentricity is fixed as either 10° or 20°). The cue position was chosen pseudorandomly (see below). The monkey had to remember the position of the cue while fixating the central spot. After 1–1.5 s, the fixation spot went off and the monkey had to make a saccadic eye movement to the remembered position. If the saccade was correct (i.e., if it landed within ±3° from the cue position), a reward (drop of water) together with a tone was delivered (rewarded trial), or only the tone was delivered (nonrewarded trial). If the saccade was incorrect, no tone was presented and the same trial was repeated. The next trial started after an intertrial interval of 3.5–4 s.

In Figure 3, we show the data from the early and late stages that express the degree to which the monkey was experienced with this task. Description of the early and late stage was given in the main text (for the monkey G). Data of the monkey (H) in the late stage is collected after the monkey (H) had had more than one and a half years of experience with the task, which corresponded to well over 600 blocks. All neurons recorded in the corresponding stage were used to show population data (Figures 3 and 5).

The error rate is computed as (number of error trials)/(number of successful trials + number of error trials), where error trials includes both "fixation break" error trials, which occur when the monkeys break fixation too early, and "incorrect saccade" error trials, which occur when the monkeys make saccade toward a wrong direction.

Recording Procedures

Eye movements were recorded using the search coil method (Enzanshi Kogyo MEL-20U) (Judge et al., 1980; Robinson, 1963). Eye positions were sampled at 500 Hz. Single unit recordings were performed using tungsten electrodes (diameter, 0.25 mm, 1–5 MΩ, measured at 1 KHz, Frederick Haer), which were driven by a hydraulic microdrive (Narishige, MO95-S). The behavioral tasks as well as storage and display of data were controlled by a computer (PC 9801RA, NEC, Tokyo). The unitary action potentials were passed through a window discriminator (Bak INC, Model DDIS-1), and the times of their occurrences were stored with a resolution of 1 ms.

Before the single unit recording experiment, we determined the recording sites in the substantia nigra obtained MR images (Hitachi, AIRIS, 0.3T). DA neurons were identified by their irregular and tonic firing with broad spike potentials. Extracellular spikes may have an initial positive component or may be followed by a prolonged positive component (Schultz and Romo, 1987). A neuron with these features was thus determined to be a DA neuron candidate (Kawagoe et al., 2003). Near the end of a long-term experimental session, we made electrolytic microlesions at the recording sites of DA neurons for later histological analysis. Later histological examination showed that the presumed DA neurons were located in the substantia nigra pars compacta (SNc) (A9) and sometimes in the area mediodorsal to the SNc (A8).

Pseudorandomization and the Probability of Reward Conditional to the PRN

The cue position was chosen by a pseudorandom schedule: within each subblock of four trials, each of all four directions was chosen randomly but always once. In result, all directions were chosen equally often in a block; the probability of a rewarded trial "R+" was 0.25 and the probability of an unrewarded trial "R−" was 0.75. This pseudorandom schedule introduced a specific probabilistic structure of the occurrence of a rewarded trial. A possible determinant of the structure was the number of trials that occurred between the last rewarded trial and the current trial, called the postreward trial number (PRN). Denoting the PRN by K, the probability conditional to PRN is given by Pr[R + |K] and Pr[R − |K]. Since Pr[R + |K] + Pr[R − |K] = 1, it suffices to compute either probability, say Pr[R + |K]. It is also sufficient to consider two consecutive subblocks to get Pr[R + |K] (in the second subblock). We have Pr[R + |K] = N_+(K)/N(K) by simply counting all possible cases for each K, denoted by N(K), and the rewarded trials for each K, denoted by N_+(K).

Each block of the experiment "formally" consists of 60 trials; however, in actual experiments, we often recorded a few more than 60 trials in a block so that the block may end at any element (from first to fourth) of the subblock (e.g., third). The first trial of the next block starts with the next trial of the subblock (e.g., fourth of the subblock). Thus, an experimental block could begin with any element of the subblock.

Models

Reinforcement Learning and the Conventional and Contextual TD Models

We simulated two types of reinforcement learning models based on temporal difference (TD) learning, using actor-critic architecture (Sutton and Barto, 1998). In TD learning, the "critic" provides the estimated reward values in each "state," which is an input to the critic, while the "actor" provides the estimated optimal action in each state. The critic estimates the expected reward value of a state, taking into account rewards obtained in the future with a discount factor. Let us denote the function of the critic by $V(x_t)$, called the value function. We used x_t to represent a state at time t. Below, we sometimes drop the subscript t when no confusion is expected (we dropped it in the main text). A state x is different between the conventional and contextual TD models. The conventional TD model uses only the current sensory input s as input and thus has $x = s$. In contrast, the contextual TD model takes as input both the current sensory input s and the context input c, which is the PRN, and thus has $x = (s, c)$.

The value function is defined by

$$V(x_t) = E\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right), \tag{4}$$

where γ is the discount factor, bounded by $0 \leq \gamma \leq 1$, E denotes taking the expectation, and r_{t+k} denotes the reward given by the state transition from the state x_{t+k} to the next state x_{t+k+1}. Different values of the discount factor give different weighting on rewards given at different times in the future.

This form of the value function conveniently gives the constraint $\gamma V(x_{t+1}) + r_t - V(x_t) = 0$. This constraint is local in that it is represented only by the variables in a state transition; one state, x_t, to the next state, x_{t+1}, with the associated reward r_t. When and if this constraint is violated, their discrepancy gives a clue for learning. Hence, TD error is defined by TD $= \gamma V(x_{t+1}) + r_t - V(x_t)$. Because TD error is local, TD learning can be performed locally in each state transition.

In Equation 4, the left-hand side $V(x_t)$ is the function of x, whereas the right-hand side is the function of rewards. Therefore, we must specify the form of function in terms of x in order to represent the expected reward (in the right-hand side). We set the form as the weighted sum of input,

$$V(x_t) = \sum_j w_j x_{tj},$$

where the summation is taken over all the dimensions of the input. Using TD error, the value function is updated in each state transition (as in Equation 2). With the form

$$V(x_t) = \sum_j w_j x_{tj},$$

the new weight is given by

$$w_j^{new} = w_j^{old} + \alpha \times \mathrm{TD} \times x_{tj},$$

where α is the learning constant small enough to prevent a perturbation. The actor is implemented by using a softmax function and learned similarly, except that its learning takes into account which action was taken at the state (Sutton and Barto, 1998).

Simulations of the Noncontextual Task

Both simulations of the noncontextual and contextual tasks used discrete time for simplicity, so that each event in each experiment was used as the current sensory input. In the noncontextual task simulation, the two states were treated as the inputs, namely, the prereward and the postreward. A binary representation was used so that the sensory input had a two-dimensional binary representation. There was no action for the model to select. In each trial, a cue indicated the transition to the prereward from the postreward (of the previous trial) and a reward delivery ($r = 1.0$) or nondelivery ($r = 0.0$), with 50% probability, indicated the transition from the prereward to the postreward.

To plot the simulation result (as in Figure 1B), we had to determine the values of two free parameters of the model, namely the learning constant α and the discount factor γ (the initial weights for the value function were randomly chosen around zero), and also had to determine the unit scale of the figure. As an exhaustive search, we ran simulations with different values of the two parameters. With each simulation result (TD errors), we tested a linear regression ($y = a \cdot x + b$), where x and y indicate the TD error and DA activity for different PRNs, respectively, and obtained the residual and the values of the coefficients. We chose the values of two parameters ($\alpha = 0.3$ and $\gamma = 0.9$) that gave the least residual, or equivalently gave the maximal correlation between x and y. The corresponding value of the coefficient a was used to determine the unit scale in Figure 1D, since the unit scale of simulation is essentially arbitrary (indicated as a.u., arbitrary unit, in the figures). We did not use the value of another coefficient b in determining the unit scaling, because we wanted to keep the origin of the TD error unchanged and also because the value of b was nearly zero (0.148).

Simulations of the Contextual Task

For the simulation of the contextual task, one trial consisted of three states (the precue, postcue, and fixation-off states). At any state transition, the model could take actions, i.e., make a saccade in one of the four directions or make no movement. Each trial started with the precue state, and the cue flash (Figure 2A) caused the transition from the precue to the postcue state. The fixation-off let the state change into the fixation-off state. In the fixation-off state, the model had to make a saccade correctly and then the next trial started. If the saccade was made in a wrong direction (or was not made at all) in the fixation-off state, or if the saccade was made at any other states, the trial was truncated and repeated again with the same target direction. The reward was given to the model for the correct saccade only in the rewarded trial; otherwise no reward was obtained.

In this task, the sensory input s to the model consisted of the information of the current state (as one of the precue, postcue, and fixation-off states), of the target direction of a trial (right, up, left, down, or unknown), and of the rewarded direction of the block (right, up, left, down, or unknown). A binary, or table look-up, representation of these inputs was used and set appropriately at each state; for example, the target direction was set as “unknown” at the precue state. The rewarded direction was set properly after a few trials in each block, when there was enough information to accurately judge the rewarded direction of the block (as observed in experiments). The context input c was the PRN (from 1 to 7), represented by a 7-dimensional binary input.

The unit scale is arbitrary in simulations. In order to determine the scale in Figures 4C and 4D, we used the minimum square loss with the equation $y = a \cdot x$, where x and y indicate the TD error and DA activity for different PRNs, respectively. The value of the coefficient a was determined and used as the scale.

In the contextual task simulation with the contextual TD model, we introduced the counting error of PRN, which is an error of incrementing PRN between trials. For example, 10% error means that incrementing PRN correctly between trials, from k (say, 2) to $k + 1$ (i.e., 3), fails with 10% probability, resulting in keeping $\text{PRN} = k$ incorrectly. In the simulated results shown (Figures 4 and 5), the counting error is set as 40%. See the Supplemental Data at http://www.neuron.org/cgi/content/full/41/2/269/DC1 for more information.

Simulated results of the PRN dependency of TD errors (Figures 4C and 4D) are only from the network that had already been well trained (that is, it could perform the task with a low error probability and the behavior of TD errors was stabilized). Since TD errors are averaged over many trials, the variance is negligible. For the trial-by-trial analysis in Figure 5, we used the well-trained network but with the identical sequence of trials occurred in the experiment. To plot the variability of TD errors (Figures 5A and 5C), we prepared ten well-trained networks, trained independently, to run ten simulations, and their means and standard errors were used. The mean was used to compute CORs. For the unit scaling in Figures 5A and 5C, the same procedure in Figures 4C and 4D was applied to the data of this experiment block. Note that COR histogram (Figures 5E–5G) remains unaffected by this unit-scaling choice because COR does not change by changing the axis scale linearly. In calculating the COR values, the first eight trials of a block were discarded to avoid any possible pseudocorrelation due to the transition between blocks. This is because DA neurons changed their responses in a very early part of a block and presumably re-acquired the information of the new rewarded cue direction in a new block (Kawagoe et al., 2003).

Acknowledgments

H.N. is supported by Grant-in-Aid for Scientific Research on Priority Areas (C) and for Young Scientists (B), and the Fund from the US-Japan Brain Research Cooperative Program of MEXT, Japan. We thank J. Lauwereyns, K. Namakura, and K. Sakai for their comments on the draft and reviewers for their comments. H.I. and H.N. thank K. Aihara for his support.

Received: May 22, 2003
Revised: November 25, 2003
Accepted: December 22, 2003
Published: January 21, 2004

References

Arbib, M.A., and Dominey, P.F. (1995). Modeling the roles of basal ganglia in timing and sequencing saccadic eye movements. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J.L. Davis, and D.G. Beiser, eds. (Cambridge, MA: MIT Press), pp. 149–162.

Barto, A.G. (1995). Adaptive critics and the basal ganglia. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J.L. Davis, and D.G. Beiser, eds. (Cambridge, MA: MIT Press), pp. 215–232.

Berns, G.S., and Sejnowski, T.J. (1998). A computational model of how the basal ganglia produce sequences. J. Cogn. Neurosci. *10*, 108–121.

Coe, B., Tomihara, K., Matsuzawa, M., and Hikosaka, O. (2002). Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J. Neurosci. *22*, 5081–5090.

Dayan, P. (2002). Motivated reinforcement learning. NIPS *14*, 11–18.

Doya, K. (2002). Metalearning and neuromodulation. Neural Netw. *15*, 495–506.

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

Graybiel, A.M., Aosaki, T., Flaherty, A.W., and Kimura, M. (1994). The basal ganglia and adaptive motor control. Science *265*, 1826–1831.

Hollerman, J.R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. *1*, 304–309.

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

Judge, S.J., Richmond, B.J., and Chu, F.C. (1980). Implantation of magnetic search coils for measurement of eye position: an improved method. Vision Res. *20*, 535–538.

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

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2003). Reward-predicting activity of dopamine and caudate neurons — a possible mechanism of motivational control of saccadic eye movement. J. Neurophysiol., in press. Published online October 1, 2003. 10.1152/jn.00721.2003.

Kobayashi, S., Lauwereyns, J., Koizumi, M., Sakagami, M., and Hikosaka, O. (2002). Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. J. Neurophysiol. *87*, 1488–1498.

Lauwereyns, J., Takikawa, Y., Kawagoe, R., Kobayashi, S., Koizumi, M., Coe, B., Sakagami, M., and Hikosaka, O. (2002a). Feature-based anticipation of cues that predict reward in monkey caudate nucleus. Neuron *33*, 463–473.

Lauwereyns, J., Watanabe, K., Coe, B., and Hikosaka, O. (2002b). A neural correlate of response bias in monkey caudate nucleus. Nature *418*, 413–417.

Leon, M.I., and Shadlen, M.N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron *24*, 415–425.

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

Lu, X., Matsuzawa, M., and Hikosaka, O. (2002). A neural correlate of oculomotor sequences in supplementary eye field. Neuron *34*, 317–325.

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science *301*, 229–232.

Meck, W.H. (1996). Neuropharmacology of timing and time perception. Brain Res. Cogn. Brain Res. *3*, 227–242.

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. *72*, 1024–1027.

Montague, P.R., Dayan, P., Person, C., and Sejnowski, T.J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. Nature *377*, 725–728.

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

Nakahara, H., Doya, K., and Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences—a computational approach. J. Cogn. Neurosci. *13*, 626–647.

Nakamura, K., Sakai, K., and Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. J. Neurophysiol. *80*, 2671–2687.

Nieder, A., and Miller, E.K. (2003). Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. Neuron *37*, 149–157.

Parent, A., and Hazrati, L.-N. (1994). Multiple striatal representation in primate substantia nigra. J. Comp. Neurol. *344*, 305–320.

Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.

Procyk, E., Tanaka, Y.L., and Joseph, J.P. (2000). Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat. Neurosci. *3*, 502–508.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning: Current Research and Theory, A.H. Black and W.F. Prokasy, eds. (New York: Applenton Century Crofts), pp. 64–99.

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

Robinson, D.A. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. IEEE Trans. Biomed. Eng. *10*, 137–145.

Romo, R., and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J. Neurophysiol. *63*, 592–606.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. J. Neurosci. *23*, 9913–9923.

Schall, J.D., Stuphorn, V., and Brown, J.W. (2002). Monitoring and control of action by the frontal lobes. Neuron *36*, 309–322.

Schlag-Rey, M., Amador, N., Sanchez, H., and Schlag, J. (1997). Antisaccade performance predicted by neuronal activity in the supplementary eye field. Nature *390*, 398–401.

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

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

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J.R., and Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J.L. Davis, and D.G. Beiser, eds. (Cambridge, MA: MIT Press), pp. 233–248.

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

Shidara, M., and Richmond, B.J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science *296*, 1709–1711.

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

Shima, K., Mushiake, H., Saito, N., and Tanji, J. (1996). Role for cells in the presupplementary motor area in updating motor plans. Proc. Natl. Acad. Sci. USA *93*, 8694–8698.

Stuphorn, V., Taylor, T.L., and Schall, J.D. (2000). Performance monitoring by the supplementary eye field. Nature *408*, 857–860.

Suri, R.E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. Exp. Brain Res. *140*, 234–240.

Suri, R.E., and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp. Brain Res. *121*, 350–354.

Suri, R.E., and Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. Neural Comput. *13*, 841–862.

Sutton, R.S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. SIGART Bulletin *2*, 160–163.

Sutton, R.S., and Barto, A.G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psychol. Rev. *88*, 135–170.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, MA: A Bradford Book).

Takikawa, Y., Kawagoe, R., and Hikosaka, O. (2002a). Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. J. Neurophysiol. *87*, 508–515.

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

Tanji, J. (2001). Sequential organization of multiple movements: involvement of cortical motor areas. Annu. Rev. Neurosci. *24*, 631–651.

Tesauro, G.J. (1994). TD-Gammon, a self-teaching backgammon program, achieves masterlevel play. Neural Comput. *6*, 215–219.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. Nature *412*, 43–48.

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