Dopamine neurons learn to encode the long-term value of multiple future rewards

Kazuki Enomoto[a,b,1], Naoyuki Matsumoto[a,c,1], Sadamu Nakai[a,d], Takemasa Satoh[a,e], Tatsuo K. Sato[a,f], Yasumasa Ueda[a], Hitoshi Inokawa[a], Masahiko Haruno[b,g], and Minoru Kimura[a,b,2]

[a]Department of Physiology, Kyoto Prefectural University of Medicine, Kyoto 602-8566, Japan; [b]Brain Science Institute, Tamagawa University, Tokyo 194-8610, Japan; [c]Faculty of Environmental and Symbiotic Sciences, Prefectural University of Kumamoto, Kumamoto 862-8502, Japan; [d]Faculty of Industrial Science and Technology, Tokyo University of Science, Yamakoshi 049-3514, Japan; [e]Division of Neurobiology, School of Life Sciences, Faculty of Medicine, Tottori University, Yonago 683-8503, Japan; [f]Department of Physiology, Institute of Ophthalmology, University College London, London EC1V 9EL, United Kingdom; and [g]Center for Information and Neural Networks, National Institute of Information and Communications Technology, Tokyo 184-8795, Japan

Edited by Ranulfo Romo, Universidad Nacional Autonoma de Mexico, Mexico City, D.F., Mexico, and approved August 5, 2011 (received for review October 6, 2010)

Midbrain dopamine neurons signal reward value, their prediction error, and the salience of events. If they play a critical role in achieving specific distant goals, long-term future rewards should also be encoded as suggested in reinforcement learning theories. Here, we address this experimentally untested issue. We recorded 185 dopamine neurons in three monkeys that performed a multistep choice task in which they explored a reward target among alternatives and then exploited that knowledge to receive one or two additional rewards by choosing the same target in a set of subsequent trials. An analysis of anticipatory licking for reward water indicated that the monkeys did not anticipate an immediately expected reward in individual trials; rather, they anticipated the sum of immediate and multiple future rewards. In accordance with this behavioral observation, the dopamine responses to the start cues and reinforcer beeps reflected the expected values of the multiple future rewards and their errors, respectively. More specifically, when monkeys learned the multistep choice task over the course of several weeks, the responses of dopamine neurons encoded the sum of the immediate and expected multiple future rewards. The dopamine responses were quantitatively predicted by theoretical descriptions of the value function with time discounting in reinforcement learning. These findings demonstrate that dopamine neurons learn to encode the long-term value of multiple future rewards with distant rewards discounted.

decision making | basal ganglia | temporal difference learning | primate

Suppose you try to win a tennis match. If you are an experienced player, you may plan long-term tactics for six-game sets that maximize your concentration and effort in the most critical games, usually in the middle of the set, to maintain an advantage over your opponent and reserve your resources for the other games of the set and match. On the contrary, if you are a beginner, you will probably just concentrate on winning each single game and will be exhausted halfway through the match. Therefore, assigning long-term reward values for individual actions is a learned intelligence for the successful achievement of distant goals.

Reinforcement learning theories propose an algorithm for subjects to learn to take actions that are most likely to yield maximum amount of total future rewards (1). For long-term judgment, a "value" is assigned to the current state of the subjects as an expected total future rewards, where kth future reward is discounted by γ^k. Here, γ is a discount factor between 0 and 1 that weights the relative contribution of future rewards to value. The values of the current and the subsequent states are linked through temporal difference (TD) error, in such a manner as follows:

$$\text{Current TD error} = \text{current reward} + \gamma \cdot \text{value of next state} - \text{value of current state} \quad [1]$$

Dopamine neurons convey the reward value signals of external events (2–4), their expectation errors (2, 3, 5–8), and signals of the motivational salience of external stimuli (5, 9, 10). The discounted value of a single reward that is expected after some delay is represented by neuronal activity in the cerebral cortex (11–13) and dopamine neurons (14, 15) in rodents and primates, and, as shown with functional brain imaging, in the cerebral cortex and striatum in humans (16, 17). It is also reported that the activity of target neurons of dopamine signals in the striatum represent the reward values of action options (18, 19) and chosen actions (19, 20) during behavioral tasks in which experimental animals learn to choose options with a higher reward probability on a trial-and-error basis and to keep choosing the option even if those choices sometimes lead to no reward. Therefore, an important question is the extent to which the dopamine neurons represent the TD error of multistep choices for rewards. Previous studies examined dopamine neuron activity by using classical and instrumental conditioning tasks for a single reward (2, 3, 8, 10, 21, 22). Some studies used multistep choice paradigms in which the subject's estimated reward is based on the histories of actions and outcomes (4–7, 15, 23, 24). However, none of these studies have examined whether and how dopamine neurons encode the TD error signals in multistep tasks for multiple future rewards.

To address this issue, we recorded the dopamine neuron activity of monkeys performing a multistep choice task as a model of the achievement of a distant goal in the natural environment (Fig. 1*A*). The monkeys first explored a set of three targets to find the rewarding one and then exploited this knowledge to receive one (i.e., two-step choice task) or two (i.e., three-step choice task) additional rewards by choosing the same target (Fig. 1*B*). During the exploration trials, the average reward probability (correct choice rate) increased from approximately 20% for the first choice (N1) to 50% for the second choice (N2) and to approximately 80% for the third choice (N3). During the exploitation trials (R1 and R2), the probability was almost 100% because the monkeys simply repeated the last rewarded choices (Fig. 1*C*). Thus, the monkeys obtained a total of two or three rewards by searching for and choosing a single target in a set of several trials. The next set was restarted following an interposed resetting signal (*SI Text*). We recorded dopamine neuron activity before and after monkeys learned the multistep choice task over several weeks, and examined whether the dopamine responses represent TD errors through learning.

Author contributions: K.E., N.M., S.N., T.S., T.K.S., and M.K. designed research; K.E., N.M., S.N., T.S., and T.K.S. performed research; K.E., N.M., S.N., T.S., T.K.S., and M.H. analyzed data; and K.E., N.M., Y.U., H.I., M.H., and M.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]K.E. and N.M. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: mkimura@lab.tamagawa.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1014457108/-/DCSupplemental.

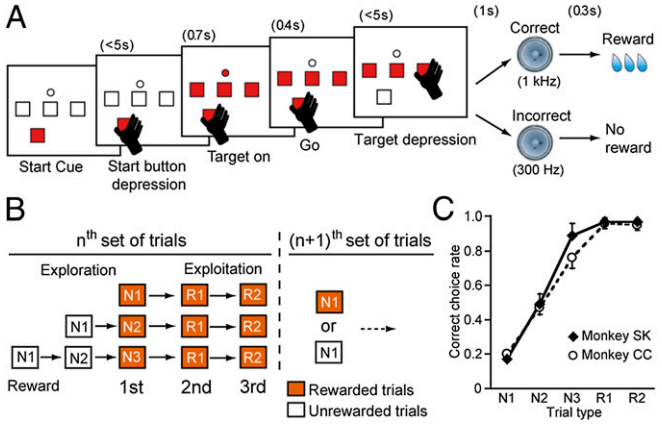

Fig. 1. Behavioral paradigms of multistep actions for rewards in monkeys. (*A*) Sequence of events during the multistep choice task. (*B*) Schematically illustrated structure of the three-step choice trials to obtain three rewards at different times. (*C*) Average correct choice rates (mean and SD, 29 d in monkey SK and 35 d in monkey CC, during the advanced stage of learning) against five types of three-step choice trials (N1, N2, N3, R1, and R2) in two monkeys.

Results

Reward Expectation During Multistep Choices for Rewards. To examine the monkey's reward expectation during individual trials of the multistep choice task, we recorded the anticipatory licking that preceded the reinforcer beeps for reward (Fig. 2*A*) as a behavioral measure in monkeys BT and CC (2, 10, 22). A previous study in our laboratory showed that the start cue responses of dopamine neurons vary in parallel with the latency of the monkey's start cue-evoked behavioral responses (5). The start cue reaction times in the present study, however, did not reflect the expected multiple future values, and the reaction times during the first choice (N1) were longer than the other trial types. Instead, during the exploration trials, the monkeys made longer and more frequent anticipatory lickings when the reward probability of the trials became higher (N1, 30%; N2, 48%; and N3, 79% in monkey BT; N1, 20%; N2, 47%; and N3, 75% in monkey CC). If the monkeys chose a target followed by no reward in the N2 trials, they might be aware that the remaining one would be a rewarding target. In this sense, the N3 trials were different from the other exploratory N1 and N2 trials. However, the monkeys chose a rewarding target in the N3 trials at much lower probabilities (75–90%) than 100%. Thus, we categorized the N3 trials as exploration trials.

Remarkably, anticipatory licks occurred for a shorter time and less frequently during the exploitation trials (R1 and R2) than during exploration trials (N1, N2, and N3), even though the reward probability was almost 100% (Tukey–Kramer test, $P < 0.05$ for N1 vs. N3 trials; $P < 0.05$ for N3 vs. R1 trials in monkey BT; $P < 0.01$ in all cases except between trials N1 and R2, N2 and N3, and R1 and R2 in monkey CC; Fig. 2 *B* and *C*). Therefore, there was a large discrepancy between the magnitude of anticipatory licking and the probability of immediate reward (Fig. 2*C*). Why did anticipatory lickings decrease during the exploitation trials? One possibility was that the monkeys expected not only the immediate reward but also the distant rewards to be obtained within the next few trials. Monkey BT may have expected two rewards, and monkey CC may have expected three rewards from the beginning of the exploration trials, an immediate reward during the N1, N2, or N3 trials, and distant rewards during the exploitation trials. When the first reward had been obtained in an exploration trial, the monkeys would expect the one or two rewards that remained in the exploitation trials.

To test this possibility parametrically, we estimated the average duration of anticipatory licking by using the "value function" of reinforcement learning theories (1, 8, 21), which defines the value of the current state as the sum of the expected future rewards discounted by the number of trials required to obtain them (*SI Text*). The discount factors were estimated to maximize the correlation coefficient (R) between the simulated value function and the normalized licking duration (Fig. 2*D*). We used second derivatives of R to examine how quickly the fit decreases around the γ-value that gave the best fit (*SI Text*). The black superimposed lines in Fig. 2*C* show the estimated value functions, which accurately approximated the normalized average durations of anticipatory licking in both monkeys ($\gamma = 0.65$ for monkey BT, $\gamma = 0.66$ for monkey CC). These results suggested that the monkeys made individual choices while expecting the sum of immediate and future rewards rather than expecting the immediately available rewards alone (other possibilities are detailed in *SI Text*).

Dopamine Neurons Encode Long-Term Reward Value as Expected Sum of Future Rewards. We recorded a total of 185 dopamine neurons in the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) of the three monkeys (Fig. S1). The activities of 51 of these neurons were examined under three-step choices

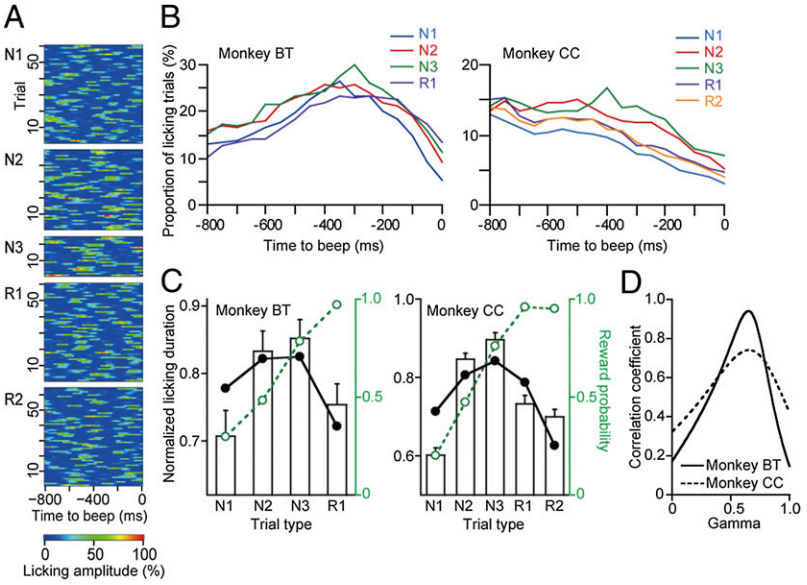

Fig. 2. Reward expectation during multistep actions measured by anticipatory licking. (*A*) The anticipatory licking movements for the 800-ms period before the reinforcer beeps (*SI Text*) in monkey CC are color-coded. (*B*) The average proportion of trials in which the amplitude of anticipatory licking exceeded the threshold (50% maximum) is plotted against the time to the reinforcer beeps in the two monkeys. (*C*) Bar graphs of the normalized licking duration (100–800 ms period before the beeps, mean and SEM; 32 sessions in monkey BT and 75 sessions in monkey CC; *SI Text*) against trial type. The average reward probability (dashed green line) and the best-fit value function derived from reinforcement learning algorithm (solid black line, $\gamma = 0.65$, $R = 0.71$, $P = 0.29$ in monkey BT; $\gamma = 0.66$, $R = 0.74$, $P = 0.16$ in monkey CC) are superimposed. (*D*) The parameter space landscape of correlation coefficients between the experimental and simulated licking duration in which R is plotted against γ. The values of the second derivatives of R are −27 for monkey BT and −6.1 for monkey CC.

for three rewards in monkeys SK and CC (Table S1). The neurons responded to the start cue of individual trials with brief increases in discharges above the baseline rate of four to five spikes per second, as shown in the activity of an example neuron presented in Fig. 3*A*. Average responses of the 25 dopamine neurons in monkey CC and 26 neurons in monkey SK became gradually larger from the N1 to N2 and N3 trials for the first immediate reward, in parallel with the increase in the reward probability of the trial (Fig. 3 *B* and *C*). In contrast, responses during the R1 and R2 trials for the second and third rewards were much smaller than the responses that would have reflected the high reward probabilities during these trials. Responses in the N3 trials were significantly greater than the responses in all the other trials (Tukey–Kramer test, $P < 0.01$ in monkey CC and $P < 0.05$ in monkey SK), which made an inverted "V" shape in the reward probability–dopamine response plot. Because anticipatory licking during the multistep choice task suggested that the monkeys may have anticipated the sum of the immediate and the future rewards from the beginning of a series of choices (Fig. 2), the dopamine responses may have reflected this anticipation.

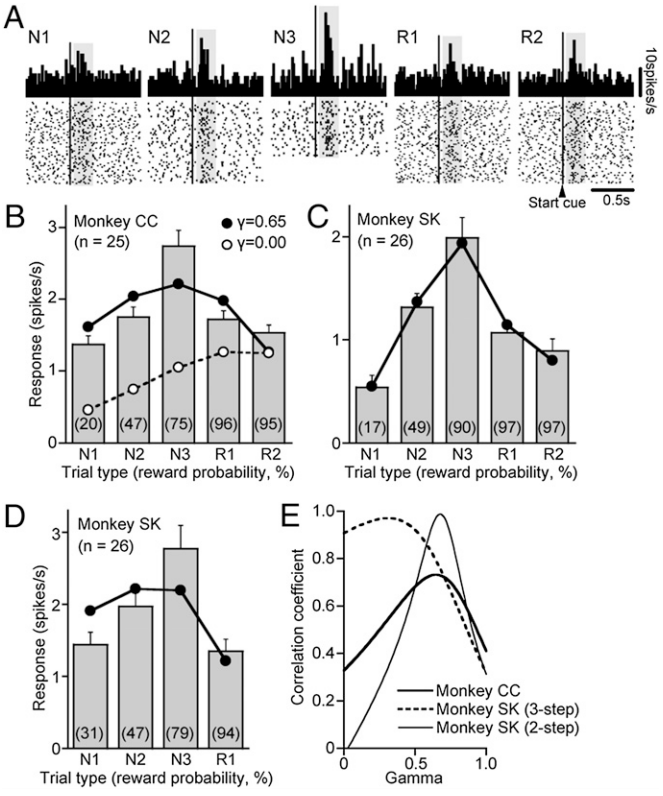

Fig. 3. Dopamine neurons encode long-term value as a sum of expected future rewards. (*A*) Example responses of a dopamine neuron to the illumination of the start cues in individual trials of the three-step choice task in monkey CC. The bin size of the spike density histogram is 15 ms. Hatched areas are the time windows for the analyses shown in *B*. (*B*) Bar graphs of ensemble average of dopamine responses (mean and SEM) above the baseline in monkey CC during the time windows (50–290 ms after the start cue) shown in *A*. The best-fit value functions ($\gamma = 0.65$, $R = 0.71$, $P = 0.18$, solid line) and reward probability of trials ($\gamma = 0.00$, $R = 0.29$, $P = 0.68$, dashed line) are superimposed. The numbers in parentheses represent the reward probability for the given trial type. (*C*) Same as in *B* but for monkey SK (40–240 ms after the start cue). The best-fit value function ($\gamma = 0.31$, $R = 0.99$, $P < 0.01$) is superimposed (Fig. S2*B*). (*D*) Same as in *B* but for monkey SK (70–260 ms after the start cue) in the two-step choice task with a fixed amount of reward. The best-fit value function is superimposed ($\gamma = 0.68$, $R = 0.71$, $P = 0.29$). (*E*) Plots of the parameter space landscape of correlation coefficients. The value of second derivative of *R* is −5.8 for monkey CC, −2.2 for monkey SK in the three-step choice task, and −36 for monkey SK in the two-step choice task.

To validate this hypothesis, we examined the magnitude of the dopamine responses by using the value function of reinforcement learning theories as the sum of the expected multiple future rewards (volume × probability) discounted by the number of future steps to obtain them in monkeys CC and SK (Fig. 3 *B*, *C*, and *E*). We also manipulated the number and magnitude of the future rewards to test the value function (i.e., TD) model of dopamine responses. Both monkeys obtained one reward during the exploration trials. During the subsequent exploitation trials, there were two rewards for monkey CC and one or two rewards for monkey SK. When the total number of rewards was reduced from three to two with a fixed amount in monkey SK, the dopamine responses in the R1 trials were still much lower than the responses for the first reward in the N3 trial (Mann–Whitney *U* test, $P < 0.001$; Fig. 3*D*). The magnitudes of dopamine responses were accurately approximated by the estimated value functions during both conditions (Fig. 3 *C* and *D*). The volumes of the three rewards (one during exploration and two during exploitation trials) were fixed for monkey CC (0.35 mL), but the volume of the distant rewards (R1 and R2, 0.2 mL) was smaller than that for the immediate rewards (N1, N2, and N3, 0.35 mL) in monkey SK. Our hypothesis was that the discount factor might be smaller when the volume of distant rewards was reduced. Indeed, the discount factor was smaller when the distant reward was reduced ($\gamma = 0.31$; Fig. 3*C*) than when it was fixed ($\gamma = 0.68$; Fig. 3*D*). In monkey CC, the neuronal discount factor ($\gamma = 0.65$) was almost identical to the behavioral discount factor that was estimated by the anticipatory licking ($\gamma = 0.66$; Fig. 2*C*, *Right*; licking was not measured in monkey SK), which indicated that the dopamine responses may faithfully represent the expectation of long-term multiple rewards. In support, there was a large discrepancy between the dopamine responses and the probability of immediate reward (i.e., value function with γ of 0.00; Fig. 3*B*, dashed line). It is also notable that the correlation coefficients between value function and dopamine neuron firing sharply decreased around the estimated γ-value (Fig. 3*E*), indicating stable and reliable estimation of γ.

In monkeys BT and CC, we examined the dopamine neuron responses to conditioning stimuli (CSs) under a classical conditioning paradigm (Fig. S5*A*) in which the CSs that signaled different reward probabilities appeared in an unpredictable order. This paradigm does not distinguish single-step from cumulative coding because the trials are independent. Both anticipatory licking (Fig. S5*B*) and the magnitude of the dopamine responses to the CSs (Fig. S5 *C* and *D*) faithfully represented the single reward value that was assigned to the stimuli.

Dopamine Neurons "Learn" to Encode Long-Term Reward Expectation During Multistep Choices for Rewards. Although the encoding of the long-term value of a series of actions is a key component process for achieving distant goals on a trial-and-error basis, it is unlikely that dopamine neurons have this ability without experiences. To examine whether the dopamine responses are established through learning, we studied responsiveness of dopamine neurons when monkeys learned the multistep choice paradigm for multiple rewards over a period of several weeks. We recorded the activities of 76 dopamine neurons (51 neurons from monkey SK and 25 neurons from monkey CC) over the course of learning the three-step choices among three alternatives. Before learning, the monkeys had mastered a simple version of the multistep choices: three rewards through three-step choices between two alternatives. Then, they started to learn the upgraded version to obtain the first of three rewards by searching for a rewarding target among three alternatives. The correct choice rate during the N3 trials increased day by day in both monkeys (Fig. 4*A*), but the correct choice rates in the other trials were stable over the course of learning (Fig. S6). The slow change in the correct choice rate during the N3 trials may result from the process of adaptively switching the task strategy. First, the monkeys were required to discard the simple strategy of choosing a different target from the last-tried unrewarded one and choosing the same target as the last-tried rewarded one that had been appropriate for the previous task of three-step choices between two alternatives. Then,

they needed to acquire a new strategy to estimate the value of three options while updating these values based on the previously tried N1 and N2 trials and their outcomes (i.e., reward history), and to choose the highest-value option.

For convenience, the learning process was divided into two stages, an early stage and a later, advanced stage when the correct choice rate in the N3 trials was greater than 80% of the highest stable rate (100% in monkey SK, 84% in monkey CC) for five consecutive days. The early stage comprised days 1 through 22 in monkey SK and days 1 through 26 in monkey CC. The advanced stage comprised days 23 through 51 in monkey SK and days 27 through 61 in monkey CC. The durations of anticipatory licking in the early stage were not significantly different among the five trial types (Tukey–Kramer test, $P > 0.47$ in all cases; Fig. 4*B*, *Left*), but there was a tendency for longer licking durations in the trials with higher reward probabilities (i.e., during the exploitation trials R1 and R2). During the advanced stage, in contrast, the longest anticipatory licking occurred during the N3 trials, and the licking duration was much shorter during all the other trials, especially in the R1 and R2 trials (Tukey–Kramer test, $P < 0.05$ for N3 vs. all other trial types; Fig. 4*B*, *Right*). This indicated that the monkeys had already acquired an explicit idea not only about the probabilities of the first, immediate reward but also about the distant two rewards.

The responses of dopamine neurons evolved over the course of learning. In the early stage, the start cue responses were small in most trials, as shown in the activity of an example neuron recorded on day 12 of learning in monkey SK (Fig. 4*C*, *Left*, and Fig. S2*A*) and the average activity of 25 neurons recorded during the early stage (Fig. 4*E*). Although the responses during the first choice (N1) tended to be small, there were no significant differences among any of the trial types (Tukey–Kramer test, $P > 0.05$). In the advanced stage, dopamine neurons showed strong responses in the N3 trials compared with the responses in the other trial types, which resulted in an inverted V-shape distribution, as shown in the activity of an example neuron recorded on day 26 of learning in monkey SK (Fig. 4*C*, *Right*, and Fig. S2*B*; Fig. 3*C* shows population data). The dopamine responses of two example neurons in the early and advanced stages were accurately approximated by the value functions with small ($\gamma = 0.04$) and large ($\gamma = 0.38$) discount factors, respectively (Fig. 4 *C* and *D*). The average responses of 25 neurons during the early stage were also approximated with a smaller discount factor ($\gamma = 0.00$; Fig. 4*E*) than during the advanced stage ($\gamma = 0.31$; Fig. 3*C*). The dopamine neuron responses during the advanced stage were characterized by larger responses for the first, immediate reward in the N2 and N3 trials compared with the responses for the second and third rewards in the R1 and R2 trials. This result favored "active" mechanisms that have evolved to encode long-term values over "passive" processes (i.e., a lowered enthusiasm for the search for reward or habituation during the exploitation trials). Therefore, the dopamine neurons learned to encode the long-term value of not only the expected immediate reward but also the distant rewards, as the monkeys learned the multistep choice paradigm for rewards.

On the contrary, in the early stage of learning, the monkeys may be uncertain about the nature of the sequence of the trials, the long-term schedule, and where they were in the sequence because of the limited experience. Because these factors reflect an internal state, the monkeys appeared to pay attention to the current choices but paid much less attention to future trials during the early stages. Thus, the fact that dopamine responses during early and advanced stages of learning were accurately approximated by value functions with small and large discount factors does not necessarily reflect adaptive change of encoding time scale. Through the process of learning, the monkeys may also have developed an action policy to accomplish the multitrial choices for long-term rewards and valuation of the long-term rewards. However, the inverted V-shape distribution of the start cue responses in well learned monkeys (Figs. 3 and 4) must have evolved through the learning of the discounted sum of the multiple future rewards rather than through action policy learning.

Dopamine Neuron Coding of Prediction Error of Immediate and Future Rewards. A considerable subset of dopamine neurons (16 of 25 neurons during the early stage, 21 of 26 neurons during the advanced stage in monkey SK, and 25 of 25 neurons during the advanced stage in monkey CC) also responded to reinforcer beep sounds after rewarding and nonrewarding choices with increased and decreased discharge rates, respectively (Fig. S3 *A* and *B*). At the time of the beeps (i.e., conditioned reinforcers) that followed individual choices, the error of reward prediction could be

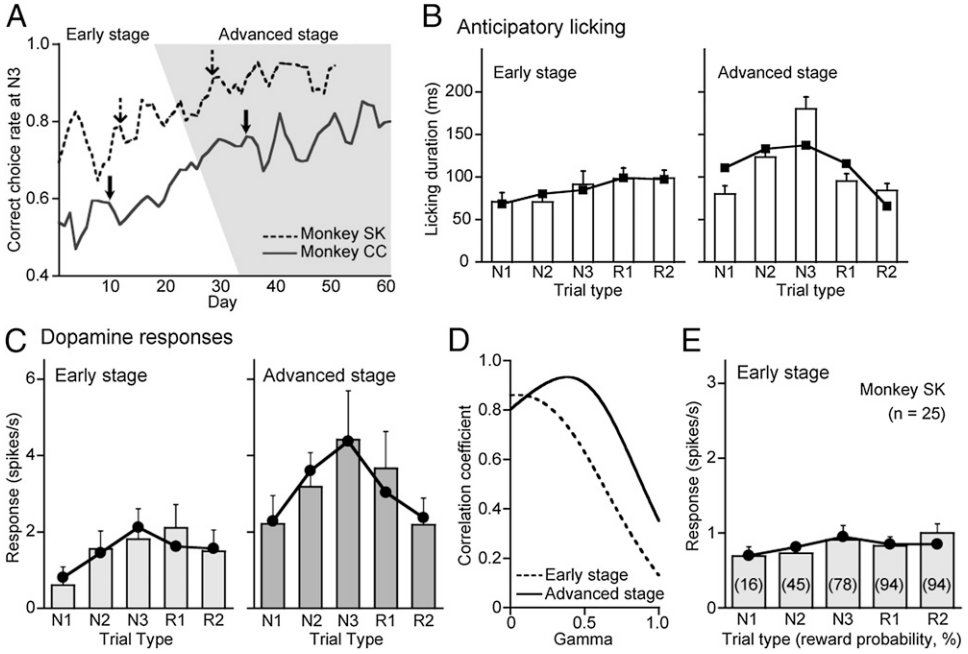

Fig. 4. Development of value coding by dopamine neurons through learning. (*A*) The adaptive increase in the correct choice rate in N3 trials through the learning of the three-step choice task for 51 to 61 d. The advanced stage of learning (correct choice rate > 0.8) is indicated by shading. (*B*) Bar graphs of the average duration of anticipatory licking on day 10 (early stage) and day 37 (advanced stage) of learning in monkey CC (mean and SEM, solid arrows in *A*). The best-fit value functions in the early stage ($\gamma = 0.05$, $R = 0.90$, $P < 0.05$) and in the advanced stage ($\gamma = 0.73$, $R = 0.69$, $P = 0.20$) are superimposed. (*C*) Bar graphs of start cue responses of an example neuron recorded on day 12 (early stage) and of another neuron on day 29 (advanced stage) in monkey SK (dashed arrows in *A*). Superimposed line plots are the best-fit value functions ($\gamma = 0.04$, $R = 0.83$, $P = 0.08$, day 12; $\gamma = 0.38$, $R = 0.91$, $P < 0.05$, day 29). (*D*) Plots of the parameter space landscape of correlation coefficients of the data in *C*. The values of second derivatives of *R* are −1.7 during the early stage and −3.0 during the advanced stage. (*E*) Bar graphs of ensemble average responses of 25 dopamine neurons (mean and SEM). Superimposed plots show the best-fit value function ($\gamma = 0.00$, $R = 0.72$, $P = 0.18$). Ensemble average responses during the advanced stage are shown in Fig. 3 *B* and *C*.

assessed. An important issue here was examining whether the magnitude of these responses represented the errors in the prediction of the sum of the immediate and distant rewards or the errors in the prediction of only the single, immediate rewards. The dopamine responses to positive reinforcers during the early stage of learning were fitted with a very small discount factor (γ = 0.00; Fig. S3*C*, *Left*; monkey SK), whereas those during the advanced stage were fitted with larger discount factors [γ = 0.31 (Fig. S3*C*, *Right*) for monkey SK; γ = 0.65 (Fig. S3*D*) for monkey CC]. These results are consistent with the hypothesis that the responses of dopamine neurons encode prediction errors of the sum of the immediate and distant rewards during the advanced stage through learning. Although dopamine neurons encoded the TD error fairly well ($R = 0.90, P < 0.05$) during the early stage of learning, it was more precise ($R = 0.98, P < 0.01$ in monkey SK; and $R = 0.99, P < 0.01$ in monkey CC) and had a larger gain (i.e., stronger discharges at the same level of error, such as in the N1 trials) during the advanced stage.

Dopamine neurons exhibited decreased discharge rates after negative reinforcers (17 of 25 neurons during the early stage, 14 of 26 neurons during the advanced stage in monkey SK, and 16 of 25 neurons during the advanced stage in monkey CC). The magnitudes of the responses in the N1, N2, and N3 trials monotonically developed with the increase in the reward probability of trials during the early and advanced stages in monkey SK, which was consistent with the encoding of negative reward prediction errors (Fig. S3*E*). The average responses in the R1 and R2 trials were not reliably estimated because of the small number of trials.

Discussion

In the present study, we provided direct evidence that dopamine neurons learn to encode the value of a series of actions as an expected sum of immediate reward and future discounted rewards, and its error, which were both parametrically estimated as the TD error in reinforcement learning theories. When values are fully learned, TD error at the trial start cue equals to the value of the cue for the next state, given that the value of the current state is zero (21). Although this has been postulated in theories, most previous studies on dopamine neurons used behavioral paradigms for a single reward. Experiments that used a multistep choice paradigm for rewards in collaboration with computational modeling allowed us to demonstrate the TD error coding by dopamine neurons. When monkeys expected the small magnitude of distant rewards in the advanced stage of learning, neuronal discount factor was smaller than when fixed amount of rewards were expected. Thus, the dopamine signals may play a role as a critical underlying component of intelligence by which humans and animals choose options that are expected to yield a large accumulation of rewards through the course of future works rather than adopt short-sighted action policies to work for an immediate reward (25, 26).

Value Coding of Single Reward with Delays and Sum of Expected Multiple Future Rewards. Previous studies have shown that the discounted value of a single delayed reward is represented by behavioral responses (27, 28) and neuronal activity (11, 12, 14, 15, 22) as a hyperbolic function (14, 29). A human imaging study (16) examined brain areas for reward prediction using different time scales in a Markov decision task that used a computational model-based regression analysis. They showed graded maps of time scales within the insula and the striatum; ventral parts of the anterior regions were involved in the prediction of immediate rewards (γ = 0), and dorsal parts of the posterior regions were involved in the prediction of future rewards (γ = 0.99). The extremely large and small discount factors are in contrast to those in the present study ($0.00 < \gamma < 0.73$). However, the time discounting of the expected values may depend critically on the behavioral context, and therefore on the behavioral tasks used, such as trial-based discounting as performed in the present study and time-based discounting as used in the previous work (16). In our multistep choice task, the value discounting of the summed future rewards should extend until the end of the exploitation trials. However, we think it is significant that the discount factor that was estimated from the monkey's anticipatory licking for the expectation of reward was almost the same as the discount factor that was estimated from the dopamine neuron response in the same monkey (γ = 0.66 and γ = 0.65, respectively, in monkey CC).

The temporal discounting in neuronal value coding may be especially useful for the valuation and economic choices of the reward events expected with short (e.g., within a few seconds) and variable delays (22). An abnormal bias toward a small, immediate reward in humans and animals, called impulsivity, has been reported to result from impairments of dopamine and reinforcement processes that mediate the effects of the immediate and delayed rewards (30–33). In contrast, the dopamine neuron coding for the sum of expected multiple future rewards may serve as an important brain mechanism that is involved in pursuit of unseen distant goals by the assignment of values to individual actions and, thus, the solving of temporal credit assignment problems in reinforcement learning theories (1, 21). In this case, the dopamine neuron coding may play crucial roles in the valuation of a chain of reward events that is expected over a longer time scale, such as several tens of minutes and hours (e.g., for a tennis match), several days and months (e.g., in a prediction of stock price changes), and even years (e.g., in life planning).

Dopamine neurons were previously shown to summate over multiple bouts of reward that are separated by a short delay and correctly treat these bouts as larger than a single reward (15). However, this does not fully account for the core prediction of reinforcement learning theory (i.e., the expected sum of future discounted rewards) that is tested in the present study. Most previous studies used TD learning models with exponential discounting, except for one recent work (34) that tried a hyperbolically discounted TD learning model. We used a standard TD model with exponential discounting because it can describe the learning recursively in a simple way. However, it would be an interesting future study to see which of the two models better fits the behavioral and dopamine neuron responses.

Previous studies have shown that dopamine neurons signal the occurrence of salient events for visually cued reward schedules (24) and the preference for advance information about upcoming rewards (23). A previous study in our laboratory suggested that, in monkeys, the responses of dopamine neurons to task start cues may be related to their motivation to work for the reward (5), based on the observation that the neuronal responses were negatively correlated with the task start times of individual trials under the same reward probability. However, it was not clear whether the dopamine neuron coding of motivation was based on an immediately expected reward or the sum of immediate and future rewards. The present study extended the previous study by finding that the behavioral coding estimated by anticipatory licking and the dopamine neuron coding of the value encompass the expected sum of immediate and future rewards. Furthermore, in the classical conditioning task, when the long-term expectation of multiple future rewards was not possible, dopamine neuron activities encoded the value of CSs in current trials. Therefore, our findings have directly demonstrated that dopamine neurons signal the sum of immediately expected reward and future, discounted rewards depending on the behavioral contexts.

A question arises regarding why the information about the sum of future rewards is coded at the start cue even if it is available at the time of the previous outcome. Indeed, the dopamine responses to action outcomes represented the prediction error of summed future rewards (Fig. S3). However, the responses of dopamine neurons to the task start cue signaled the sum of multiple future rewards. There are probably multiple potential reasons that explain the coding of the expected long-run value signal at the start of trials in the context of our behavioral task. Temporal uncertainty may play a significant role. The time intervals between the previous outcome and the start cue of the current trials were substantial, and they were not fixed, but varied considerably, between 6.5 s and 8.5 s. This prevented the monkeys from predicting the precise timing of the start cue. Moreover, measurements of anticipatory lickings after the start cue (Fig. 2) suggested that monkeys used the information about the sum of future rewards that they had obtained from the previous outcome to perform the

current choice. It would be helpful to signal the occurrence of the start cue of individual trials for the goal-directed behavior. This view is consistent with previous results that showed that the start cue responses of dopamine neurons are modulated by the "previous outcomes" (6, 7). Nevertheless, in the multistep choice paradigm of the present study, response magnitudes of the first exploitation (R1) trials were not different depending on how the first reward was obtained during the exploration trials; rewarded at the first exploration trial (N1), after one exploration trial without reward (N2), and after two exploration trials without reward (N3). Two-way ANOVA considering the last rewarded trial (N1, N2, and N3) and the index of neurons in our data set revealed no significant main effect of the last trial type [$F(2,1157) = 2.15$; $P = 0.12$ in monkey CC; $F(2,1111) = 2.53$, $P = 0.08$ in monkey SK in the three-step choice task]. This supported our results that the prediction of future rewards is the driving force of the inverted V-shape distribution of dopamine responses rather than the postdictive evaluation of past rewards.

Value Coding and Uncertainty of Reward and Decision. We quantitatively examined the possible involvement of uncertainty of reward availability (2). Because the maximum neuronal responses in the N3 trials occurred at reward probabilities much higher than 50% in the two monkeys that participated in the three-step choice task (75% in monkey CC and 90% in monkey SK; Fig. 1*C*), it is unlikely that uncertainty was a major determinant of the dopamine responses in our multistep choice paradigms. Even when reward uncertainty was greatly reduced in a control task in which monkeys were instructed which a single target to choose (Fig. S4*A*), the maximum responses in the N3 trials did not undergo any significant changes despite almost 100% reward probability (Mann–Whitney *U* test, $P = 0.89$; Fig. S4*B*). In addition, anticipatory, tonic increases in firing until the potential time of a probabilistic reward, which was previously reported to represent reward uncertainty (2), was not observed in the classical conditioning paradigm during the delay period between the CS presentation and the occurrence of reinforcer (Fig. S5*E*). One possible explanation for this discrepancy may be the differences in behavioral conditioning: in our paradigm, the monkeys depressed the "hold" button before the CS was presented, but in the previous report (2), the monkeys were conditioned in a standard classical conditioning paradigm without arm movement (*SI Text* explains another possibility).

Methods

The experiments were approved by the Animal Care and Use Committee of the Kyoto Prefectural University of Medicine and were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Three Japanese monkeys sat in a primate chair. They made multistep choices for rewards (Fig. 1*A*). The monkeys depressed a start button after it was illuminated. Then, three target buttons were simultaneously turned on, followed by the "go" stimulus after a short delay. The monkeys released the start button and depressed one of the three target buttons. If the chosen target was a rewarding one, a high-pitched beep sounded as a positive reinforcer and a drop of reward water was delivered; if the target was not a rewarding one, a low-pitched tone (a negative reinforcer) sounded and no reward was given. When the monkeys had hit a rewarding button, they could obtain one (two-step choice task) or two (three-step choice task) additional rewards by repeatedly choosing the same button as in the last trial. Therefore, the monkeys made a series of choices. The first choice explored among three alternatives in a trial-and-error manner; then, this knowledge was exploited for one (monkey BT) or two (monkeys CC and SK) more rewards (Fig. 1*B*). This instrumental, reward-pursuing choice task mimics somewhat the natural foraging behavior of monkeys.

We examined whether and how the responses of dopamine neurons to the start cue and the magnitude of anticipatory licking represented future reward values by estimating the value function of reinforcement learning theories as the sum of immediate and expected multiple future rewards discounted by the number of steps to obtain them (further details are provided in *SI Text*).

ACKNOWLEDGMENTS. The authors thank R. Sakane for technical assistance; K. Samejima, H. Yamada, Y. Hori, and K. Doya for their comments on an early version of the manuscript; and the anonymous reviewers for their valuable comments and advice during the course of revisions. This study was supported by Grant-in-Aid 17022032 for Scientific Research on Priority Areas (to M.K.) and by the Development of Biomarker Candidates for Social Behavior carried out under the Strategic Research Program for Brain Sciences from the Ministry of Education, Culture, Sports, Science and Technology of Japan (M.K.).

1. Sutton RS, Barto AG (1998) *Reinforcement Learning* (MIT Press, Cambridge, MA).
2. Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299:1898–1902.
3. Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* 43:133–143.
4. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci* 9:1057–1063.
5. Satoh T, Nakai S, Sato T, Kimura M (2003) Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci* 23:9913–9923.
6. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129–141.
7. Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O (2004) Dopamine neurons can represent context-dependent prediction error. *Neuron* 41:269–280.
8. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
9. Redgrave P, Prescott TJ, Gurney K (1999) Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci* 22:146–151.
10. Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459:837–841.
11. Roesch MR, Taylor AR, Schoenbaum G (2006) Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron* 51:509–520.
12. Kim S, Hwang J, Lee D (2008) Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron* 59:161–172.
13. Tsujimoto S, Sawaguchi T (2005) Neuronal activity representing temporal prediction of reward in the primate prefrontal cortex. *J Neurophysiol* 93:3687–3692.
14. Kobayashi S, Schultz W (2008) Influence of reward delays on responses of dopamine neurons. *J Neurosci* 28:7837–7846.
15. Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci* 10:1615–1624.
16. Tanaka SC, et al. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci* 7:887–893.
17. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306:503–507.
18. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. *Science* 310:1337–1340.
19. Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. *Neuron* 58:451–463.
20. Pasquereau B, et al. (2007) Shaping of motor responses by incentive values through the basal ganglia. *J Neurosci* 27:1176–1183.
21. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936–1947.
22. Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. *Nat Neurosci* 11:966–973.
23. Bromberg-Martin ES, Hikosaka O (2009) Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* 63:119–126.
24. Ravel S, Richmond BJ (2006) Dopamine neuronal responses in monkeys performing visually cued reward schedules. *Eur J Neurosci* 24:277–290.
25. Kacelnik A (1997) Normative and descriptive models of decision making: time discounting and risk sensitivity. *Ciba Found Symp* 208:51–67.
26. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–291.
27. Mazur JE (1987) An adjusting procedure for studying delayed reinforcement. *Quantitative Analyses of Behavior*, eds Commons ML, Mazur JE, Nevin JA, Rachlin H (Lawrence Erlbaum, Hillsdale, NJ),Vol 5.
28. Kable JW, Glimcher PW (2010) An "as soon as possible" effect in human intertemporal decision making: behavioral evidence and neural mechanisms. *J Neurophysiol* 103:2513–2531.
29. Louie K, Glimcher PW (2010) Separating value from choice: Delay discounting activity in the lateral intraparietal area. *J Neurosci* 30:5498–5507.
30. Williams J, Dayan P (2005) Dopamine, learning, and impulsivity: A biological account of attention-deficit/hyperactivity disorder. *J Child Adolesc Psychopharmacol* 15:160–179.
31. Cardinal RN, Winstanley CA, Robbins TW, Everitt BJ (2004) Limbic corticostriatal systems and delayed reinforcement. *Ann N Y Acad Sci* 1021:33–50.
32. Denk F, et al. (2005) Differential involvement of serotonin and dopamine systems in cost-benefit decisions about delay or effort. *Psychopharmacology (Berl)* 179:587–596.
33. Pine A, Shiner T, Seymour B, Dolan RJ (2010) Dopamine, time, and impulsivity in humans. *J Neurosci* 30:8888–8896.
34. Alexander WH, Brown JW (2010) Hyperbolically discounted temporal difference learning. *Neural Comput* 22:1511–1527.