Behavioral/Systems/Cognitive

Serotonin Affects Association of Aversive Outcomes to Past Actions

Saori C. Tanaka,[1,2,3] **Kazuhiro Shishida,**[3,5] **Nicolas Schweighofer,**[2,3,4] **Yasumasa Okamoto,**[3,5] **Shigeto Yamawaki,**[3,5] **and Kenji Doya**[2,3,6]

[1]Institute of Social and Economic Research, Osaka University, Osaka 567-0047, Japan, [2]Department of Computational Neurobiology, Advanced Telecommunication Research Institute International Computational Neuroscience Laboratories, Kyoto 619-0288, Japan, [3]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kyoto 619-0288, Japan, [4]Department of Biokinesiology and Physical Therapy, University of Southern California, Los Angeles, California 90089-9006, [5]Department of Psychiatry and Neurosciences, Hiroshima University, Hiroshima 734-8551, Japan, and [6]Neural Computation Unit, Okinawa Institute of Science and Technology, Okinawa 904-2234, Japan

Impairment in the serotonergic system has been linked to action choices that are less advantageous in a long run. Such impulsive choices can be caused by a deficit in linking a given reward or punishment with past actions. Here, we tested the effect of manipulation of the serotonergic system by tryptophan depletion and loading on learning the association of current rewards and punishments with past actions. We observed slower associative learning when actions were followed by a delayed punishment in the low serotonergic condition. Furthermore, a model-based analysis revealed a positive correlation between the length of the memory trace for aversive choices and subjects' blood tryptophan concentration. Our results suggest that the serotonergic system regulates the time scale of retrospective association of punishments to past actions.

Introduction

We must often learn to choose appropriate actions based on delayed rewards and punishments. In the game of chess, for instance, a move that takes the opponent's queen may appear to be a good one, but it may later turn out to be a critical mistake when one's king is lost as a consequence. Making progress in chess, or in any situations involving learning from delayed reward and punishment (Dickinson et al., 1992), requires solving the "temporal credit assignment problem," that is, linking the delayed outcomes to those actions responsible for these outcomes. This problem can be solved either via "prospective" learning of the value of future outcomes that will result from current actions, or via "retrospective" learning of the association between the present outcome with past actions. The inability, or reduced ability, to solve the temporal credit assignment problem leads to irrational behaviors, such as impulsive choice of an immediate small reward over a delayed large reward or avoidance of an immediate small punishment at the price of a later larger punishment. Clinical reports and animal experiments suggest that serotonin dysfunction is one of the leading causes of impulsive behaviors (Wogar et al., 1993; Evenden and Ryan, 1999; Mobini et al., 2000). In a previous study, we demonstrated that subjects under low serotonin levels showed impulsive choices because of reduced ability to predict future outcomes (Schweighofer et al., 2008). Here, we study how humans can solve the retrospective temporal credit assignment problems and whether serotonin has a role in retrospective learning.

The computational theory of "reinforcement learning" (Sutton and Barto, 1998) helps us formalize the prospective and retrospective ways to solve the temporal credit assignment problem and quantify the ability of the subjects in the different ways of learning. In the prospective way, the sum of future outcomes is estimated as the "value function" and actions are reinforced according to the temporal difference (TD) error between value function and actual outcomes. The alternative, retrospective, way is to maintain decaying "eligibility traces" for executed actions and at the time of outcome, reinforce the actions in proportion to the eligibility traces. In the prospective way, the time span of action-outcome associations is regulated by the "temporal discounting factor," often denoted by γ; in the retrospective way such time span is regulated by the "trace decay factor," often denoted λ. A low setting of either of these factors lets the learner neglect action-outcome association with a long interval and thus can cause impulsive choices (see Materials and Methods).

To test the effect of serotonin on the time span of retrospective temporal credit assignment, we developed a monetary choice task that is difficult to solve in a prospective way, and examined subjects' choices under different serotonin levels by dietary manipulation of the levels of tryptophan, a precursor of serotonin. We observed slow learning of delayed punishment at low serotonin levels. A computational analysis using

Received May 29, 2009; revised Aug. 25, 2009; accepted Nov. 9, 2009.

This research was supported in part by Core Research for Evolutional Science and Technology (CREST), Science and Technology of Japan, and Osaka University Global Center of Excellence Program. A part of this study is the result of "Development of biomarker candidates for social behavior" performed under the Strategic Research Program for Brain Sciences by the Ministry of Education, Culture, Sports, Science and Technology of Japan. We thank G. Okada, K. Ueda, A. Kinoshita, T. Mantani, N. Shirao, M. Sekida, H. Yamashita, H. Tanaka, O. Yamashita, K. Samejima, F. Ohtake, and M. Kawato for helpful discussions and technical advice.

Correspondence should be addressed to Saori C. Tanaka, Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan. E-mail: xsaori@iser.osaka-u.ac.jp.

DOI:10.1523/JNEUROSCI.2799-09.2009

 0270-6474/09/2915669-06$15.00/0

reinforcement learning model showed an effect of serotonin on the retrospective temporal credit assignment for delayed punishment: lower serotonin levels correlated with faster decay of eligibility traces.

Materials and Methods

Subjects and serotonin manipulation

Thirty-eight right-handed males (age range, 20–26 years; 22.0 ± 2.1 years, mean ± SD) gave their informed consent to participate in the study, which was conducted with the approval of the Institutional Review Board of Advanced Telecommunication Research Institute International (ATR) and Hiroshima University. On the day of the screening, a psychiatrist interviewed each volunteer to assess them for psychiatric problems using the Structured Clinical Interview (SCID) for DSM-IV, and each volunteer underwent a health examination, including a blood test, urine test, chest x-ray, and an electrocardiogram, to screen for health problems. To evaluate the personality of the volunteer, a psychiatrist administered the Temperament and Character Inventory (TCI), the neo Five Factor Inventory (neo-FFI), and the Beck Depression Inventory (BDI). We excluded 16 participants who had health and/or psychiatric problems.

All subjects participated on 1 d for screening and task training, and 22 subjects participated on 3 d for experiments under the three different tryptophan conditions (trp−, trp+, and control conditions). On the three experimental days, subjects consumed one of three amino acid drinks: one contained a standard amount of tryptophan (control; 2.3 g per 100 g of amino acid mixture), one contained excess tryptophan (trp+; 10.3 g), and one did not contain any tryptophan (trp−; 0 g). These experiments took place over an interval of more than 1 week to completely remove the effects of tryptophan dietary control on the last experiment day. The experiment was a counterbalanced, placebo-controlled, double-blind, within-subject design in which the controller prepared a counter-balanced schedule of the three tryptophan conditions for each subject. To maximize the pharmacological impact, the subject was instructed to consume only the low-protein diet we provided (<35 g/d total) beginning from 24 h before the experiment, and were instructed to fast overnight before each experiment day. Dietary tryptophan depletion is known to reduce the level of central serotonin metabolites in CSF (Young et al., 1985; Carpenter et al., 1998; Williams et al., 1999), and dietary tryptophan loading increases the level of CSF serotonin metabolites (Young and Gauthier, 1981; Bjork et al., 2000).

On each experiment day, two venous blood samples were obtained to determine the plasma free-tryptophan concentration, which was shown to correlate with the CSF serotonin level (Young and Gauthier, 1981; Young et al., 1985; Carpenter et al., 1998; Williams et al., 1999; Bjork et al., 2000). The first blood sample was obtained before consumption of the amino acid drink to determine the baseline plasma free tryptophan level, and the second one was taken 6 h after consumption of the amino acid drink to determine the effect of dietary manipulation of tryptophan on the plasma-free tryptophan level. After the second venipuncture, the subjects performed the task.

Amino acid mixtures

We prepared amino acid mixtures comprising the following quantities of 14 amino acid partially dissolved in 350 ml of water: L-tryptophan, 10.3 g (trp+), 2.3 g (control), or 0 g (trp−); 5.5 g of L-alanine; 4.9 g of L-arginine; 3.2 g of glycine; 3.2 g of L-histidine; 8.0 g of L-isoleucine; 13.5 g of L-leucine; 11.0 g of L-lysine monohydrochloride; 5.7 g of L-phenylalanine; 12.2 g of L-proline; 6.9 g of L-serine; 6.5 g of L-threonine; 6.9 g of L-tyrosine; and 8.9 g of L-valine. This aqueous suspension was flavored with 10 ml of chocolate syrup; in addition, 2.7 g of L-cysteine and 3.0 g of L-methionine were administered in a small amount water along with each of the trp−, trp+ and control drinks due to their unpalatability in the beverage. On each experiment day, all subjects received the same amino acid mixture except for the amount of tryptophan.

Behavioral task

On each experimental day, subjects performed a decision-making task 6 h after consumption of the amino acid drink. In each trial, the subject chose one of two fractal images displayed on the screen by pressing a button (Fig. 1*A*). Depending on the selected image (Fig. 1*B*), a monetary feedback with different outcomes (10, 40, −10, or −40 yen) was displayed either immediately after the button was pressed or three trials later (Fig. 1*C*). For example, +40(0) denoting gaining 40 yen within current trial (immediate reward) and −10(3) denoting losing 10 yen after 3 trials (delayed punishment).

At each trial, two fractal images were displayed side by side on the screen. We prepared 16 pairs, counterbalancing the number of times of appearance of each image. The 16 pairs of images were presented in pseudo-random order. Each pair was presented as a scheduled trial number: each of six pairs (+10(0) vs +40(0); +10(3) vs +40(3); −10(0) vs −40(0); −10(3) vs −40(3); +10(0) vs +40(3); −10(0) vs −40(3)) was presented in 10 trials during a single session, and each of 10 pairs (+40(0) vs +40(3); +10(0) vs +10(3); −10(0) vs −10(3); −40(0) vs −40(3); +10(3) vs +40(0); −10(3) vs −40(0); +10(0) vs −10(0); +40(0) vs −40(0); +10(3) vs −10(3); +40(3) vs −40(3)) was presented in five trials during a single session.

The subjects were not informed of the stimulus-outcome associations shown in Figure 1*B* and received money after the experiment in proportion to the total outcome that the subject earned during the experiment. To maximize the total outcome, the subjects needed to learn the possible stimulus-outcome associations by correctly assigning the credit of the present feedback to the chosen images that caused the present feedback.

Each subject performed 110 trials during a single session, and six sessions on each experiment day. It took ~28 min for subjects to complete six sessions. At the beginning of each session, the session number was displayed on the screen for 2.5 s. On the screening day, each subject practiced the test session under the same task settings used in each experimental day except for images, and we confirmed that all subjects understood the task setting. We prepared different images for each subject and each experiment day.

Data analysis

Reinforcement learning model. In temporal difference (TD) learning, the "value" V of state s at time t is defined as the sum of future outcomes and an action a taken at time t is reinforced according to the TD error,

$$\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t)), \quad (1)$$

where γ is a discount factor ($0 \leq \gamma < 1$). Although a TD learning framework has been successfully applied to a variety of problems, a critical constraint is that the future rewards can be consistently predicted from the state representation s. When the environment has unobservable states, TD learning can be poor.

In the framework for retrospective learning, the eligibility trace for an action is incremented when it is chosen, and decays by a coefficient λ:

$$e_i(t) = \begin{cases} \lambda e_i(t-1) + 1 & (a(t) = a_i), \\ \lambda e_i(t-1) & (a(t) \neq a_i). \end{cases} \quad (2)$$

The trace-decay parameter λ ($0 \leq \lambda < 1$) controls the time scale of temporal credit assignment. An obvious drawback of this method is the lack of selectivity: a reward or punishment is associated with all of the preceding actions simply with immediacy weighting by the parameter λ. In practical reinforcement learning applications, both prospective and retrospective methods are often combined; this is known as TD(λ) learning (Sutton, 1988).

Eligibility trace model for association learning based on a delayed reinforcer. To examine the effect of the trace-decay parameter of the eligibility trace on subjects' behavior, we generated artificial choice data using the eligibility trace model with varying λ. We defined the value function of each image s_i by $V(s_i(t))$, and eligibility trace by $e(s_i(t))$ at trial t. The eligibility traces for all images decay by λ, and the

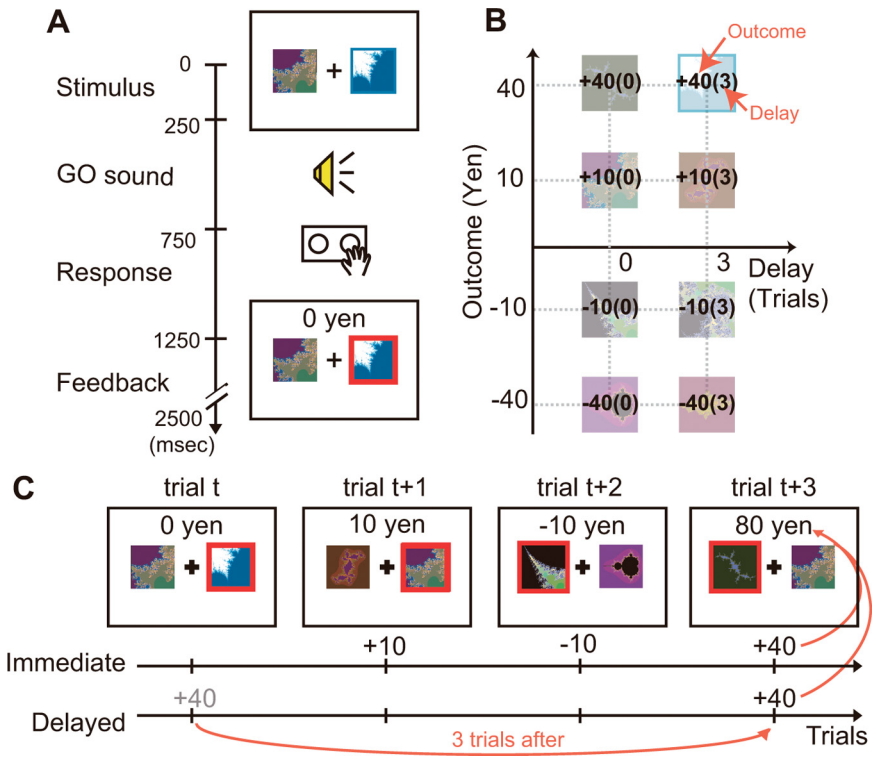

Figure 1. ***A***, Experimental task. Two fractal images were displayed on the screen in each trial. When the subject heard a beep, the subject chose one of two fractal images by pressing the corresponding button within 1 s of the beep. The outcome was displayed on the screen. A single trial lasted 2.5 s. At the next trial, a new pair of fractal images was displayed. ***B***, Outcome-delay mapping of fractal images. On each experiment day, we used eight fractal images. Each of the eight images had been assigned a different outcome (40, 10, −10, −40 yen) and delay (0 trial, 3 trials). At each trial, two fractal images were chosen from these eight images in a pseudo-random order. ***C***, If the subject selected a delay 0 image, the outcome was displayed during feedback duration in the present trial (see trials $t+1$ and $t+2$). If the subject chose a delay 3 image, the outcome was displayed three trials later (trial t and $t+3$). If the outcomes from delay 0 and delay 3 images appeared in the same trial, the sum of the immediate and delayed outcomes was displayed (trial $t+3$). When no outcome was delivered, "0 yen" was displayed (trial t).

eligibility trace for the one image chosen at trial t is incremented by 1, as follows:

$$e(s_i(t)) = \begin{cases} \lambda e(s_i(t-1)) + 1 & (s_i(t) = s(t)), \\ \lambda e(s_i(t-1)) & (s_i(t) \neq s(t)). \end{cases} \quad (3)$$

The value function of image s_i at trial t was updated by the REINFORCE algorithm (Williams, 1992):

$$V(s_i(t+1)) = V(s_i(t)) + \alpha[r(t) - V(s(t))]e(s_i(t)), \quad (4)$$

where α is the learning rate ($0 \leq \alpha < 1$), $r(t)$ is an outcome displayed at feedback timing, and $V(s(t))$ is the value of the chosen image at trial t. In this model, the value function of each image can be learned indirectly by applying the eligibility trace. Here, we used Soft Max as the action selection strategy,

$$P(s_{\text{right}}(t)) = \frac{\exp(\beta V(s_{\text{right}}(t)))}{\exp(\beta V(s_{\text{right}}(t))) + \exp(\beta V(s_{\text{left}}(t)))}. \quad (5)$$

This function determines the probability of selecting the image displayed on the right side of the screen at trial t, where V means the value of the image displayed at the right or left sides at trial t, and β ($0 \leq \beta < 1$) is the 'inverse temperature' parameter, which determines the randomness of action selection.

Figure S1 (available at www.jneurosci.org as supplemental material) shows an example of the time course of the eligibility trace of one stimulus. For $\lambda = 0$, the eligibility trace was a spike-like pattern (supplemental Fig. S1*A*, available at www.jneurosci.org as supplemental material); thus, TD error was used to update V for only the present selected image. For $\lambda = 0.8$, the eligibility trace was sustained over several trials with temporal decay (supplemental Fig. S1*B*, available at www.jneurosci.org as supplemental material); thus, TD error was used to update the V of not only the present image, but also past images. For excessively large λ (=0.99), the eligibility trace was not discounted for a long period of time (supplemental Fig. S1*C*, available at www.jneurosci.org as supplemental material); thus, TD error was used to update the V, even of images that were visited in the distant past.

Model comparison. To evaluate whether the retrospective learning model can explain a subject's behavior, we compared the log likelihood of the subjects' action with that in three models: the retrospective model (REINFORCE algorithm, see Eqs. 3 and 4), the prospective model (TD(0)), and the combined model (TD(λ)). In TD models, we assumed that the subject had complete memory and knowledge of the images chosen during the previous three trials. The TD error was computed by:

$$\delta(t) = r(t) + V'(t) - V(t), \quad (6)$$

$$V'(t) = \sum_{T=t-3}^{t} V(s_T)\,|_{\,T-(t-3)-delay(s_T)>0}$$

$$V(t) = \sum_{T=t-3}^{t} V(s_T)\,|_{\,T-(t-3)-delay(s_T)\geq 0}$$

where $delay(s_t)$ is the delay length of the image chosen at trial t. The value function of image s_i at trial t was updated in the TD(λ) as

$$V(s_i(t+1)) = V(s_i(t)) + \alpha\delta(t)e(s_i(t)), \quad (7)$$

and in the TD(0) as

$$V(s_i(t+1)) = V(s_i(t)) + \alpha\delta(t). \quad (8)$$

In each model, we used Soft Max as the action selection strategy (Eq. 5).

Estimation of subject's meta-parameters. We estimated subjects' three meta-parameters in this model, α, β and λ, maximizing the log likelihood of the subjects' action at each tryptophan level. We defined different eligibility trace decay parameters λ for reward and punishment based on our behavioral results that serotonin differentially affected the choice probability for delayed reward and punishment. We performed multiple regression analysis with two explanatory variables: the concentration of plasma-free tryptophan and the index of subjects. All subjects' values for the concentration of plasma-free tryptophan were less than the detection limit (0.5 nmol/ml); thus, we used 0.5 nmol/ml in multiple regression analysis.

Statistical analysis of data. Each multiple comparison was performed after a repeated-measures ANOVA with three tryptophan conditions ($n = 21$). In all cases, we found a significant main effect of tryptophan condition ($p < 0.05$). In the statistical test of simulated choice probability with estimated parameters, we had an a priori hypothesis based on our behavioral results found in Figure 2 that the choice probability of smaller delayed punishment under the trp− condition was smaller than under both the control and trp+ conditions. In this case we could skip a repeated-measures ANOVA test and apply t test twice for target pairs without any corrections.

Results

Serotonin manipulation results

Table S1, available at www.jneurosci.org as supplemental material shows the plasma-free tryptophan levels of the subjects before and 6 h after consumption of each amino acid drink. Except in one subject (subject #2, who was omitted from subsequent analyses), 6 h after consumption, the plasma free-tryptophan level significantly decreased in the trp− condition ($p < 0.0001$, paired t test) and significantly increased in the trp+ condition ($p <$

0.0001, paired t test) compared with the respective levels before consumption. Based on previous studies of dietary tryptophan depletion (Young et al., 1985; Carpenter et al., 1998; Williams et al., 1999) and loading (Young and Gauthier, 1981; Bjork et al., 2000), we assumed that there were significant decreases and increases in central serotonin levels, respectively.

Behavioral results

Figure S2 (available at www.jneurosci.org as supplemental material) shows choice patterns at each pair in early (trial 1–110) and latter (551–660) trials in the control condition. The size of arrowhead of each orange line show averaged choice probabilities of connected images. In early trials, subjects clearly showed choice preferences in the pairs with delay 0 (pairs 1, 7, 13, and 14). It took a longer time to learn the optimal choices in the delayed pairs (pairs 2, 8, 15, and 16) than in the immediate pairs. We also found similar choice patterns both in trp− and trp+ conditions.

We found that retrospective learning better explained subjects' behavior than prospective learning ($p < 0.0001$ for multiple comparison with Bonferroni correction; see supplemental Fig. S3, available at www.jneurosci.org as supplemental material). To learn this task in a prospective way, subjects would need to maintain a state representation consisting of the sequence of his past choices $s(t) = \{a(t-1), a(t-2), \ldots\}$ and to learn the appropriate values for this high-dimension state vector, requiring a heavy load of updating working memories and adjusting many parameters. On the other hand, by using retrospective learning, subjects can simply rely on the decaying memories of past actions.

Based on our hypothesis of an effect of serotonin on learning with delayed outcomes, we compared subjects' choice probabilities for the four pairs of interest (immediate rewards, immediate punishments, delayed rewards, and delayed punishments) at three stages of learning (early: first half of the first session, middle: the second session, late: the last session) under three tryptophan conditions (Fig. 2). We found that the optimal choice probability for the delayed punishment pair (Fig. 2, right bottom panel: −10(3) vs −40(3)) was significantly lower in the middle stage under the trp− condition (choice probability of −10(3) over −40(3), 0.681 ± 0.0642, mean ± SEM) than in the control condition (0.829 ± 0.0421, $p = 0.046$ for multiple comparison with Bonferroni correction) and the trp+ condition (0.833 ± 0.0475, $p = 0.033$ for multiple comparison with Bonferroni correction). We did not find a significant effect of tryptophan conditions in the early and late stages, indicating a slower learning of delayed punishments under the trp− condition. In contrast, we did not find any significant effects of tryptophan conditions on the choice probabilities for other three pairs.

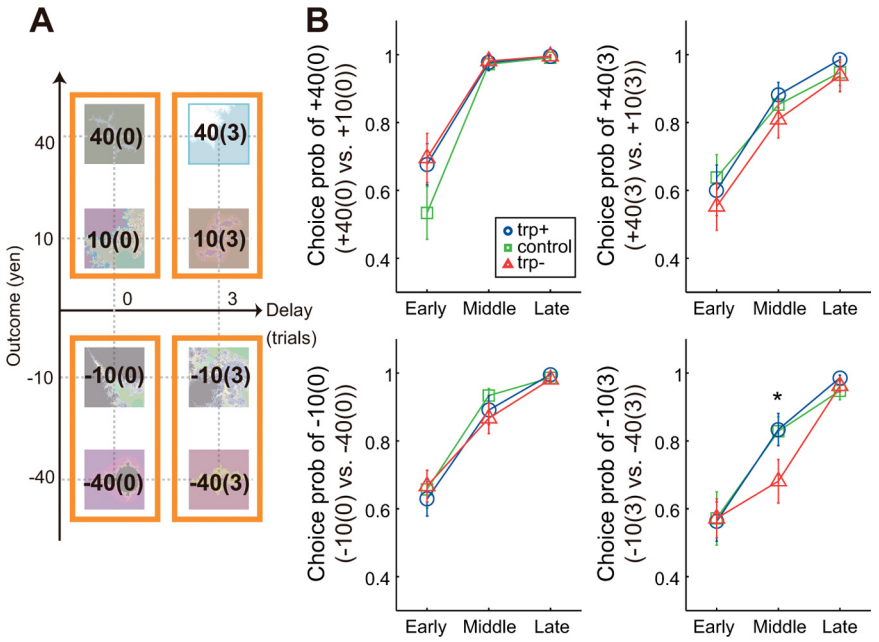

Figure 2. Choice probabilities of the optimal image (in ***A***) for the four pairs with the same delay and different magnitude. ***B*** (left top panel), Immediate reward (+10(0) vs +40(0)) (right top panel) delayed reward (+10(3) vs +40(3)) (left bottom panel) immediate punishment (−10(0) vs −40(0)), and (right bottom panel) delayed punishment (−10(3) vs −40(3)). The optimal image was defined as the larger reward or smaller punishment. In each panel, we plotted the averaged choice probabilities of the optimal image in each pair during the early stage (first half of the first session), middle stage (the second session), and late stage (the last session). Each error bar shows the SEM ($n = 21$). *$p < 0.05$, multiple comparisons with Bonferroni correction between trp− and other conditions.

Model-based behavioral analyses

To clarify the effect of serotonin on action learning, we analyzed the subjects' choice behavior using a computational model of temporal credit assignment (Sutton and Barto, 1998). We estimated each subject's parameters of learning (learning rate α, inverse temperature β, and trace decay factor λ) (Doya, 2002) so that the likelihood of reproducing the subject's action sequence is maximized (Samejima et al., 2005; Tanaka et al., 2006). Given the differential effect of tryptophan conditions on learning from delayed rewards and punishments (Fig. 2, right panels), we estimated separate trace decay factors for rewards ($\lambda+$) and punishments ($\lambda-$). Figure 3*A* shows the estimated parameters at each tryptophan condition. We found that the estimated $\lambda-$ was significantly smaller under the trp− condition than under the trp+ condition ($p = 0.047$ for multiple comparison with Bonferroni correction). To take into account the individual variability in the effect of dietary manipulation, we performed a regression analysis of the estimated parameters and the blood tryptophan concentrations of each subject (supplemental Fig. S4, available at www.jneurosci.org as supplemental material). We observed a significant positive correlation between the estimated $\lambda-$ and tryptophan concentration ($p = 0.0172$, $R^2 = 0.525$). We did not observe a significant effect of tryptophan conditions on other parameters, learning rate α ($p = 0.806$, $R^2 = 0.265$), inverse temperature β ($p = 0.0565$, $R^2 = 0.476$), and trace decay factor for reward $\lambda+$ ($p = 0.676$, $R^2 = 0.297$).

We ran simulations of the eligibility model with estimated meta-parameters. The results in Fig. 1*B* show simulated choice probabilities averaged across 21 subjects in each tryptophan condition. As in our behavioral result, simulations show slow learning of delay punishments under the low serotonin condition (Fig. 3*B*): in the middle stage, the probability of smaller delayed punishment choice in the delayed punishment pair under the trp− condition (0.6143 ± 0.05313) is lower than under both the control condition (0.7095 ± 0.04875, $p = 0.037$) and the trp+ (0.7476 ± 0.04344, $p = 0.038$) condition (Fig. 3*B*, right bottom panel; a priori comparison (uncorrected one-tailed paired t test) based on our behavioral results shown in Fig. 2). The time course

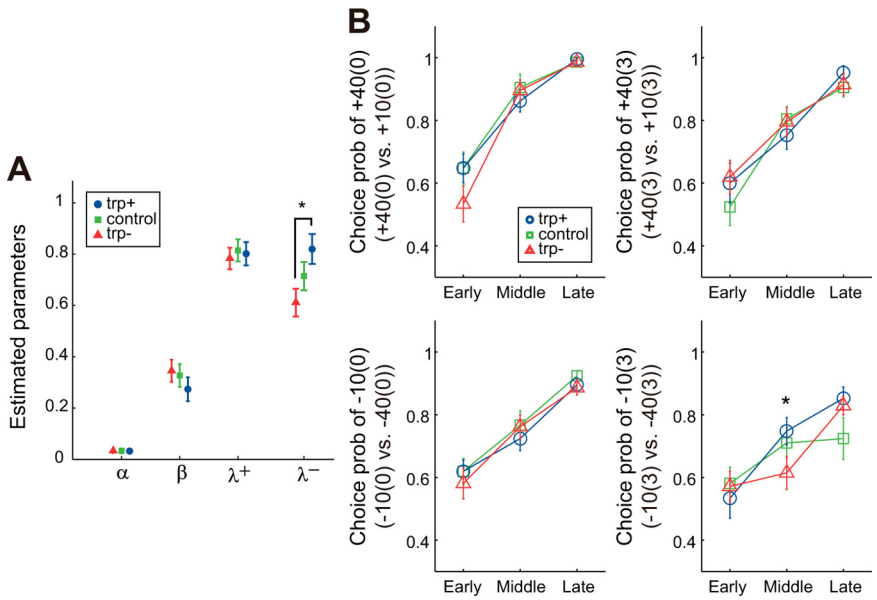

Figure 3. ***A***, Estimated parameters in each tryptophan condition (red triangle, trp− condition; green square, control condition; blue circle, trp+ condition). α, learning rate; β, inverse temperature; $\lambda+$, trace decay factor for reward; $\lambda-$, trace decay factor for punishment. *$p < 0.05$ in multiple comparison with Bonferroni correction. ***B***, Simulated choice probabilities for the optimal image in the four pairs with estimated subjects' parameters. Each error bar shows the SEM ($n = 21$). *$p < 0.05$, a priori comparisons (uncorrected one-tailed paired t test) between trp− and other conditions.

of the estimated value function explained the subjects' actual choice with the delayed punishment pair well (supplemental Fig. S5, available at www.jneurosci.org as supplemental material).

Discussion

Compared with controls, tryptophan-depleted subjects showed slower learning from delayed punishments, but showed no difference in learning from immediate punishments or delayed rewards. A computational model-based analysis revealed a correlation between faster decay of eligibility trace for punishments and lower blood tryptophan levels. The model simulation using the parameters estimated from subjects' behaviors replicated the less optimal choice of delayed punishments in the middle stage of learning under tryptophan-depleted condition and explained the slower learning from delayed punishments as the difficulty of associating punishments with choices in the longer past under low serotonin levels.

We did not find a significant effect of tryptophan loading on choice behavior, as was the case in our previous study using the same tryptophan manipulation (Schweighofer et al., 2008). One possible reason is variable effects of tryptophan loading on central serotonin levels, which is likely given the large variance among the subjects in the concentrations of plasma-free tryptophan in the tryptophan-loaded condition (supplemental Table S1, available at www.jneurosci.org as supplemental material). Regression analysis between the estimated eligibility trace decay factor and the measured concentration of plasma-free tryptophan corrected for the subject-wise variance and showed a significant correlation.

We found a significant serotonergic effect on action learning based on delayed punishment, but not delayed reward. A possible reason for the difference is a sampling bias; as learning proceeds, aversive stimuli are less frequently selected than rewarding stimuli, which can cause apparently slower learning of punishment. By comparing the choice probabilities of rewards and punishments in terms of the number of experience rather than experimental stages, we could however rule out such an explanation (see supplemental Fig. S6, available at www.jneurosci.org as supplemental material).

A possible confounding factor in discriminating the effects of gains and losses is the "house money effect," which means that after a gain, subsequent losses that are smaller than the original gain can be integrated with the prior gain, mitigating the influence of loss aversion and facilitating risk-seeking (Thaler, 1991). Although trial-by-trial losses resulted in a decrement in the gain that the subject would receive at the end of the study, we expect that the house money effect was minimal in our task for the following reasons. First, subjects received visual feedback of the outcome for only the present trial and were not informed of the cumulative gain. Because subjects experienced gains and losses of variable sizes in a random order while trying to learn the cue-outcome association, it would have been very difficult for them to keep track of how much they had gained or lost so far. Because subjects did not receive any initial "house money" at the beginning of the task, subjects' cumulative gains were often negative during the task although all subjects received positive reward (~800 Japanese yen) at the end of their experiments.

In the model-based analysis, we estimated separate trace decay factors for rewards and punishments although subjects had no knowledge about association between stimuli and losses or gains in the beginning of the task. A possible mechanism behind the separate eligibility traces for gains and losses is multiple copies of eligibility traces in the brain; different brain areas (or networks) are specialized in learning from gains and losses using separate eligibility traces with different trace decay parameters. For the same cue, multiple eligibility traces are activated with different decay parameters, and the effective decay parameters can be different depending on whether the cue is associated with a gain or a loss. This multiple system does not require any knowledge about states at the start of the task from the subject. Such an implementation is consistent with the biological findings that even among dopamine neurons there are specialization for positive and negative rewards (Matsumoto and Hikosaka, 2009), and that there are multiple subsystems with different delay discounting in the striatum (Tanaka et al., 2004, 2006, 2007).

Although previous studies have demonstrated an effect of serotonin on learning based on aversive stimuli (Deakin and Graeff, 1991; Harvey, 1996; Buhot, 1997), differential effects of serotonin on immediate and delayed punishments have not been explored. In our previous study, we demonstrated a serotonergic effect on the time span of prospective evaluation of future rewards, but not punishments (Tanaka et al., 2007; Schweighofer et al., 2008). The present study is the first to demonstrate the serotonergic modulation of the time span of retrospective association of punishments to past events. Distinct regulation of the time spans of prospective and retrospective learning from rewards and punishments can be realized by separate serotonergic projection pathways to different brain areas. While the serotonergic projections from the dorsal raphe nucleus mainly target the

striatum and the frontal cortex, which have been shown to be involved in reward predictive learning (Kawagoe et al., 1998; Shidara et al., 1998; Tremblay et al., 1998; Corbit and Balleine, 2003; Matsumoto et al., 2003; McClure et al., 2003; O'Doherty et al., 2003; Samejima et al., 2005; Hampton et al., 2006), those from the median raphe nucleus mainly target the limbic system, which have been shown to be involved in aversive learning and memory (Kim and Fanselow, 1992; Kim et al., 1993). Serotonergic modulation of the time span of prospective and retrospective learning may enable consistent regulation of both systems, and thus facilitate effective action learning.

References

Bjork JM, Dougherty DM, Moeller FG, Swann AC (2000) Differential behavioral effects of plasma tryptophan depletion and loading in aggressive and nonaggressive men. Neuropsychopharmacology 22:357–369.

Buhot MC (1997) Serotonin receptors in cognitive behaviors. Curr Opin Neurobiol 7:243–254.

Carpenter LL, Anderson GM, Pelton GH, Gudin JA, Kirwin PD, Price LH, Heninger GR, McDougle CJ (1998) Tryptophan depletion during continuous CSF sampling in healthy human subjects. Neuropsychopharmacology 19:26–35.

Corbit LH, Balleine BW (2003) The role of prelimbic cortex in instrumental conditioning. Behav Brain Res 146:145–157.

Deakin JFW, Graeff FG (1991) 5-HT and mechanism of defence. J Psychopharmacol 5:305–315.

Dickinson A, Watt A, Griffiths WJH (1992) Free-operant acquisition with delayed reinforcement. Q J Exp Psychol 45B:241–258.

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

Evenden JL, Ryan CN (1999) The pharmacology of impulsive behaviour in rats VI: the effects of ethanol and selective serotonergic drugs on response choice with varying delays of reinforcement. Psychopharmacology (Berl) 146:413–421.

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

Harvey JA (1996) Serotonergic regulation of associative learning. Behav Brain Res 73:47–50.

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

Kim JJ, Fanselow MS (1992) Modality-specific retrograde amnesia of fear. Science 256:675–677.

Kim JJ, Rison RA, Fanselow MS (1993) Effects of amygdala, hippocampus, and periaqueductal gray lesions on short- and long-term contextual fear. Behav Neurosci 107:1093–1098.

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

Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459:837–841.

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.

Mobini S, Chiang TJ, Ho MY, Bradshaw CM, Szabadi E (2000) Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. Psychopharmacology (Berl) 152:390–397.

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

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

Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, Yamawaki S, Doya K (2008) Low-serotonin levels increase delayed reward discounting in humans. J Neurosci 28:4528–4532.

Shidara M, Aigner TG, Richmond BJ (1998) Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. J Neurosci 18:2613–2625.

Sutton RS (1988) Learning to predict by the methods of temporal differences. Machine Learning 3:9–44.

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

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

Tanaka SC, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S, Doya K (2006) Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. Neural Netw 19:1233–1241.

Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, Doya K (2007) Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. PLoS ONE 2:e1333.

Thaler RH (1991) Quasi rational economics. New York: Russel Sage Foundation.

Tremblay L, Hollerman JR, Schultz W (1998) Modifications of reward expectation-related neuronal activity during learning in primate striatum. J Neurophysiol 80:964–977.

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8:229–256.

Williams WA, Shoaf SE, Hommer D, Rawlings R, Linnoila M (1999) Effects of acute tryptophan depletion on plasma and cerebrospinal fluid tryptophan and 5-hydroxyindoleacetic acid in normal volunteers. J Neurochem 72:1641–1647.

Wogar MA, Bradshaw CM, Szabadi E (1993) Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers. Psychopharmacology (Berl) 111:239–243.

Young SN, Gauthier S (1981) Effect of tryptophan administration on tryptophan, 5-hydroxyindoleacetic acid and indoleacetic acid in human lumbar and cisternal cerebrospinal fluid. J Neurol Neurosurg Psychiatry 44:323–328.

Young SN, Smith SE, Pihl RO, Ervin FR (1985) Tryptophan depletion causes a rapid lowering of mood in normal males. Psychopharmacology (Berl) 87:173–177.