NAIST-IS-DD0361017

Doctoral Dissertation

Functional Model of Serotonin in Human Reward System Based on Reinforcement Learning Theory

Saori Tanaka

February 13, 2006

Department of Information Systems
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Science.

Thesis Committee:

Professor Shin Ishii	(Supervisor)
Professor Mitsuo Kawato	(Co-supervisor)
Professor Kotarou Minato	(Co-supervisor)
Associate Professor Kenji Doya	(Co-supervisor)
Associate Professor Tomohiro Shibata	(Co-supervisor)

強化学習理論からみたヒトの報酬系における

セロトニンの機能モデル

田中沙織

内容梗概

我々ヒトをふくむ動物は，変動する環境の中において生き残るために，可能な多くの選択肢の中から行動を選択している．このような行動選択に大きな影響を与えるのが「報酬」である．「強化学習理論」は，報酬予測と実際の報酬の誤差を，行動の良し悪しを評価する学習信号として用い，長い目で見てより多くの報酬を得られるような行動則を探索的に学習する理論的枠組みである．強化学習をヒトをふくむ動物の行動学習モデルとして考える場合，実装上の重要な問題である学習パラメータの設定はどのように行われるのか．我々は報酬予測の時間スケールパラメータである減衰率に着目した．動物実験や離床事例にみられるセロトニン減少に伴う衝動的行動が，低い減衰率を用いたモデルで説明できることから，セロトニンがこの減衰率を調整するという仮説をたてた．この仮説を実証するため，計算論的モデルに基づいた実験タスクを用いて，ヒトを対象とした一連の行動・脳活動計測実験を行った．異なる時間スケールでの報酬予測を必要とする課題実行中の脳活動と，強化学習モデルの出力との相関を調べた結果，異なる時間スケールの報酬予測には線条体を通る異なるネットワークが関わることを示した．またセロトニンレベルを人為的に操作することで，線条体の報酬予測関連活動がセロトニンによって制御されることを示した．また衝動性の原因として時間遅れを伴う報酬による行動学習の障害も考えられることから，セロトニンが行動学習に与える影響を調べたところ，報酬と過去の行動を関連付けるときの時間スケールを制御していることを示した．これらの結果から，報酬予測や行動学習におけるセロトニンの機能について，具体的な脳内メカニズムを提唱する．

キーワード

強化学習理論，減衰率，セロトニン，fMRI

*奈良先端科学技術大学院大学情報科学研究科情報システム学専攻学位論文，
NAIST-IS-DD0361017, 2006 年 2 月 13 日.

Functional Model of Serotonin in Human Reward System Based on Reinforcement Learning Theory

Saori Tanaka

Abstract

To survive in variable environment, animals need to make appropriate choice from among alternatives. Reinforcement learning is theoretical framework of learning a policy maximizing total outcomes in the long run by trial and error, and has been focused as computational model explaining animal and human action learning behaviors. Because impulsive behaviors caused by central serotonergic impairment can be explained by the reinforcement learning model with shorter time scales, we hypothesize that serotonin controls the time scale of reward prediction. To test this hypothesis, we performed a series of experiments using model-inspired tasks and brain imaging method with human subjects. By regression analysis of output of reinforcement learning model and human brain activities, we demonstrated that parallel cortico-striatum loops were involved in reward prediction at different time scales. Further, by manipulation of central serotonin levels, we revealed that reward predictive activities in the stratum were differentially modulated by serotonin. In recent experiment, based on the possibility that immature learning of delayed outcomes can cause impulsive behaviors, we tested the serotonergic effects on action learning. We showed that serotonin also affected the time scale of temporal credit assignment in association learning between action and delayed outcome. From these results, we suggest the detailed functional model of serotonin in reward-based learning system.

Keywords:

Reinforcement learning theory, discounting factor, serotonin, fMRI

.*Doctoral Dissertation, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0361017, February 13, 2006.

Contents

Chapter 1

General Introduction

Chapter 2

Prediction of Immediate and Future Rewards Differentially Recruits Cortico-Basal Ganglia Loops

Chapter 5

Chapter 6

Chapter 1

General Introduction

In our everyday life, we perform various actions. In the morning, I brush my teeth, wash my face, and ride my bicycle to the office, feeling the fresh wind. In my office, I attempt to gauge my boss' mood from his e-mail. In the cafeteria, I pick lunch from the menu, and feel disappointed with a meal less palatable than expected. After work, I make dinner with what I have in the fridge, and soak myself in a hot bath while contemplating the events of the day. These actions all need various functions, for example, sequential movement using acquired motor skills, estimation of others' internal models, decision-making based on reward prediction, working memory, short-term/long-term memory, and so on. In these various behaviors, we should act by our own free will. But is it possible to explain our "will" in objective terms? Computational neuroscience has been succeeding in objective descriptions of our biological phenomena, such as behavior of molecules in cells, neuron firing, and trajectory generation of arm movement. We adopt a "reinforcement learning theory" as a candidate for computational model of animal's reward-based action learning and decision-making, and try to elucidate its brain mechanism. In this introduction, we start by explaining a reinforcement learning theory and our hypotheses for learning mechanism in the brain.

1. Reinforcement learning theory and brain mechanism

1.1. Reinforcement learning theory

"Reinforcement" means enhancement of the association between stimulus and response in behavioral psychology studies. Studies on animals have demonstrated that stimuli resulting in reward or punishment generated predictive responses such as expecting a reward or avoiding punishment (Thorndike 1911). In such instrumental conditioning, reward (or punishment) acts as a reinforcer that reinforces the association between a particular stimulus and response. There are two prominent types of reward. One is primary reward: the essential needs of life, like thirst, hunger, sleep, and sex. The other is secondary reward; it is not reward in itself, but something that can generate primary reward, like money.

Reinforcement learning is a theoretical model of instrumental learning under an ideal environment (Sutton and Barto 1998). In reinforcement learning theory, the learning problem is to find a policy $a = \mu(s)$ that maximizes the predicted amount of future reward with temporal discounting, formulated as the "value function"

$$V(t) = E[r(t + 1) + \gamma r(t + 2) + \gamma^2 r(t + 3) + \dots],$$

under the environment in which action $a(t)$ at state $s(t)$ results in reward $r(t)$ and state changes $s(t+1)$. The "discount factor" γ $(0 \leq \gamma < 1)$ controls the time scale of prediction; while only the immediate reward $r(t + 1)$ is considered with $\gamma = 0$, rewards in the more distant future are taken into account with γ closer to 1. Thus, reinforcement learning is learning the optimal value function

$$V^{\mu}(s(t)) = \mathrm{E}[r(t+1)+ \gamma r(t+2)+...]$$

that fulfills the Bellman equation

$$V(s) = \max_{a} \mathrm{E}[r(s, a)+ \gamma V(s(s, a))].$$

Any deviation from the prediction is given by the temporal difference (TD) error

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

which is a crucial learning signal for reward prediction that acts by updating the older value function in proportion to it, and for action learning by increasing the probability of taking $a(t)$ at state $s(t)$, $P(a(t)|s(t))$, at a positive TD value, or a decreasing probability at a negative TD value.

1.2. Reinforcement learning model of the basal ganglia

Recently, reinforcement learning theory has received attention as a computational model explaining animal and human action learning behavior, and has been successfully used for explaining reward-predictive activities of the midbrain dopaminergic system as well as those of the cortex and the striatum (Houk, Adams et al. 1995; Schultz, Dayan et al. 1997; Doya 2000; Berns, McClure et al. 2001; McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003).

Schultz and his colleagues recorded the activities of dopamine neurons in the substantial nigra during a conditioning task in which monkeys learned lever response at the timing of condition stimulus (CS) trial and error (Schultz, Dayan et al. 1997). Results showed that dopamine neurons encoded the error signal of reward prediction; although in the earlier phase dopamine neurons were strongly activated when receiving a juice reward, they were also activated at the timing of CS after learning. Furthermore, omission of a reward resulted in a dip in the activity of dopamine neurons after learning. These dopamine neurons’ behaviors can be explained well by the reinforcement learning model.

What, then, is the mechanism that generates information needed for computing TD error in dopamine neurons, such as reward prediction and reward itself, in the brain? Previous studies proposed a brain mechanism for the reinforcement learning model in reward-based action learning (Houk, Adams et al. 1995; Doya 2000). These models were based on anatomical findings that there are parallel loop organizations between the cerebral cortex, striatum, globus pallidus, thalamus, and cerebral cortex (Alexander, DeLong et al. 1986; Middleton and Strick 2000). We would like to explain the anatomical background and findings of the basal ganglia and parallel loop organization.

1.2.1. Histological structure of the basal ganglia

The striatum receives input from almost all areas of the cerebral cortex. This information is sent to the globus pallidus internal segment and the SNr, thalamus, and is sent back to the cortex again. The striatum consists of the putamen the caudate nucleus, and it is located in the center of the brain. The striatum is the largest sub-cortical structure, measuring about 10 cm^3 in humans. The nucleus accumbens is defined as part of ventral region of the striatum. The human striatum contains a vast number of projection neurons (medium spiny neurons), about 10^8, and a small number of inter-neurons, about 6×10^5. The projection neurons in the striatum receive glutamatergic input from the cerebral cortex and dopaminergic input from the substantial nigra pars compacta (SNc).

Although the putamen and the caudate have similar histological structures, two characteristically different structures, those of the striosome and matrix, are distributed in mosaic-like patterns. The striosome is an embryologically older structure; it is developed under dopaminergic projection from the SNc, and after that the matrix develops. The projection neurons in the striosome send output to the dopamine neurons in the SNc, while the projection neurons in the matrix send output the GABAergic projection neurons in the globus pallidus internal segment and the substantial nigra pars reticulata (SNr). These structures receive input from different cortical areas. Although the striosome receives input mainly from the prefrontal area, the insula, and the amygdala, the matrix receives input from the motor area, the somatosensory area, and the cingulate cortex (Graybiel 1990).

1.2.2. Cortico-basal ganglia loops

In the brain, several cortex-basal ganglia-thalamus-cortex loops work in parallel. These loops are subdivided into five main categories depending on cortical function: The motor loop via the putamen, the oculomotor loop via the caudate, the prefrontal loop via the caudate, the lateral orbitofrontal loop via the posterior part of the putamen and caudate, and the anterior cingulate loop via the ventral part of the striatum. Different parts of the striatum receive input from different cortical areas, and the topography of neural fibers arising from the frontal lobe to the striatum is demonstrated by injection of antidromic tracer into the striatum (Haber, Kunishio et al. 1995). While the orbitofrontal and medial prefrontal cortex send fibers to the ventral part of the striatum, the dorsolateral prefrontal cortex and motor area send fibers to the dorsal part of the striatum.

In a reinforcement learning model of the basal ganglia (Fig. 1, Doya 2000), state *s* from the environment is represented in the cerebral cortex. In the striatum, neurons in the striosome encode the value of state *V*, and neurons in the matrix encode the value Q_1, Q_2 … of action candidates a_1, a_2… Information about *V* is sent to the substantia nigra pars compacta from the striosome, and TD error is computed from *V* and reward *r*. From the action values calculated in the striatum, the optimal action maximizing action value is chosen via the globus pallidus, substantia nigra pars reticulata, thalamus, and finally sent to the spinal cord and other brain areas as the output. TD error δ is sent to the striatal neurons from the SNc, and updates *V* and *Q* by changing the synaptic plasticity of striatal neurons. The reinforcement learning models of the basal ganglia well explain not only dopaminergic activities (Schultz, Dayan et al. 1997; Bayer and Glimcher 2005) but also previous reward-related experimental results in the basal ganglia (Kawagoe, Takikawa et al. 1998; Shidara, Aigner et al. 1998; Tremblay, Hollerman et al. 1998; Samejima, Ueda et al. 2005).

To elucidate the brain mechanism for reward-based learning, it is effective to observe behaviors of brain networks and substance systems by using the basal ganglia model based on reinforcement learning theory.

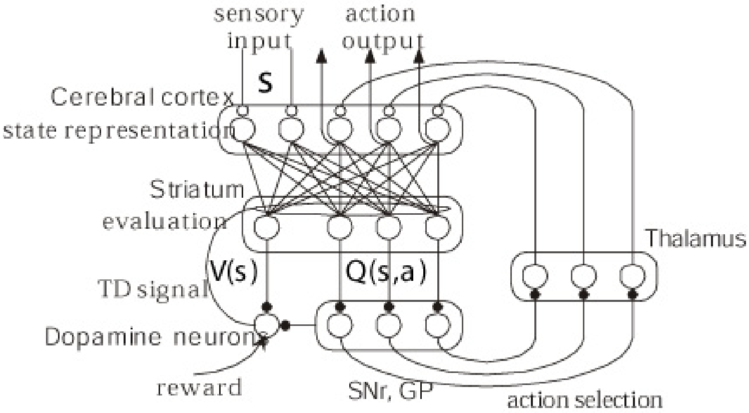

Figure1: A schematic diagram of the cortico-basal ganglia loop (Doya, 2000).

2. Meta-learning and neuromodulators

Based on reinforcement learning theory and previous clinical reports and animal experiments, Doya proposed the “meta-learning hypothesis” (Doya 2002). In reinforcement learning theory, there are meta-parameters, such as learning rate α, inverse temperature β, and discount factor γ. For appropriate learning, these parameters are finely tuned depending on the task setting and environment, but it is a difficult problem. Both humans and animals can learn novel actions under variable environments because there are “meta-learning” systems – controlling meta-parameters – in the brain. Doya suggested that the neuromodulators, like dopamine and serotonin, are candidates for this system, and proposed the following hypotheses.

- Serotonin controls discount factor γ
- Noradrenalin controls inverse temperature of action selection β
- Acetylcholine controls learning rate α

Please refer to Doya (2002) for a detailed explanation of each hypothesis. Here, we focus on the serotonin hypothesis, and explore serotonergic functions in the human reward system by behavior, brain imaging, and pharmacological experiments.

3. Serotonin and the time scale of reward prediction

3.1. Physiology of serotonin

In the central nervous system, serotonin is involved in many kinds of physical and psychological functions including sleep, pain, fear, mood disorders, impulsivity, drug abuse, etc. There are several subtypes of serotonin receptors, e.g. 5-HT$_{1A}$, 5-HT$_{1B}$, 5-HT$_{1D}$, 5-HT$_{2A}$, 5-HT$_{2C}$, 5-HT$_{3}$, 5-HT$_{4}$, 5-HT$_{6}$, and 5-HT$_{7}$. We would like to summarize the features of typical receptors in the human central nervous system. The 5-HT$_{1}$ family (5-HT$_{1A}$, 5-HT$_{1B}$, 5-HT$_{1D}$, 5-ht$_{1E}$, and 5-ht$_{1F}$) comprises G-protein-coupled receptors (coupling with G$_{i}$) and reduces the intracellular cAMP concentration. The 5-HT$_{1A}$ receptors are densely distributed in the limbic areas, such as the hippocampus, the amygdala, and the septum. They are autoreceptors observed in the cell body and dendrite of the post synapse. 5-HT$_{1B}$ receptors are densely distributed in the striatum, globus pallidus, and SNr. 5-HT$_{1D}$ receptors are mainly distributed in the dorsal raphe, striatum, nucleus accumbens, and hippocampus. 5-HT$_{1B/1D}$ receptors modulate serotonin release as autoreceptors in nerve endings.

The 5-HT$_2$ family (5-HT$_{2A}$, 5-HT$_{2B}$ and 5-HT$_{2C}$) conjugates with G protein Gq, and IP3 and DG are generated. IP3 binds to IP3 receptors in the endoplasmic reticulum and results in enhancement of calcium mobilization. 5-HT2A receptors are mainly distributed in the cerebral cortex, hippocampus, and basal ganglia. 5-HT2C receptors are especially densely distributed in the basal ganglia. 5-HT3 receptors are serotonergic cation channels, and their activation increases cell membrane permeability to Na^+, K^+, and Ca^{2+} from the extracellular to the intracellular environment. 5-HT4, 5-HT6, and 5-HT7 receptors, coupling with G protein Gs, increase intracellular cAMP. 5-HT4 receptors are observed in the striatum, nucleus accumbens, globus pallidus, and substantial nigra in high density. These 5-HT receptor subtypes are differentially distributed with different intracellular communication pathways. Previous studies, using agonist, antagonist, and knockout mice of particular receptor subtypes, suggest that each subtype is involved in various mental functions, including depression, mania, ADHD, and impulsivity (Matsui et al., 1997).

3.2. Neural substrates of impulsivity

In reinforcement learning theory, a small value of discount factor γ may cause "impulsive" behavior. In clinical cases, "impulsivity" is one of the prominent symptoms of depression, attention-deficit hyperactivity disorder (ADHD), drug abuse, and a brain lesion of the orbitofrontal cortex (OFC) or medial prefrontal cortex (mPFC) (Ainslie 1975; Evenden 1999). There are two definitions of impulsivity. One is "motor impulsivity", and the other is "cognitive impulsivity". Motor impulsivity can be seen as impairment to motor inhibition in the GO/NO-GO task. Cognitive impulsivity is defined by impulsive choice, abnormal frequent choices of immediate-small rewards against delayed-large rewards. Impulsive choices at small settings of γ happen because the delayed reward is discounted more heavily than the immediate reward. Thus, in our study, we focus on cognitive impulsivity.

Serotonergic involvement in cognitive impulsivity has been suggested in clinical reports and many animal studies; lesions of the serotonergic system produced by selective injection to the dorsal raphe, from which serotonergic projections are sent to broad brain areas, caused impulsive choice in rats (Wogar, Bradshaw et al. 1993; Bizot, Le Bihan et al. 1999; Mobini, Chiang et al. 2000) and injection of serotonin agonist decreased impulsive choices (Poulos, Parker et al. 1996; Bizot, Le Bihan et al. 1999). From these experimental results, functional models of serotonin, where serotonin controls the slope parameter of delay discounting, were proposed in previous studies (Ho, Mobini et al. 1999; Mobini, Chiang et al. 2000).

Lesion of particular brain areas also caused impulsive choices in rats. Cardinal and his colleagues demonstrated that the core of the nucleus accumbens is involved in impulsivity (Cardinal, Pennicott et al. 2001); an AcbC lesioned rat chose immediate-small rewards more frequently than large-delayed rewards delivered after several tens of seconds. Mobini and his colleagues reported that an OFC lesion also caused impulsive choices (Mobini, Body et al. 2002).

4. Aim of this thesis

Although there have been many studies examining delay discounting or impulsivity from various perspectives (e.g., serotonergic function examined following injection of serotonergic neurotoxins, antagonists and agonists or the function of particular brain areas examined following lesions — or in clinical cases — of brain damage), there has been no model systematically explaining these results. Thus, we propose following hypotheses of serotonergic function in delayed reward prediction.

1. There are distinct neural pathways for reward prediction at different time scale

2. These pathways are regulated by serotonergic modulation

We hypothesize that different cortico-basal ganglia loops are involved in reward prediction at different time scales simultaneously, and one of these time scales is chosen by serotonergic modulation on parallel loops and used in actual action selection. The advantage of this parallel organization is that, even in a variable environment, it is not necessary to learn a novel value function with an appropriate time scale for the present environment. These hypotheses can explain previous studies on lesioned brains and serotonergic manipulations: if brain regions, in which there are loops involved in reward prediction at longer time scales, are lesioned, subjects can only predict rewards at shorter time scales. Further, in cases of a low serotonin level, impulsive behaviors may be generated by enhancing loops involved in reward prediction at shorter time scales or by suppressing loops involved in reward prediction at longer time scales.

To explore the relationship between the serotonergic system and the specific brain areas, non-invasive imaging is a powerful tool. Functional magnetic resonance imaging (fMRI) has been used to measure brain activity non-invasively. This method can show the effect of the serotonergic system on brain activity related to reward prediction. To test our hypotheses, we performed a series of experiments as follows.

Experiment 1:

To test the first hypothesis that different brain areas are involved in reward prediction at different time scales, we measured subjects' brain activity while they learned actions related to immediate and delayed monetary rewards. We demonstrated that different brain areas were activated in short-term and long-term reward prediction. By applying model-based analysis using reinforcement learning theory, we found that different cortico-striatum loops were involved in reward prediction and prediction error at different time scales.

Experiment 2:

A shorter time scale of reward prediction may cause a lack of far-sighted behavior. In an unpredictable environment, we also cannot predict distant future rewards. To understand the brain mechanisms involved in reward prediction under predictable and unpredictable environments, we measured brain activities during a Markov decision with regular and random state-transition rules with monetary reward. In predictable condition, brain areas involved in short-term reward prediction were more strongly activated. On the other hand, in unpredictable condition, brain areas involved in long-term reward prediction were more strongly activated. We estimated the value of selected actions using Bayesian estimation from each subject's action sequence, and suggested that different loops are involved in actual action selection and reward prediction at different time scales within the striatum.

Experiment 3:

We demonstrated parallel organization in cortico-striatum loops for reward prediction at different time scales in the above two experiments. In this experiment, to elucidate the effects of serotonin on these parallel mechanisms, we controlled subjects' serotonin levels by dietary tryptophan (precursor of serotonin) manipulation, and measured brain activities during choice tasks between immediate-small against delayed-large juice reward, at different serotonin levels. Using a regression analysis of reward prediction signals, we found that while the activities in the ventral part of the striatum strongly correlated with short-term reward prediction at low serotonin levels, those of the dorsal part strongly correlated with long-term reward prediction at high serotonin levels. The result supports the possibility that serotonin controls the time scale of reward prediction by differentially regulating the activities within the striatum.

Experiment 4:

Impulsive choice, that is, preference for immediate-small reward over delayed-large reward, might be caused by a low setting of the time scale of reward prediction and immature learning of actions that produce a delayed-large reward. To test whether serotonin affects the learning of stimulus-outcome associations based on delayed outcomes, we developed a novel task in which subjects needed to learn the possibly delayed stimulus-outcome association by correctly assigning credit from the present outcome to previously selected stimuli, and analyzed the behavior under central serotonin manipulation. We found significantly slower learning of delayed small punishment compared to delayed large punishment at low serotonin levels. Based on the reinforcement learning model, we estimated subjects' learning parameters that maximize the likelihood of their actions, and found that the estimated trace decay parameter at low serotonin levels was smaller than at high serotonin levels.

We provide a detailed explanation of each experiment in the following chapters.

Chapter 2

Prediction of Immediate and Future Rewards Differentially Recruits Cortico-Basal Ganglia Loops

Evaluation of both immediate and future outcomes of an action is a critical requirement for intelligent behavior. We investigated brain mechanisms for reward prediction at different time scales in an fMRI experiment using a Markov decision task. When subjects learned actions from immediate rewards, significant activity was found in the lateral orbitofrontal cortex and the striatum. When subjects learned to acquire large future rewards despite small immediate losses, the dorsolateral prefrontal cortex, inferior parietal cortex, dorsal raphe nucleus, and cerebellum were also activated. Computational model-based regression analysis using the predicted future rewards and prediction errors estimated from subjects' performance data revealed graded maps of time scale within the insula and the striatum, where ventroanterior parts were responsible for predicting immediate rewards and dorsoposterior parts for future rewards. These results suggest differential involvement of the cortico-basal ganglia loops in reward prediction at different time scales.

1. Introduction

In our daily life, we make decisions based on the prediction of rewards at different time scales; for example, a decision to undertake hard daily exercises to achieve a major future goal, or to resist a sweet temptation if it may lead to a future disaster. Patients with damage in the prefrontal cortex often have trouble in daily decision making, which requires assessment of future outcomes (Bechara, Damasio et al. 2000; Mobini, Body et al. 2002). Lesions in the core of the nucleus accumbens in rats result in the choice of a small immediate reward rather than a larger future reward (Cardinal, Pennicott et al. 2001). Low activity of the central serotonergic system is associated with impulsive behaviors in humans (Rogers, Everitt et al. 1999), and animal experiments have shown that lesions in the ascending serotonergic pathway cause the choice of small immediate rewards as opposed to larger future rewards (Evenden and Ryan 1996; Mobini, Chiang et al. 2000). A possible mechanism underlying these observations is that different sub-loops of the topographically organized cortico-basal ganglia network are specialized for reward prediction at different time scales and that they are differentially activated by the ascending serotonergic system (Doya 2002).

To test whether there are distinct neural pathways for reward prediction at different time scales, we developed a *Markov decision task,* in which an action does not only affect the immediate reward but also the future states and rewards, and we analyzed subjects' brain activities using functional MRI. Recent functional brain imaging studies have shown the involvement of specific brain areas, such as the orbitofrontal cortex (OFC) and the ventral striatum, in prediction and perception of rewards (Berns, McClure et al. 2001; Breiter, Aharon et al. 2001; O'Doherty, Deichmann et al. 2002; O'Doherty, Dayan et al. 2003). However, in previous studies, rewards were given either independent of subject's actions or as a function of the current action. Our Markov decision task probes decision making under a dynamic context with small losses followed by a large positive reward. The results of the block-design analysis suggest differential involvement of brain areas in decision making by prediction of rewards at different time scales. By analyzing subjects' performance data according to a theoretical model of reinforcement learning, we revealed a gradient of activation within the insula and the striatum for prediction of rewards at different time scales.

2. Methods

2.1. Subjects.

Twenty healthy, right-handed volunteers (18 males and 2 females), aged 22 to 34 years gave informed consent to participate in the experiment, which was conducted with the approval of the ethics and safety committees of ATR and Hiroshima University.

2.2. Behavioral task.

In the *Markov decision task* (**Fig. 1**), one of three states is visually presented to the subject using three different figures, and the subject selects one of two actions by pressing one of two buttons using their right hand (**Fig. 1a**). In the SHORT condition (**Fig. 1b**), action a_1 results in a small positive reward $+r_1$ (10, 20, or 30 yen, with equal probabilities), while action a_2 results in a small loss $-r_1$, at any of the three states. Thus, the optimal behavior is to collect small positive rewards at each state by taking action a_1. In the LONG condition (**Fig. 1c**), however, the reward setting is changed so that action a_2 gives a large positive reward $+r_2$ (90, 100, or 110 yen) at state s_3, and action a_1 gives a large loss $-r_2$ at state s_1. Thus, the optimal behavior is to receive small losses at states s_1 and s_2 to obtain a large positive reward at state s_3 by taking action a_2 at each state. There were two control conditions: NO condition, where the reward was always zero, and RANDOM condition, where the reward was positive ($+r_1$) or negative ($-r_1$) with equal probability regardless of state and action.

Subjects performed four trials in a NO condition block, 15 trials in a SHORT condition block, four trials in a RANDOM condition block, and 15 trials in a LONG condition block. A set of four condition blocks (NO, SHORT, RANDOM, LONG) was repeated four times (see **Fig. 3a**). Subjects are informed of the current condition at the beginning of each condition block by a text on the screen, for example, "SHORT condition" (“instruction step”, first slide in **Fig. 1a**); thus, the entire experiment consisted of 168 steps (152 trial steps and 16 instruction steps), taking about 17 minutes. The mappings of the three states to the three figures and the two buttons to the two actions are randomly set at the beginning of each experiment, so that subjects must learn the amount of rewards associated with each figure-button pair in both SHORT and LONG conditions. Furthermore, in the LONG condition, subjects have to learn the subsequent figure for each figure-action pair and take into account the amount of reward expected from the subsequent figure in selecting a button.

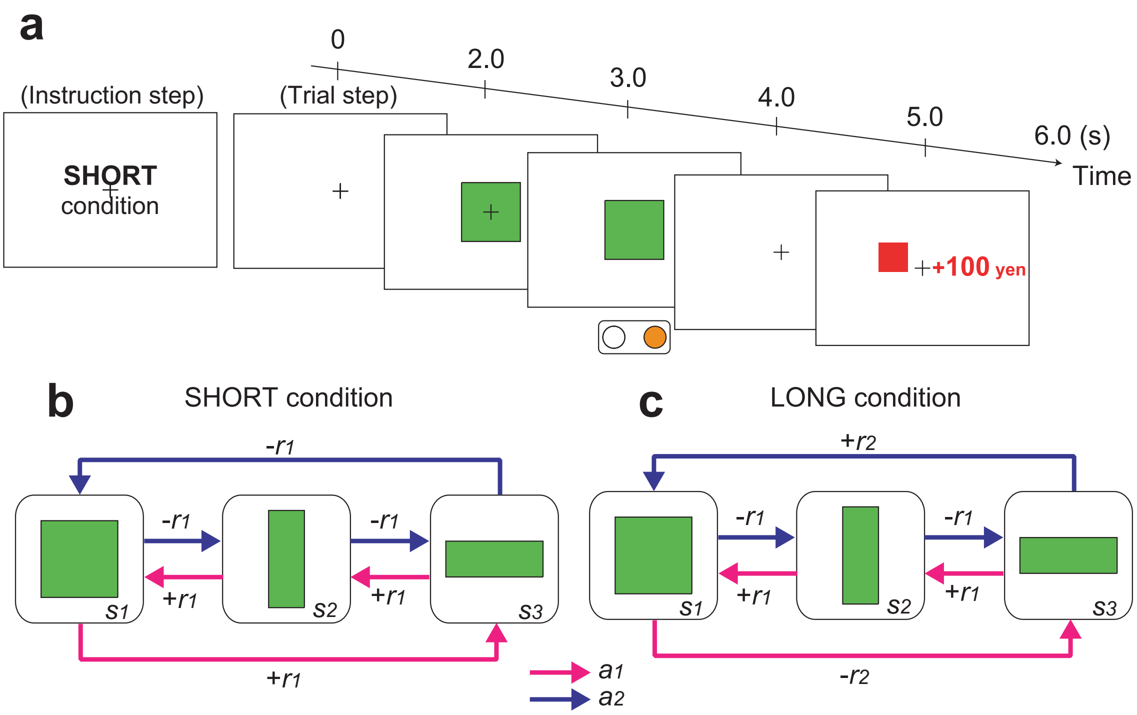

Figure 1: Experimental design. (**a**) Sequences of stimulus and response events in the task. At the beginning of each condition block, the condition is informed by displaying character (6 s), such as the "SHORT condition" (instruction step). In each trial step, a fixation point is presented on the screen, and after 2 seconds, one of three figures (square, vertical rectangle, and horizontal rectangle) is presented. As the fixation point vanishes after 1 second, the subject presses either the right or left button within 1 second. After a short delay (1 s), a reward for the current action is presented by a number and the past cumulative reward is shown by a bar graph. Thus, one trial takes six seconds. (**b** and **c**) The rules of the reward and state transition for action a_1 (red arrow) and action a_2 (magenta arrow) in the SHORT (**b**) and LONG (**c**) conditions. The small reward r_1 is either 10, 20, or 30 yen, with equal probability, and the large reward r_2 is either 90, 100, or 110 yen. The rule of state transition is the same for all conditions; $s_3 \rightarrow s_2 \rightarrow s_1 \rightarrow s_3$ … for action a_1, and $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_1 \rightarrow$ … for action a_2. Although the optimal behaviors are opposite (SHORT: $a_{1,}$ LONG: a_2), the expected cumulative reward during one cycle of the optimal behavior is 60 yen in both the SHORT (+20 × 3) and LONG (– 20 – 20 + 100) conditions.

2.3. fMRI imaging.

A 1.5-Tesla scanner (Shimadzu-Marconi, MAGNEX ECLIPSE, Japan) was used to acquire both structural T1-weighted images (TR = 12 ms, TE = 4.5 ms, flip angle = 20 deg, matrix = 256 × 256, FoV = 256 mm, thickness = 1 mm, slice gap = 0 mm) and T2*-weighted echo planar images (TR = 6 s, TE = 55 ms, flip angle = 90 deg, 50 transverse slices, matrix = 64 × 64, FoV = 192 mm, thickness = 3 mm, slice gap = 0 mm) with blood oxygen level-dependent (BOLD) contrast.

Because the aim of the present study was to specify brain activities for reward prediction in multiple trial steps, we acquired functional images every 6 seconds (TR = 6 s) in synchronization with single trial. Although shorter TR and event-related paradigm are often used in experiments that aim to distinguish brain activities for events within a trial, such as conditioned stimuli, action and reward (Breiter, Aharon et al. 2001; Knutson, Adams et al. 2001; Knutson, Fong et al. 2003; O'Doherty, Dayan et al. 2003), analysis of those finer events in time were not the focus of the current study. With this longer TR, signal in a single scan contains a mixture of responses for reward predictive stimulus and reward feedback. However, because of the progress of learning and the stochasticity of the amount of reward, the time courses of reward prediction $V(t)$ and prediction error $\delta(t)$ over the 168 trial steps were significantly different with each other. Thus, we could separate activities for reward prediction and outcomes by using both reward prediction $V(t)$ and reward outcome $r(t)$ in multiple regression analysis as described below.

2.4. Data analysis.

The data were pre-processed and analyzed with SPM99 (Friston *et al.*, 1995; Wellcome Department of Cognitive Neurology, London, UK). The first two volumes of images were discarded to avoid T1 equilibrium effects. The images were realigned to the first image as a reference, spatially normalized with respect to the Montreal Neurological Institute EPI template, and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum).

We conducted two types of analysis. One was block-design analysis using four boxcar regressors covering the whole experiment convolved with a hemodynamic response function as the reference waveform for each condition (NO, SHORT, RANDOM, and

LONG). We did not find substantial differences between SHORT vs. NO and SHORT vs. RANDOM contrasts, or between LONG vs. NO and LONG vs. RANDOM contrasts. Thus, we report here only the results with the NO condition as the control condition. The other method was multivariate regression analysis using explanatory variables, representing the time course of the reward prediction $V(t)$ or reward prediction error $\delta(t)$ at six different timescales γ, estimated from subjects' performance data (described below).

In both analyses, images of parameter estimates for the contrast of interest were created for each subject. These were then entered into a second-level group analysis using a one-sample t test at a threshold of $P < 0.001$, uncorrected for multiple comparisons (random effects analysis) and extent threshold of 4 voxels. Statistical analysis at $P < 0.05$, correction for a small volume (SVC) was performed using a region of interest (ROI) of the striatum (including the caudate and putamen), which was defined anatomically based on a normalized T1 image.

2.5. Procedures of performance-based regression analysis.

The time course of reward prediction $V(t)$ and reward prediction error $\delta(t)$ were estimated from each subject's performance data, i.e. state $s(t)$, action $a(t)$, and reward $r(t)$, as follows.

Reward prediction: To estimate how much of a forthcoming reward a subject would have expected at each step during the Markov decision task, we took the definition of the value function

$$V(t) = E[r(t + 1) + \gamma r(t + 2) + \gamma^2 r(t + 3) + \ldots] \qquad (1)$$

and reformulated it based on the recursive structure of the task. Namely, if the subject starts from a state $s(t)$ and comes back to the same state after k steps, the expected cumulative reward $V(t)$ should satisfy the consistency condition

$$V(t) = r(t + 1) + \gamma r(t + 2) + \ldots + \gamma^{k-1} r(t + k) + \gamma^k V(t).$$

Thus, for each time t of the data file, we calculated the weighted sum of the rewards acquired until the subject returned to the same state and estimated the value function for that episode as

$$\hat{V}(t) = \frac{\left[r(t+1) + \gamma r(t+2) + \cdots + \gamma^{k-1} r(t+k)\right]}{1-\gamma^k}.$$

The estimate of the value function $V(t)$ at time t was given by the average of all previous episodes from the same state as at time t

$$V(t) = \frac{1}{L}\sum_{l=1}^{L} \hat{V}(t_l),$$

where $\{t_1, \ldots, t_L\}$ are the indices of time visiting the same state as $s(t)$, i.e. $s(t_1) = \ldots = s(t_L) = s(t)$.

Reward prediction error: the TD error

$$\delta(t) = r(t) + \gamma V(t) - V(t-1), \tag{2}$$

was calculated from the difference between the actual reward $r(t)$ and the temporal difference of the estimated value function $V(t)$.

We separately calculated the time courses of $V(t)$ and $\delta(t)$ during SHORT and LONG conditions; we concatenated data of four blocks in the SHORT condition, and calculated $V(t)$ and $\delta(t)$ as described above. We used the same process for the LONG condition data. During the NO and RANDOM conditions, the values of $V(t)$ and $\delta(t)$ were fixed to zero. Finally, we reconstructed the data corresponding to the real time course of experiment. Examples of the time course of these variables are shown in Figure 2. We used one of these, $V(t)$ and $\delta(t)$, as the explanatory variable in a regression analysis by SPM. To remove any effects of factors other than reward prediction, concurrently we used possibly relevant explanatory variables, namely the four box-car functions representing each condition (NO, SHORT, RANDOM, and LONG). Because the immediate reward prediction $V(t)$ with $\gamma = 0$ and reward outcome $r(t)$ can coincide if learning is perfect, we included the reward outcome $r(t)$ in regression analysis with $V(t)$.

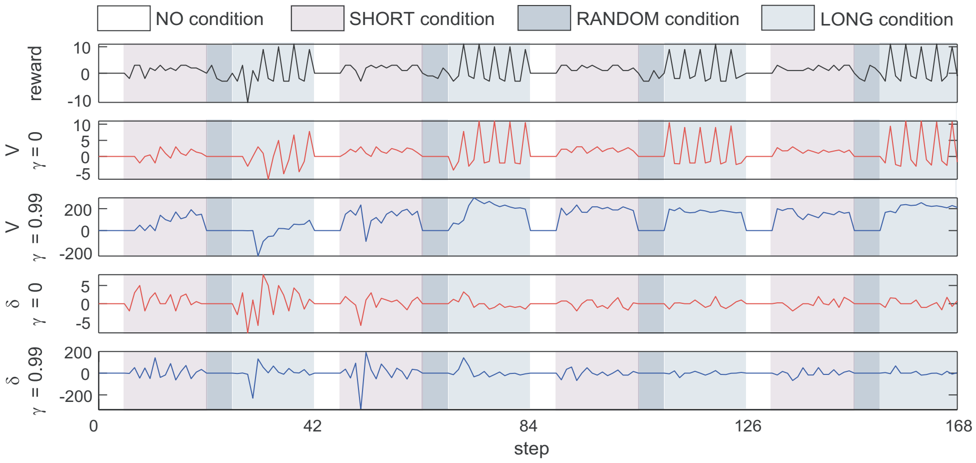

Figure 2: An example of the time series of the explanatory variables for one subject. Waveforms of the explanatory variables showing reward outcome *r*, reward prediction *V* (γ = 0 and 0.99), and reward prediction error δ (γ = 0 and 0.99) of one subject. It can be seen that the time series for both *V* and δ at γ = 0 reflect short timescale tracking of *r*, whereas at γ = 0.99 it reflects long timescale tracking of *r*(*t*). Line colors correspond to the color map of γ shown in Figure 6. In early blocks, reward prediction *V* was low and variable, and δ fluctuated widely. As the task proceeded, the level of *V* increased and δ converged toward zero. Differences in the time courses of *V* at different values of γ can also be seen. With a small γ (= 0), the time course of *V* had sharp peaks representing immediate rewards, while at a large γ (= 0.99), *V* was stable in both SHORT and LONG conditions.

Thus, the significant correlation with *V*(*t*) (**Fig. 7a** and **b**) should represent a predictive component rather than a reward outcome.

The amplitude of explanatory variables $\delta(t)$ with all γ had a decreasing trend (**Fig. 2**). This causes a risk that areas that are activated early in trials, e. g. those responsible for general attentiveness or novelty, have correlations with $\delta(t)$. Because our aim in regression analysis was to clarify the brain structures for reward prediction at specific time scales, we removed the areas that had similar correlation to $\delta(t)$ at all settings of γ from considerations in Figure 7 and Table 3.

To compare the results of regression analysis with six different values of γ, we used display software that can overlay multiple activation maps in different colors on a single brain structure image. When a voxel is significantly activated in multiple values of γ, it is shown by a mosaic of multiple colors, with apparent subdivision of the voxel (**Fig. 7**).

3. Results

3.1. Behavioral results

In the *Markov decision task* (**Fig. 1**; see Methods for details), one of three states is visually presented to the subject using three different figures, and the subject selects one of two actions by pressing one of two buttons (**Fig. 1a**). For each state, the subject's action affects not only the reward given immediately but also the state subsequently presented (**Fig. 1b** and **c**).

While the rule of state transition is fixed during the entire experiment, the rules of reward delivery are changed according to task conditions. In the SHORT condition, action a_1 gives a small positive reward $+r_1$ (20 yen average) and action a_2 gives a small loss $-r_1$ at all three states (**Fig. 1b**). The optimal behavior for maximizing the total outcomes is to collect small positive rewards by taking action a_1 at each state. In the LONG condition, while action a_2 at state s_3 gives a big bonus $+r_2$ (100 yen average), action a_1 at state s_1 results in a big loss $-r_2$ (**Fig. 1c**). The optimal behavior is to receive small losses at state s_1 and s_2 to obtain a large positive reward at state s_3 by taking action a_2 at each state, opposite to the optimal behavior in the SHORT condition; this behavior produces a net positive outcome during one cycle. Thus in the LONG condition, the subject has to select an action by taking into account both the immediate reward and the future reward expected from the subsequent state, while the subjects need to consider only the immediate outcome in the SHORT condition. Subjects performed 15 trials in a SHORT condition block and 15 trials in a LONG condition block; four condition blocks were performed (see Methods for Behavioral task and **Fig. 3a** for Task schedule).

All subjects successfully learned the optimal behaviors: taking action a_1 in the SHORT condition (**Fig. 3b**) and action a_2 in the LONG condition (**Fig. 3c**). Cumulative rewards within each 15 trials in the SHORT (Fig. 2d) and LONG (Fig. 2e) conditions also indicate successful learning. It can be seen from the single subject data in the LONG condition (**Fig. 3e**, orange) that the subject learned to lose small amounts ($-r_1$) twice to get a big bonus ($+r_2$). The average cumulative reward in the last block was 254 yen in the SHORT condition and 257 yen in the LONG condition, which was 84.7 % and 85.7 %, respectively, of the theoretical optimum of 300 yen.

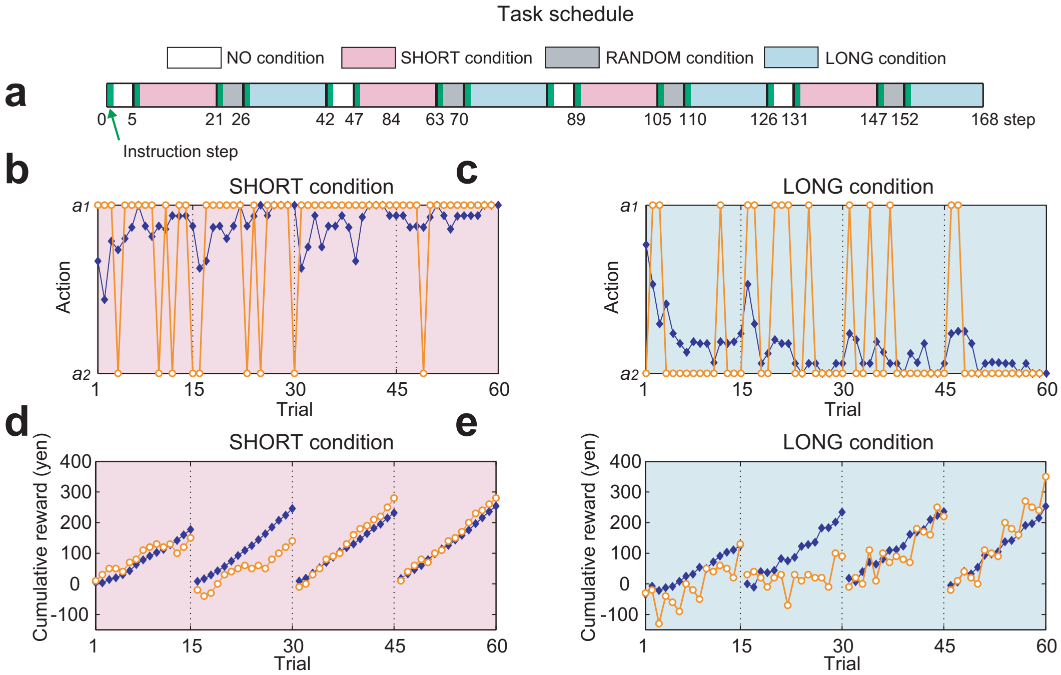

Figure 3: Task schedule and behavioral results. (**a**) A set of four condition blocks, NO (four trials), SHORT (15 trials), RANDOM (four trials), and LONG (15 trials), is repeated four times. At the beginning of each condition block, the task condition is presented to the subject (instruction step); thus, the entire experiment consisted of 168 steps (152 trial steps and 16 instruction steps). (**b** and **c**) The selected action of a representative single subject (orange) and the group average ratio of selecting a_1 (blue) in the (**b**) SHORT and (**c**) LONG conditions. (**d** and **e**) The accumulated reward in each block of a representative single subject (orange) and the group average (blue) in the (**d**) SHORT and (**e**) LONG conditions. To clearly show the learning effects, data from four trial blocks in the SHORT and LONG conditions are concatenated, with the dotted lines indicating the end of each condition block.

3.2. Block-design analysis

First, in order to find the brain areas that are involved in immediate reward prediction, we compared brain activity during the SHORT condition and the NO condition, in which reward was always zero. In the SHORT vs. NO contrast, a significant increase in activity was observed in the lateral OFC (**Fig. 4a**), the insula and the occipitotemporal area (OTA) (**Fig. 4b**), as well as in the striatum, the globus pallidus (GP) (**Fig. 4c**) and

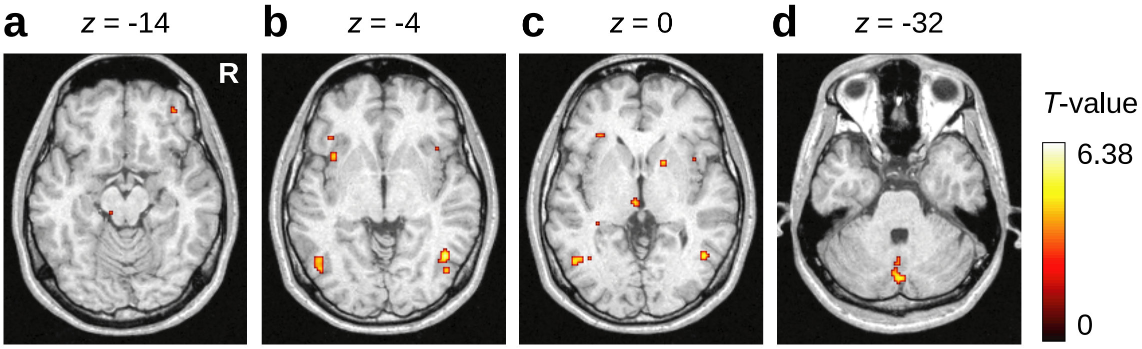

Figure 4: Brain areas activated in the SHORT vs. NO contrast (p < 0.001, uncorrected; extent threshold of 4 voxels). (**a**) Lateral OFC. (**b**) Insula. (**c**) Striatum. (**d**) Medial cerebellum.

the medial cerebellum (**Fig. 4d**) (threshold of $P < 0.001$, uncorrected for multiple comparisons). These areas may be involved in reward prediction that only takes into account an immediate outcome. In the LONG condition, subjects need to predict both immediate and future rewards for optimal actions. Thus, in order to reveal the areas that are specific to future reward prediction, we compared the brain activity during LONG and SHORT conditions. In the LONG vs. SHORT contrast, a robust increase in activity was observed in the ventrolateral prefrontal cortex (VLPFC), the insula, the dorsolateral prefrontal cortex (DLPFC), the dorsal premotor cortex (PMd), the inferior parietal cortex (IPC) (**Fig. 5a**), the striatum, GP (**Fig. 5b**), the dorsal raphe nucleus (**Fig. 5c**), the lateral cerebellum (**Fig. 5d**), the posterior cingulate cortex, and the subthalamic nucleus ($P < 0.001$, uncorrected). Especially, activations in the striatum were highly significant (threshold at $P < 0.05$, corrected for a small volume when using the region of interest of the striatum anatomically defined). These areas are specifically involved in decision making based on the prediction of reward in multiple steps in the future. In the LONG vs. NO contrast, the activated areas were approximately the union of the areas activated in the SHORT vs. NO and LONG vs. SHORT contrasts. These results were consistent with our expectation that both immediate and future reward prediction were required in the LONG condition. The results of block-design analysis, including the LONG vs. NO contrast, are summarized in the Table 1. Activities in both SHORT and LONG conditions were stronger in the first two blocks, when subjects were involved in active trial and error, than in the last two blocks when the subjects' behaviors became repetitive.

We compared the activities in the SHORT vs. NO contrast and the LONG vs. SHORT contrast in three regions (Fig. 6); namely the lateral prefrontal cortex (lateral OFC and

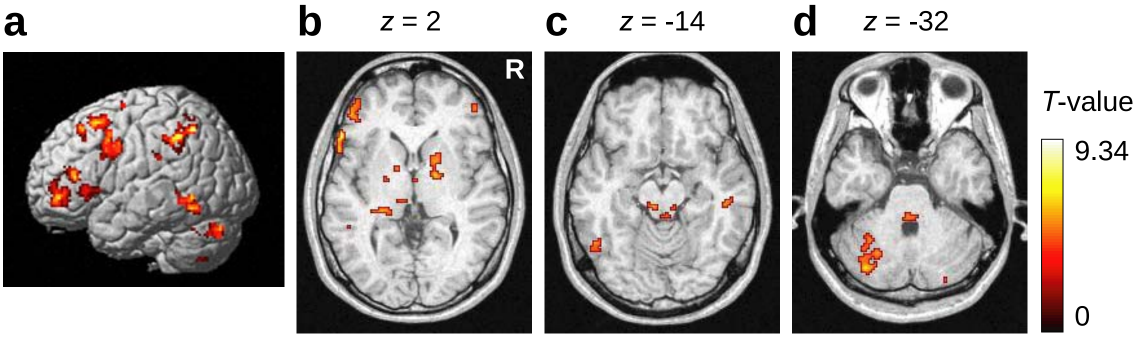

Figure 5: Brain areas activated in the LONG vs. SHORT contrast ($p < 0.0001$, uncorrected; extent threshold of 4 voxels for illustration purposes). (**a**) DLPFC, IPC, PMd. (**b**) GP, striatum. (**c**) Dorsal raphe nucleus. (**d**) Left lateral cerebellum.

VLPFC), the insula, and the anterior striatum, where significant activities were found in both contrasts. In the lateral PFC (**Fig. 6a**), although the activities in lateral OFC for the SHORT vs. NO contrast (red) and in the VLPFC for the LONG vs. SHORT contrast (blue) were close in location, they were clearly apart on the cortical surface. Activities in the insula were also separated (**Fig. 6b**). In the anterior striatum (**Fig. 6c**), we found limited overlaps between the two contrasts (green). In all three areas, activities in the SHORT vs. NO contrast were found in the ventral parts, while activities in the LONG vs. SHORT contrast were found in the dorsal parts. These results of block-design analysis suggest differential involvement of brain areas in predicting immediate and future rewards.

3.3. Performance-based multiple regression analysis

In order to further clarify the brain structures specific to reward prediction at different time scales, we estimated how much reward the subjects should have predicted on the basis of their performance data and used their time courses as the explanatory variables of regression analysis. We took the theoretical framework of temporal difference (TD) learning (Sutton 1998), which has been successfully used for explaining reward-predictive activities of the midbrain dopaminergic system as well as those of the cortex and the striatum (Houk, Adams et al. 1995; Schultz, Dayan et al. 1997; Doya 2000; Berns, McClure et al. 2001; McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003). In TD learning theory, the predicted amount of future reward starting from a state $s(t)$ is formulated as the "value function"

$$V(t) = E[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \ldots] \qquad (1)$$

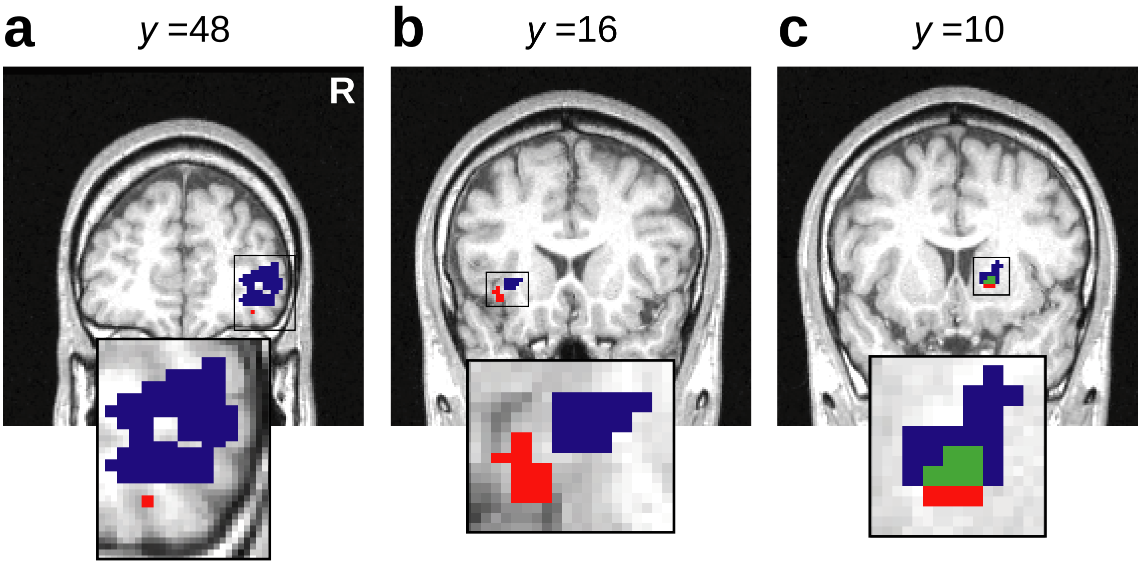

Figure 6: Comparison of brain areas activated in the SHORT vs. NO contrast (red) and the LONG vs. SHORT contrast (blue). These figures show activation maps focused on (**a**) the lateral OFC (red: (x, y, z) = (38, 46, -14), blue: (46, 47, 3)), (**b**) the insula (red: (-36, 13, -4), blue: (-30, 18, 1)), and (**c**) the striatum (red: (18, 10, 0), blue: (18, 12, 3)) where we observed significant activation in both contrast. The overlapped area is indicated in green.

Any deviation from the prediction is given by the TD error

$$\delta(t) = r(t) + \gamma V(t) - V(t-1), \qquad (2)$$

which is a crucial learning signal for reward prediction and action selection. The "discount factor" γ ($0 \leq \gamma < 1$) controls the time scale of prediction; while only the immediate reward $r(t+1)$ is considered with $\gamma = 0$, rewards in the longer future are taken into account with γ closer to 1.

We estimated the time courses of reward prediction $V(t)$ and prediction error $\delta(t)$ from each subject's performance data and used them as the explanatory variables in multiple regression analysis with fMRI data (see Methods). In our Markov decision task, the minimum value of γ needed to find the optimal action in the LONG condition is 0.36, while any small value of γ is sufficient in the SHORT condition. From the results of block-design analysis, we assumed that different cortico-basal ganglia network are specialized for reward prediction at different time scales and that they work in parallel, depending on the requirement of the task. Thus, we varied the discount factor γ as 0, 0.3, 0.6, 0.8, 0.9, and 0.99: small γ for immediate reward prediction and large γ for long future reward prediction. An example of these time courses is shown in Figure 2.

We observed a significant correlation with reward prediction $V(t)$ in the medial prefrontal cortex (mPFC: including the anterior cingulate cortex (ACC) and the medial OFC) (**Fig. 7a**) and bilateral insula (**Fig. 7b**), left hippocampus, and left temporal pole ($P < 0.001$, uncorrected; see Table 2). These figures show the correlated voxels within these areas using a gradient of colors for different discount factor γ (red for $\gamma = 0$, blue for $\gamma = 0.99$). The activities of the mPFC, temporal pole, and hippocampus correlated with reward prediction with a longer time scale ($\gamma \geq 0.6$). Furthermore, in the insula, we found a graded map of activities for reward prediction at different time scales (**Fig. 7b**). While the activity in the ventroanterior part correlated with reward prediction at a shorter time scale, the activity in the dorsoposterior part correlated with reward prediction at a longer time scale.

We also found significant correlation with reward prediction error $\delta(t)$ with a wide range of time scale in the basal ganglia (**Fig. 7c**) ($P < 0.001$, uncorrected; see Table 3 and Methods). Again, we found a graded map, which had a short time scale in the ventroanterior part and a long time scale in the dorsoposterior part.

The red and blue lines in Figure 7b and c show the vertical positions of activity peaks in the SHORT vs. NO and LONG vs. SHORT contrasts, respectively, in the insula and the striatum (**Fig. 6b** and **c**). The coincidence of the ventroanterior-dorsoposterior maps and the ventroanterior-dorsoposterior shifts in activities indicate that, while the ventroanterior parts with smaller γ were predominantly active in the SHORT condition, the dorsoposterior parts with larger γ became more active in the LONG condition.

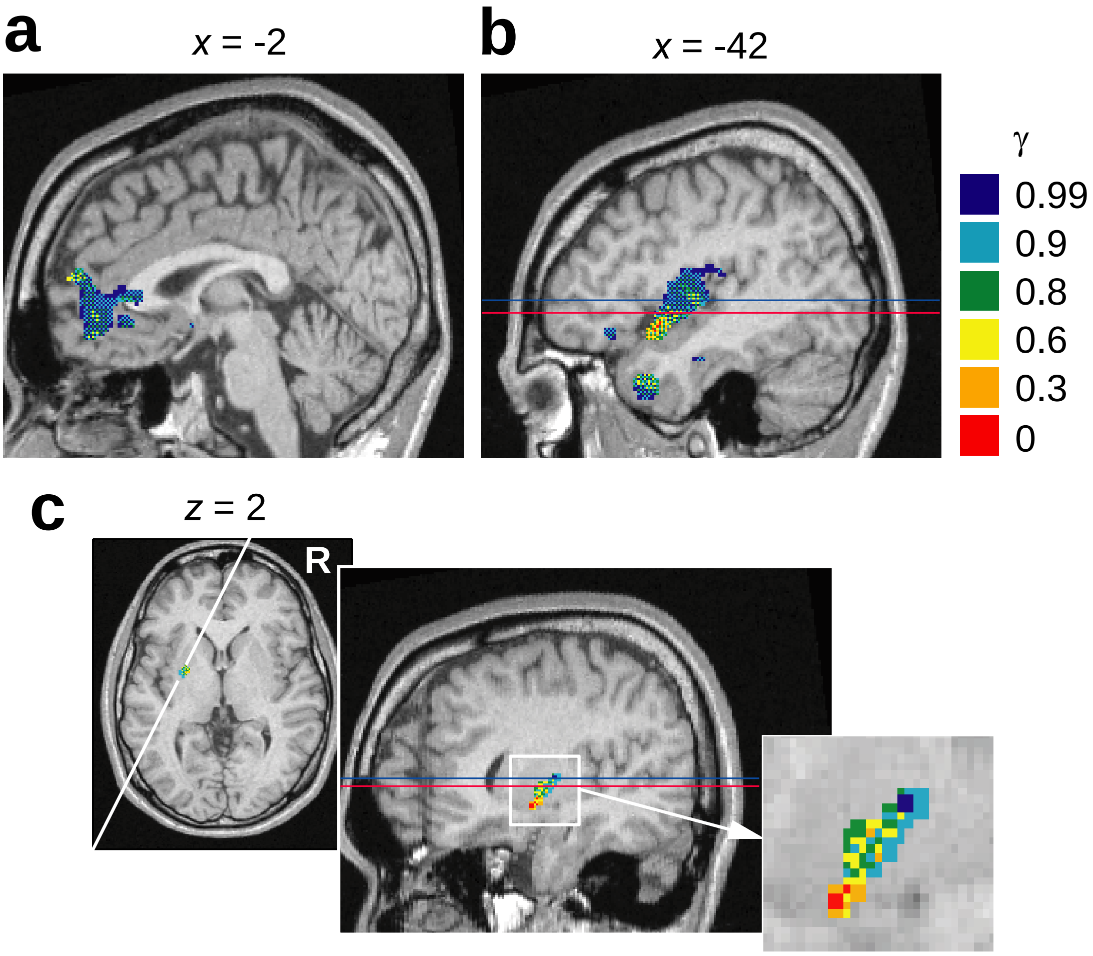

Figure 7: Voxels with a significant correlation (height threshold of p < 0.001, uncorrected; extent threshold of 4 voxels) with reward prediction $V(t)$ and prediction error $\delta(t)$ are shown in different colors for different settings of the discount factor. Voxels correlated with two or more regressors are shown by a mosaic of colors. (**a** and **b**) Significant correlation with reward prediction $V(t)$. (**a**) mPFC. (**b**) Insula. (**c**) Significant correlation with reward prediction error $\delta(t)$ restricted to region of interest of the striatum (slice at white line in horizontal slice at z = 2 mm). Note the ventroanterior to dorsoposterior gradient with the increase in γ both in the insula and the striatum. Red and blue lines correspond to the z-coordinate levels of activation peaks in the insula and striatum shown in Figure 5b and c (red for the SHORT vs. NO and blue for the LONG vs. SHORT contrasts).

4. Discussion

4.1. Consistency between results of block-design and model-based regression analyses

The results of the block-design and performance-based regression analyses suggest differential involvement of brain areas in action learning by prediction of rewards at different time scales. Both block-design and performance-based regression analyses found activities in the insula and the anterior striatum. Activations of the ventral part in the SHORT vs. NO contrast and the dorsal part in the LONG vs. SHORT contrast in each area (**Fig. 6**) are consistent with the ventroanterior-dorsoposterior maps of the discount factor γ found in performance-based regression analysis (**Fig. 7**).

4.2. Gradient maps of time scale of reward prediction in the insula and striatum

The insula takes a pivotal position in reward processing by receiving primary taste and visceral sensory input (Mesulam and Mufson 1982) and sending output to the OFC (Cavada, Company et al. 2000) and the striatum (Chikama, McFarland et al. 1997). Previous studies showed that the insula is activated with anticipation of primary reward (O'Doherty, Deichmann et al. 2002) and that insular lesion causes deficits in incentive learning for primary reward (Balleine and Dickinson 2000). Our results confirm the role of the insula in prediction of non-primary, monetary reward (Knutson, Fong et al. 2003), and further suggest heterogeneous organization within the insula. Previous imaging studies also showed involvement of the insula, especially ventroanterior part, in processing of aversive outcomes (O'Doherty, Critchley et al. 2003; Ullsperger and von Cramon 2003). Thus a possible interpretation of the activation of the insula in LONG condition is that it was due to the losses that subjects acquired before getting a large reward. However, we also ran a regression analysis using losses and found significant correlation in ventroanterior part of insula. Anatomical and physiological studies of insula also showed involvement of its ventroanterior part in perception of aversive stimuli (Mesulam and Mufson 1982). Thus we argue that the activation of dorsoposterior insula is not simply due to losses in LONG condition.

Previous brain imaging and neural recording studies suggest a role for the striatum in prediction and processing of reward (Schultz, Dayan et al. 1997; Koepp, Gunn et al. 1998; Elliott, Friston et al. 2000; Breiter, Aharon et al. 2001; Knutson, Adams et al.

2001; O'Doherty, Deichmann et al. 2002; Pagnoni, Zink et al. 2002; Elliott, Newman et al. 2003; Knutson, Fong et al. 2003; Haruno, Kuroda et al. 2004). Consistent with previous fMRI studies (Berns, McClure et al. 2001; McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003), our results showed striatal activities correlated with the error of reward prediction. The reinforcement learning models of the basal ganglia (Houk, Adams et al. 1995; Schultz, Dayan et al. 1997; Doya 2000) posit that the striatum learns reward prediction and action selection based on the reward prediction error $\delta(t)$ represented by the dopaminergic input. Correlation of the striatal activity with reward prediction error $\delta(t)$ could be due to dopamine-dependent plasticity of cortico-striatal synapses (Reynolds and Wickens 2002).

4.3. Possible roles of other activated areas in reward prediction

In lateral OFC, DLPFC, PMd, IPC, and dorsal raphe, we found significant activities in the block-design analyses, but there was not strong correlation in regression analyses. This may be because these areas perform functions that are helpful for reward prediction and action selection, but their activities do not directly represent the amount of predicted reward or prediction error at a specific time scale.

In reinforcement learning theory, an optimal action selection is realized by taking the action a that maximizes the 'action value' $Q(s, a)$ at a given state s. The action value is defined as

$$Q(s, a) = E[\, r(s, a) + \gamma V(s'(s, a))] \qquad (3)$$

and represents the expected sum of the immediate reward $r(s, a)$ and the weighted future rewards $V(s'(s, a))$, where $s'(s, a)$ means the next state reached by taking an action a at a state s (Sutton 1998; Doya 2000). According to this framework, we can see that prediction of immediate reward $r(s, a)$ is helpful for action selection based on rewards at either short or long time scales, i.e. with any value of discount factor γ. On the other hand, prediction of state transition $s'(s, a)$ is helpful only in long-term reward prediction with positive γ.

In the lateral OFC, we observed significant activity in both the SHORT vs. NO and the LONG vs. NO contrasts (Table 1), but no significant correlation with reward prediction $V(t)$ or reward prediction error $\delta(t)$ in regression analysis. This suggests that the lateral

OFC takes the role of predicting immediate reward $r(s, a)$, which is used for action selection in both SHORT and LONG conditions, but not in the NO condition. This interpretation is consistent with previous studies demonstrating the OFC's role in prediction of rewards, immediately following sensorimotor events (Tremblay and Schultz 2000; Critchley, Mathias et al. 2001), and action selection based on reward prediction (Rogers, Owen et al. 1999; Rolls 2000; O'Doherty, Critchley et al. 2003).

In the DLPFC, PMd, and IPC, there were significant activities in both the LONG vs. NO and the LONG vs. SHORT contrasts (Table 1) but no significant correlation with either $V(t)$ or $\delta(t)$. A possible interpretation is that this area is involved in prediction of future state $s'(s, a)$ in the LONG condition but not in the SHORT or NO conditions. This interpretation is consistent with previous studies showing the role of these cortical areas in imagery (Hanakawa, Honda et al. 2002), working memory and planning (Baker, Rogers et al. 1996; Owen, Doyon et al. 1996).

The dorsal raphe nucleus was activated in the LONG vs. SHORT contrast, but not correlated with $V(t)$ or $\delta(t)$. In consideration of its serotonergic projection to the cortex and the striatum and serotonin's implication with behavioral impulsivity (Evenden and Ryan 1996; Rogers, Everitt et al. 1999; Mobini, Chiang et al. 2000), a possible role for the dorsal raphe nucleus is to control the effective time scale of reward prediction (Doya 2002). Its higher activity in the LONG condition, where a large setting of γ is necessary, is consistent with this hypothesis.

4.4. Possible neural mechanism for reward prediction at different time scales

Let us consider the present experimental results in light of the anatomy of cortico-basal ganglia loops. The cortex and the basal ganglia both have parallel loop organization, with four major loops (limbic, cognitive, motor, and oculomotor) and finer, topographic sub-loops within each major loop (Middleton and Strick 2000). Our results suggest that the areas within the limbic loop (Haber, Kunishio et al. 1995), namely the lateral OFC and ventral striatum, are involved in immediate reward prediction. On the other hand, areas within the cognitive and motor loops (Middleton and Strick 2000), including the DLPFC, IPC, PMd, and dorsal striatum, are involved in activated in future reward prediction. The connections from the insula to the striatum are topographically organized, with the ventral/anterior, agranular cortex projecting to the ventral striatum and the dorsal/posterior, granular cortex projecting to the dorsal striatum (Chikama, McFarland et al. 1997). The graded maps shown in Figure 6b and c are consistent with this topographic cortico-striatal organization and suggest that areas that project to the more dorsoposterior part of the striatum are involved in reward prediction at a longer time scale. These results are consistent with the observations that localized damages within the limbic and cognitive loops manifest as deficits in evaluation of future rewards (Eagle, Humby et al. 1999; Bechara, Damasio et al. 2000; Rolls 2000; Cardinal, Pennicott et al. 2001; Pears, Parkinson et al. 2003) and learning of multi-step behaviors (Hikosaka, Nakahara et al. 1999). The parallel learning mechanisms in the cortico-basal ganglia loops used for reward prediction at a variety of time scales may have the merit of enabling flexible selection of a relevant time scale appropriate for the task and the environment at the time of decision making.

A possible mechanism for selection or weighting of different cortico-basal ganglia loops with an appropriate time scale is serotonergic projection from the dorsal raphe nucleus (Doya 2002), which was activated in the LONG vs. SHORT contrast. Although serotonergic projection is supposed to be diffuse and global, differential expression of serotonergic receptors in the cortical areas and in the ventral and dorsal striatum (Mijnster, Raimundo et al. 1997; Compan, Segu et al. 1998) would result in differential modulation. The mPFC, which had significant correlation with reward prediction $V(t)$ at long time scales ($\gamma \geq 0.6$), may regulate the activity of the raphe nucleus through reciprocal connection (Celada, Puig et al. 2001; Martin-Ruiz, Puig et al. 2001). This interpretation is consistent with previous studies using experimental tasks that require long-range prospects for problem solving, such as the gambling problem (Bechara, Damasio et al. 2000) or delayed reward task (Mobini, Body et al. 2002), that showed involvement of the medial OFC. Future studies using the Markov decision task under pharmacological manipulation of the serotonergic system should clarify the role of serotonin in regulating the time scale of reward prediction.

Recent brain imaging and neural recording studies reported involvement of a variety of cortical areas and the striatum in reward processing (Koepp, Gunn et al. 1998; Rogers, Owen et al. 1999; Elliott, Friston et al. 2000; Hikosaka and Watanabe 2000; Berns, McClure et al. 2001; Breiter, Aharon et al. 2001; Critchley, Mathias et al. 2001; Knutson, Adams et al. 2001; O'Doherty, Deichmann et al. 2002; Pagnoni, Zink et al. 2002; Shidara and Richmond 2002; Elliott, Newman et al. 2003; Knutson, Fong et al. 2003; Matsumoto, Suzuki et al. 2003; McClure, Berns et al. 2003; O'Doherty, Critchley et al. 2003; O'Doherty, Dayan et al. 2003; Haruno, Kuroda et al. 2004). Although some neural recording studies have used experimental tasks that require multiple trial steps for getting rewards (Hikosaka and Watanabe 2000; Shidara and Richmond 2002), none of the previous functional brain imaging studies addressed the issue of reward prediction at different time scales, and considered only rewards immediately following stimuli or actions. We could extract specific functions of OFC, DLPFC, mPFC, insula and cortico-basal ganglia loops by developing a novel Markov decision task and a reinforcement learning model-based regression analysis method. Our regression analysis not only extracted brain activities specific to reward prediction, but also revealed a novel topographic organization in reward prediction (**Fig. 7**). The combination of our Markov decision task with event-related fMRI and magnetoencephalography (MEG) should further clarify the functions used for reward prediction and perception at different time scales, and at finer spatial and temporal resolutions.

Table 1. **Areas significantly activated in the block-design analysis.**

	SHORT vs. NO		LONG vs. NO		LONG vs. SHORT	
	Area (BA)	T-value (Tal x, y, z)	Area (BA)	T-value (Tal x, y, z)	Area (BA)	T-value (Tal x, y, z)
Cerebral cortex	lOFC (11)	3.98 (38, 46, -14)	lOFC (11)	5.86 (-46, 50, -8)	VLPFC (10)	6.71 (-48, 43, -2)
	Insula (13)	4.43 (-36, 13, -4)		5.69 (42, 46, -11)		6.06 (46, 47, 3)
	OTA (37)	5.15 (-48, -62, 1)	DLPFC (46)	5.92 (-42, 35, 9)	DLPFC (46)	5.77 (-40, 35, 9)
		6.32 (46, -68, -3)	Insula (13)	4.99 (-34, 19, -8)	Insula (13)	4.93 (-30, 18, 1)
			mPFC (9)	7.54* (4, 40, 26)	PCC (23)	6.36 (-10, -26, 29)
			PMd (6)	6.56 (-40, 5, 27)	PMd (6)	6.75 (-42, 2, 31)
				6.68 (38, 3, 24)	IPC (40)	6.79 (-49, -43, 41)
			IPC (40)	6.07 (-55, -21, 40)		6.06 (48, -41, 35)
				6.61 (51, -31, 40)		
Basal ganglia	Putamen	4.54 (18, 10, 0)	Putamen	7.72* (14, 10, 0)	Putamen	5.87† (18, 12, 3)
	Medial GP	3.96 (-16, -10, -8)	Caudate head	7.89* (-4, 4, -2)		5.99† (-12, 0, 4)
			Lateral GP	7.69* (-20, -8, 0)	Lateral GP	6.38† (-20, -8, 0)
					STN	5.13 (-8, -12, -4)
Brainstem					Dorsal raphe	5.27 (4, -35, -10)
Cerebellum	Vermis	4.98 (0, -75, -23)	Vermis	6.3 (0, -60, -26)	Hemisphere	8.05* (14, -52, -39)
			Hemisphere	6.94 (-34, -69, -20)		7.49* (-28, -62, -31)

*p < 0.05, corrected for the whole brain.

†p < 0.05, corrected for a small volume restricted to the striatum.

All other regions are p < 0.001, uncorrected for the whole brain.

Extent threshold of 4 voxels.

The numbers in parentheses show the Brodmann area (BA).

Abbreviations: Tal, Talairach coordinates; lOFC, lateral orbitofrontal cortex; OTA, occipitotemporal area; PCC, posterior cingulate cortex; STN, subthalamic nucleus; VLPFC, ventrolateral prefrontal cortex.

Table 2. **Areas with significant correlation with reward prediction *V*(*t*) estimated with different discount factors *γ*.**

	$\gamma = 0$	$\gamma = 0.3$	$\gamma = 0.6$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
Area (BA)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value Tal (x, y, z)
Insula cortex (13)	4.02 (–42,7,–9)	4.84 (–42,7,–10)	6.06 (–42,4,–7)	7.04 (–42,–10,4)	7.84* (–42,–10,4)	7.82* (–42,–10,4)
		4.24 (44,–4,–5)	5.5 (42,–4,–3)	6.53 (42,–2,–3)	6.74 (42,–2,–3)	6.73 (40,–4,–1)
mPFC / ACC (11/9/32/24)			4.45 (–2,46,–16)	6.3 (–6,44,–6)	7.41 (–4,44,–6)	7.79* (–4,44,–6)
Hippocampus			3.66 (–30,–18,–14)	4.57 (–30,–20,–16)	4.79 (–30,–20,–16)	4.84 (–30,–20,–16)
Temporal pole (38)			5.01 (–44,10,–31)	5.42 (–44,10,–31)	5.26 (–44,10,–32)	5.01 (–44,10,–32)

*$p < 0.05$, corrected for the whole brain.

All other regions are $p < 0.001$, uncorrected for the whole brain.

Extent threshold of 4 voxels.

Table 3. **Voxels with significant correlation with reward prediction error *δ*(*t*) estimated with different discount factors *γ*.**

	$\gamma = 0$	$\gamma = 0.3$	$\gamma = 0.6$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)	T-value (Tal x, y, z)
Putamen	4.55 (–24, 4, –7)	4.58 (–26, 5, –9)	4.44 (–26, 2, –2)	4.66 (–26, 2, –2)	5.23 (–28, –6, 6)	4.15 (–28, –6, 6)

All areas are $p < 0.001$, uncorrected for the whole brain.

Extent threshold of 4 voxels.

Chapter 3

Brain mechanism of reward prediction in predictable and unpredictable environment

To understand the brain mechanisms involved in reward prediction under predictable and unpredictable environments, we measured brain activities during a Markov decision with regular and random state transition rules. In predictable condition, the dorsolateral prefrontal cortex and dorsal striatum were more strongly activated. On the other hand, in unpredictable condition, the orbitofrontal cortex and ventral striatum were more strongly activated. By a regression analysis with a reinforcement learning model, we reconfirmed that ventral parts of the striatum and insula were involved in reward prediction at smaller time scales, while dorsal parts were involved in reward prediction at longer time scales (Tanaka et al., 2004). We estimated the value of selected actions using Bayesian estimation from each subject’s action sequence, and we found a significant correlation in the medial prefrontal cortex (mPFC), caudate head, and globus pallidus. These results suggest a role of mPFC in integrating parallel cortico-basal ganglia loops specialized for reward prediction at different time scales.

1. Introduction

In our daily life, we make decisions based on the prediction of outcomes of actions in the given environment. If the environment has a regular rule of dynamics, we take the best action by long-term reward prediction of future outcomes. On the other hand, if the environmental dynamics is highly stochastic, we cannot predict long future outcomes and the best we can do is to act according to immediate outcomes. Thus, in selecting an action to maximize the total outcome, the predictability of the dynamics of the environment is a critical factor.

In this study, to clarify the brain mechanism for reward prediction under different predictability of the environmental dynamics, we designed a novel Markov decision with predictable or unpredictable state transition rules, and measured subjects' brain activities by using functional magnetic resonance imaging (fMRI). By block-design analysis, we found that the orbitofrontal cortex and ventral striatum were more strongly activated in unpredictable dynamics, while the dorsolateral prefrontal cortex and dorsal striatum were more strongly activated under predictable dynamics.

Further analysis by a TD learning model reconfirmed ventro-dorsal map of prediction time scale in the cortico-basal ganglia loops (Tanaka et al., 2004). Moreover, to explore how actual actions were selected based on the predictions by these parallel loops, we estimated the prediction time scale and the corresponding values of subjects' actions using Bayesian estimation from the subjects' action sequence. We found significant correlation between the estimated action value signal and the BOLD signal in the medial prefrontal cortex (mPFC), caudate head, and the globus pallidus (GP). These results suggest that the mPFC plays an important role in integrating the predictions by different cortico-basal ganglia loops specialized in different time scales.

2. Methods

2.1. Behavioral Task

In the Markov decision task (Fig. 1), markers on the corners of a square present four states, and the subject selects one of two actions by pressing a button (a_1 = left button, a_2 = right button) (Fig. 1A).The action determines both the amount of reward and the movement of the marker (Fig. 1B). In the REGULAR condition, the next trial is started

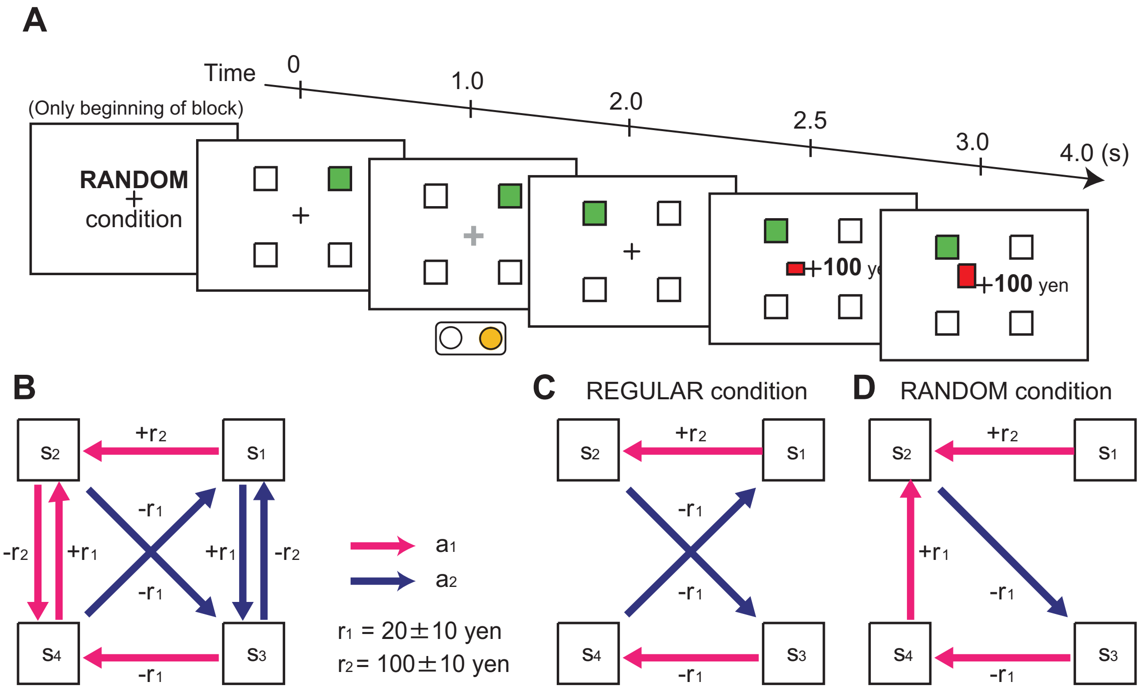

Figure 1: (A) Sequence of stimulus and response events in the Markov decision task. At the beginning of each trial block, the condition is informed by displaying character, such as "RANDOM condition". First, one of four squares representing present state turns green (0s). As the fixation point turns green (1s), the subject presses either the right or left button within 1 second. After 1s delay, the green square changes its position (2s), and then a reward for the current action is presented by a number (2.5s) and a bar graph showing cumulative reward during the block is updated (3.0s). One trial takes four seconds. (B) The rule of the reward and marker movement. (C) In the REGULAR condition, the optimal behavior is to receive small negative rewards $-r_1$ (-10, -20, or -30 yen) at states s_2, s_3, and s_4 to obtain a large positive reward $+r_2$ (90, 100, or 110 yen) at state s_1. (D) In the RANDOM condition, the next trial is started from random state. Thus, the optimal behavior is to select a larger reward at each state.

from the marker position at the end of the previous trial. Therefore, in order to maximize the reward acquired in a long run, the subject has to select an action by taking into account both the immediate reward and the future reward expected from the subsequent state. The optimal behavior is to receive small negative rewards at states s_2, s_3, and s_4 to obtain a large positive reward at state s_1 (Fig. 1C). In the RANDOM condition, next trial is started from a random marker position so that the subject has to consider only immediate reward. Thus, the optimal behavior is to collect a larger reward at each state (Fig. 1D). In the baseline condition (NO condition), the reward is always zero. Subjects performed five trials in the NO condition, 32 trials in the RANDOM condition, five trials in the NO condition, and 32 trials in the REGULAR condition in one trial block in this order. These blocks were repeated four times; thus, the entire experiment consisted of 312 trials, taking about 20 minutes. In order to learn the optimal behaviors, the discount factor γ has to be larger than 0.3425 in REGULAR condition, while it can be arbitrarily small in RANDOM condition.

2.2. An fMRI imaging

Eighteen healthy, right-handed volunteers (13 males and 5 females), gave informed consent to take part in the study, with the approval of the ethics and safety committees of ATR and Hiroshima University. A 1.5-Tesla scanner (Marconi, MAGNEX ECLIPSE, Japan) was used to acquire both structural T1-weighted images (TR = 12 ms, TE = 4.5 ms, flip angle = 20 deg, matrix = 256 × 256, FoV = 256 mm, thickness = 1 mm, slice gap = 0 mm) and T2*-weighted echo planar images (TR = 4 s, TE = 55 msec, flip angle = 90 deg, 38 transverse slices, matrix = 64 × 64, FoV = 192 mm, thickness = 4 mm, slice gap = 0 mm, slice gap = 0 mm) with blood oxygen level-dependent (BOLD) contrast.

2.3. Data analysis

The data were preprocessed and analyzed with SPM99 (Friston et al., 1995; Wellcome Department of Cognitive Neurology, London, UK). The first three volumes of images were discarded to avoid T1 equilibrium effects. The images were realigned to the first image as a reference, spatially normalized with respect to the Montreal Neurological Institute EPI template, and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum). Images of parameter estimates for the contrast of interest were created for each subject. These were then used for a second-level group analysis using a one-sample t-test across the subjects (random effects analysis).

2.4. Regression analysis of $V(t)$ and $\delta(t)$ with fixed γ

To estimate how much forthcoming reward a subject would have expected at each step during the Markov decision task, we estimated reward prediction signal based on the temporal difference (TD) learning model. The theoretical framework of TD learning (Sutton and Barto, 1998) successfully explains reward-predictive activities of the

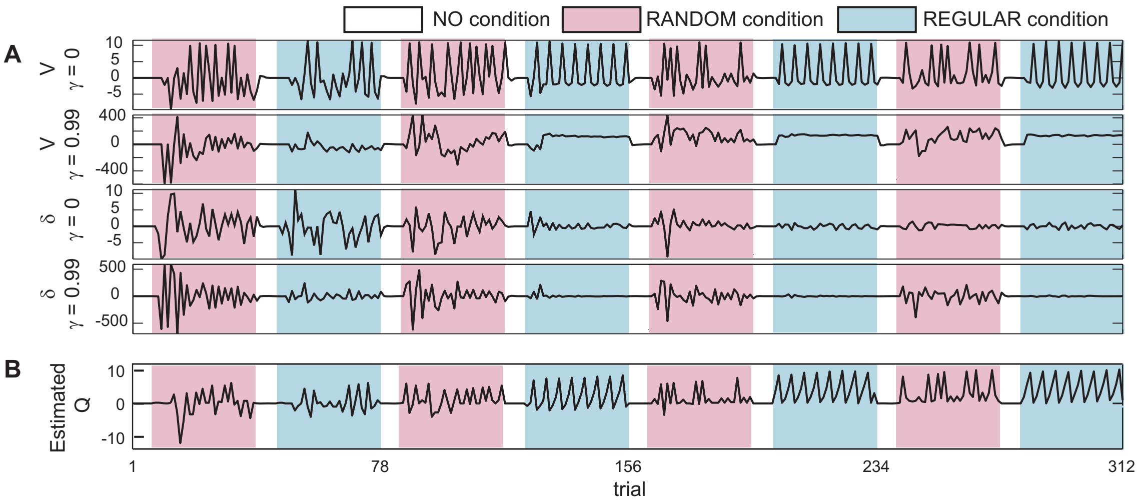

Figure 2: Time series of explanatory variables. Waveforms of the explanatory variables showing reward prediction V(t) (γ = 0 and 0.99), reward prediction error δ(t) (γ = 0 and 0.99), and estimated action value function Q(t) of one subject, convolved with a hemodynamic response function.

midbrain dopaminergic system as well as those of the cortex and the striatum (Berns et al., 2001; Doya, 2000; Houk et al., 1995; O'Doherty et al., 2003; Schultz et al., 1997). In TD learning theory, the predicted amount of future reward starting from a state $s(t)$ is formulated as the "value function,"

$$V(t) = \mathrm{E}[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \ldots], \qquad (1)$$

and learning is based on the TD error,

$$\delta(t) = r(t) + \gamma\, V(t) - V(t\text{-}1). \qquad (2)$$

The “discount factor” γ controls the time scale of prediction; while only the immediate reward $r(t{+}1)$ is considered with $\gamma = 0$, rewards in the longer-term future are taken into account with γ closer to 1. In LEGULAR condition, state transition rule was predictable, thus subjects could predict future reward $r(t{+}2)$, $r(t{+}3)$, …, this means long-term reward prediction. On the other hand, in RANDOM condition, state transition rule was unpredictable, thus subjects could only predict immediate reward $r(t{+}1)$, this means short-term reward prediction.

We calculated $V(t)$ and $\delta(t)$ in the same way as our previous study (Tanaka et al., 2004) Figure 2A shows an example of these time courses. To avoid the effects arising from other functions, we concurrently used possibly relevant explanatory variables. Because the immediate reward prediction $V(t)$ with $\gamma = 0$ and reward outcome $r(t)$ can coincide if learning is perfect, we included the reward outcome $r(t)$ in regression analysis with $V(t)$. Thus, the significant correlation with $V(t)$ should represent the predictive component rather than the reward outcome.

2.5. Baysian estimation of action value *Q*(*t*)

The action value $Q(s,a)$ is also the reward expectation of the subject, but it depended on not only the visited state s but also action a in the Markov Decision Task. The expectation of the reward at the state $s(t)$ and action selection $a(t)$ is

$$Q(s(t), a(t)) = \mathrm{E}[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \ldots | \, s(t), a(t)]. \tag{3}$$

One possible TD learning algorithm for the action value is Sarsa algorithm in which the action value is updated by temporal difference error of action value

$$\delta(t) = r(t) + \gamma \, Q(s(t), a(t)) - Q(s(t-1), a(t-1)), \tag{4}$$

$$Q(s(t), a(t)) = Q(s(t), a(t)) + \alpha \, \delta(\mathrm{t}). \tag{5}$$

We assume that the subjects used updating action value by Sarsa algorithm and that the action values of each condition (REGURAR and RANDOM) are stored independently. Thus, we use the separate estimation of $Q(s,a)$ for RANDOM and REGURAR condition. First, we cut subject's actions and rewards sequence into chunks of each session. Second, the data sequences are connected for each condition. The sequence data for REGULAR and RANDOM condition are used to estimate each action value $Q_{\mathrm{REG}}(s,a)$ for REGULAR condition and $Q_{\mathrm{RAN}}(s,a)$ for RANDOM condition, independenly. The Baysian estimation method for hidden variable of learning agents (Samejima et al 2004) is used to estimate the action values, $Q_{\mathrm{REG}}(s,a)$ and $Q_{\mathrm{RAN}}(s,a)$ for the four states and two actions and also the meta-parameters, such as the discount factor γ, the learning rate α, and the action selection randomness β. Then, the estimated $Q_{\mathrm{REG}}(s,a)$ and $Q_{\mathrm{RAN}}(s,a)$ is cut into each sessions again and reconnect as the order of actual sequence of experience for the subject. Finally, the reconnected estimation sequence of $Q(s,a) = Q(s(t), a(t))$ for the visited state $s(t)$ and the selected action $a(t)$ is used for the regression analysis with the subjective hidden variable.

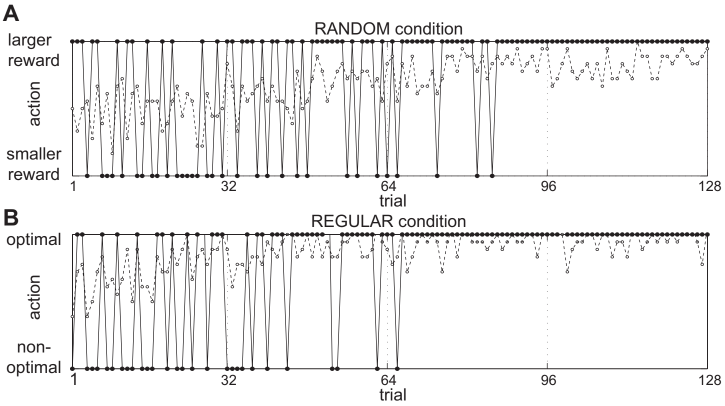

Figure 3: The selected action of a representative single subject (solid line) and the group average ratio of selecting optimal action (dashed line) in (A) RANDOM and (B) REGULAR conditions.

3. Results

3.1. Behavioral results

Figure 3 summarizes the learning performance of a representative single subject (solid line) and group average (dashed line) during fMRI measurement. All subjects successfully learned to take larger immediate rewards in the RANDOM condition (Fig. 3A). In the REGULAR condition, Fourteen subjects successfully learned to take a large reward at s_1 after small punishment at s_2, s_3, and s_4 (Fig. 3B). Other two subjects fell into the 3-states cycle; they took a large reward at $s_{1,}$ a large punishment at s_2, and a small punishment at s_4. Other two subjects could not optimal action sequence. In fMRI data analysis, we included all subjects' data.

3.2. fMRI results: Block-design analysis

To find the brain areas that were involved in reward prediction under unpredictable state transition, we compared the brain activity in RANDOM vs. NO comparison. We observed significant activation in the inferior parietal cortex (IPC), dorsal premotor area (PMd), lateral OFC, and the ventral part of the striatum ($p < 0.001$, uncorrected; Fig. 4A). To find the brain areas that were involved in reward prediction under predictable state transition, we compared the brain activity in REGULAR vs. NO comparison. We observed significant activation in the IPC, PMd, dorsolateral prefrontal cortex (DLPFC), and the dorsal part of the striatum ($p < 0.001$, uncorrected; Fig. 4B). To find the brain areas that were differentially activated under predictable and unpredictable state transition, we compared brain activity of both test conditions directly. In a RANDOM vs. REGULAR comparison, we observed a significant activation in the lateral OFC (p <

0.001, uncorrected; Fig. 4C), while in REGULAR vs. RANDOM comparison, we observed significant activation in the DLPFC (p < 0.001, uncorrected; Fig. 4D). In block-design analysis, we found the brain areas that more strongly activated under predictable and unpredictable state transition rules.

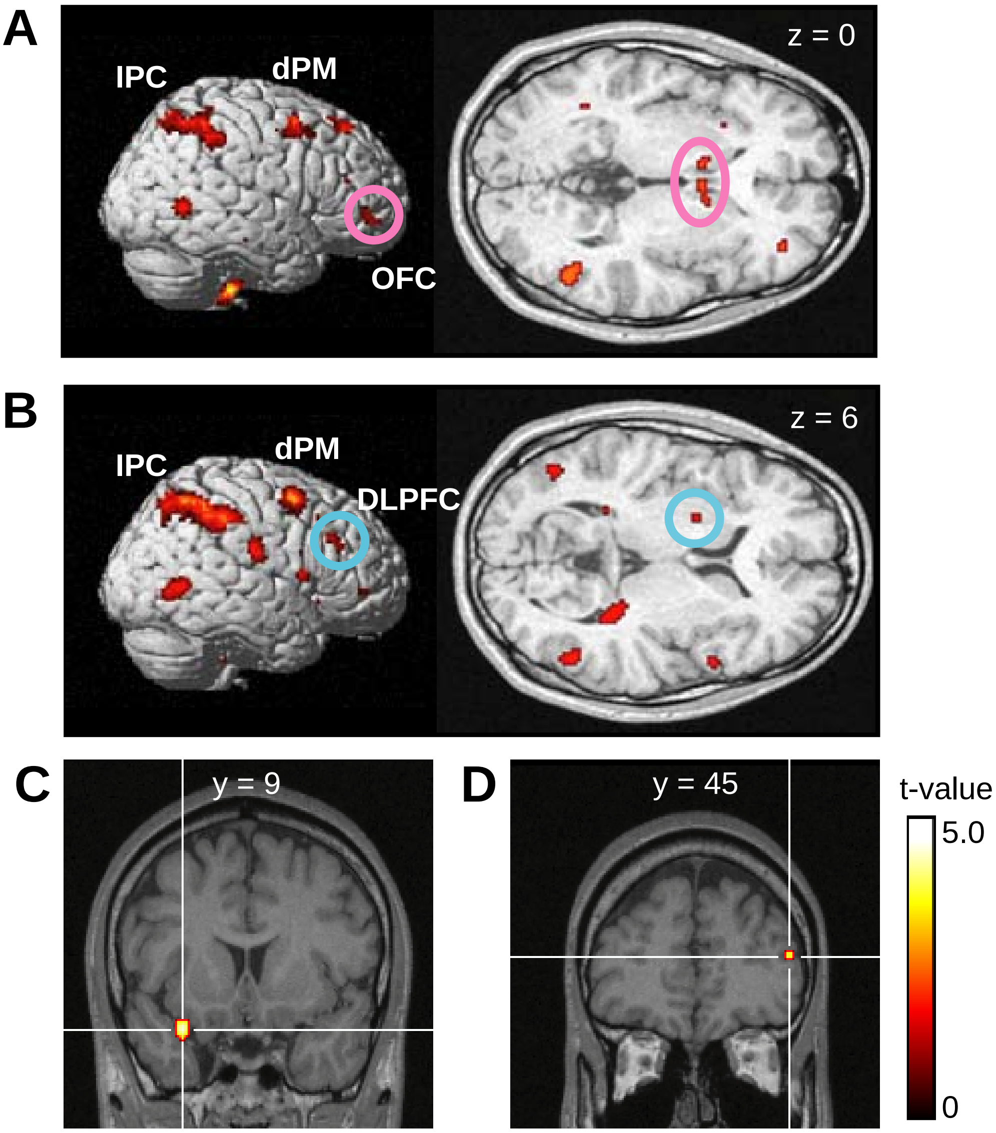

Figure 4: (A) In RANDOM vs. NO comparison, significant activation were observed in the IPC, PMd, OFC, and the ventral part of the striatum (slice at z = 0). (B) In REGULAR vs. NO comparison, significant activation were observed in the IPC, PMd, DLPFC, and the dorsal part of the striatum (slice at z = 6). (C) In RANDOM vs. REGULAR comparison, significant activation was observed in the lateral OFC ((x, y, z) = (-32, 9, -21), peak t-value = 4.90). (D) In REGULAR vs. RANDOM comparison, significant activation was observed in the DLPFC ((x, y, z) = (46, 45, 9), peak t-value = 4.06). All results were applied threshold at p < 0.001, uncorrected for multiple comparison, *n* = 18.

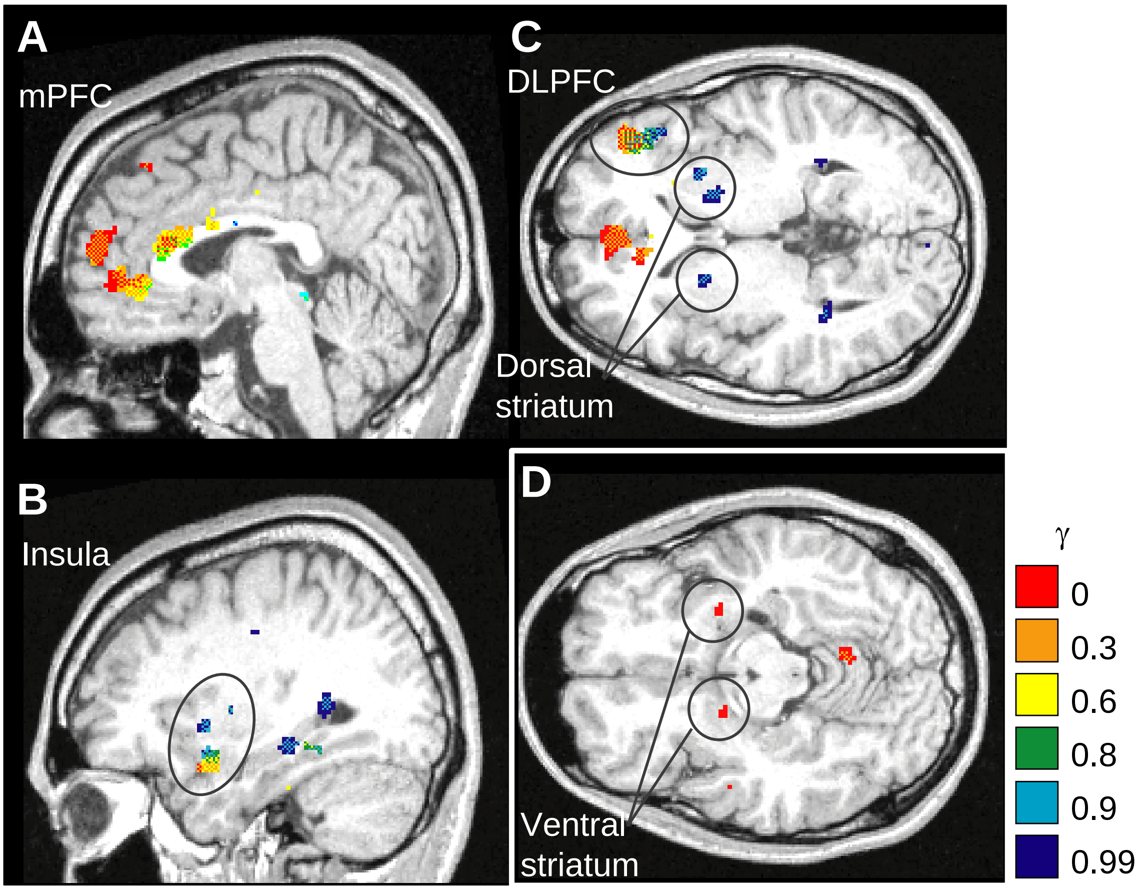

Figure 5: Voxels with a significant correlation ($p < 0.001$, uncorrected) with reward prediction $V(t)$ and prediction error $\delta(t)$ are shown in different colors for different settings of the time scale parameter γ. Voxels correlated with two or more regressors are shown by a mosaic of colors. Significant correlation with reward prediction $V(t)$ was observed in the (A) MFC, (B) insula, (C) DLPFC, and dorsal striatum. (E) Significant correlation with reward prediction error $\delta(t)$ at $\gamma = 0$ was observed in the ventral striatum.

3.3. Regression analysis of V(t) and δ(t) with fixed γ

To clarify the brain areas that were involved in reward prediction at different time scales, we performed regression analysis of BOLD signal with reward prediction signal $V(t)$ and error signal $\delta(t)$ with different settings of time scales γ. We observed significant correlation with reward prediction $V(t)$ in the mPFC ($0 \leq \gamma \leq 0.8$) (Fig. 5A), ventromedial insula (Fig. 5B), DLPFC (all γ), dorsal striatum ($\gamma \leq 0.9$) (Fig 5C) ($p < 0.001$, uncorrected). We also found significant correlation with reward prediction error $\delta(t)$ in the ventral striatum ($\gamma = 0$) (Fig. 5D), IPC, PMd, cerebellum (all γ) ($p < 0.001$, uncorrected).

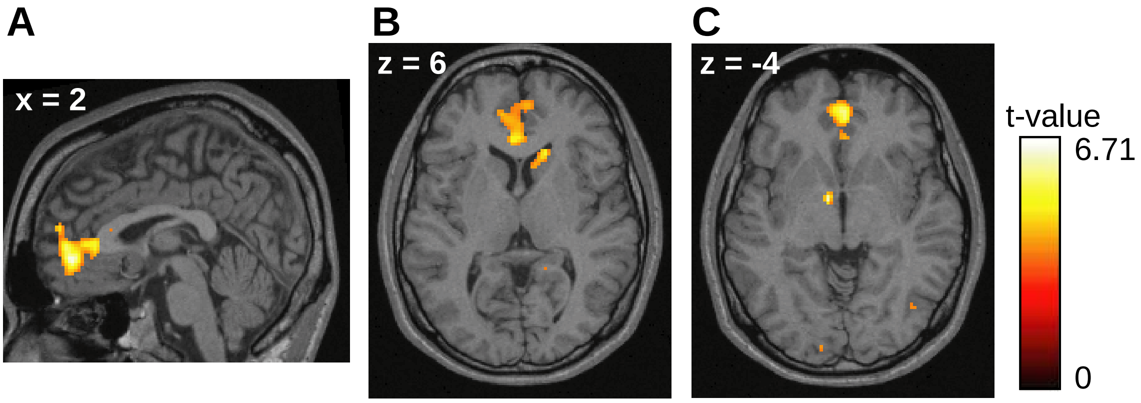

Figure 6: Significant (p < 0.001, uncorrected) correlation with estimated reward prediction V(t) was observed in the (A) mPFC, (B) caudate head, and (C) GP.

3.4. Bayesian estimation of subjective hidden variables

In the above regression analysis, we used several fixed time-scale parameters γ, and found the parallel mechanism for reward prediction at different time scales. What, then, is the relationship between this parallel mechanism and actual action selection? We estimated the subjective reward prediction signal and metaparameters, which are hidden internal variables, using Bayesian estimation from sequence of observable variables state $s(t)$, action $a(t)$, and reward $r(t)$ (Samejima et al., 2004). We used the time course of estimated value function for actually selected action $Q(s(t),a(t))$ (Fig. 2B) as explanatory variables in regression analysis, and we found a significant (p < 0.001, uncorrected) correlation of the estimated reward prediction signal $Q(t)$ in the mPFC (Fig. 6A), the caudate head (Fig. 6B), and the globus pallidus (GP) (Fig. 6C). This result suggests that these areas were more specific to action selection process.

4. Discussion

4.1. Results of block-design analysis

The results of the block-design analysis suggests differential involvement of neural pathways in reward prediction under predictable and unpredictable state transition. Activities of the lOFC, in which show more strongly activation under unpredictable state transition rule than predictable rule (Fig. 4C), were consistent with previous studies that the OFC is involved in reward prediction within a short delay and reward outcome (Breiter et al., 2001; Elliott et al., 2000; Knutson et al., 2001; Koepp et al., 1998; O'Doherty et al., 2002; Pagnoni et al., 2002; Rogers et al., 1999b). Activities of the DLPFC, in which show more strongly activation under predictable state transition rule than unpredictable rule (Fig. 4D), were consistent with previous studies that the DLPFC is involved in reward prediction at a longer time scale (McClure et al., 2004). Our results of block-design analysis were also consistent with model-based regression analysis; the ventral part of the striatum strongly activated in RANDOM condition was also significantly correlated with reward prediction error at smaller time scales (Fig. 5D), and the dorsal part of the striatum strongly activated in LEGULAR condition was also significantly correlated with reward prediction at longer time scales (Fig. 5C). In lateral OFC, PMd, and IPC, we found significant activations in the block-design analyses, but we did not find strong correlation in regression analyses. This may be because these areas perform functions that are helpful for reward prediction, but their activities do not directly represent reward prediction at a specific time scale.

4.2. Topography in the striatum

The ventral part of the striatum was involved in reward prediction error $\delta(t)$ at the shortest time scale ($\gamma = 0$), while the dorsolateral part of the striatum correlated with reward prediction $V(t)$ at longer time scales ($0.9 \leq \gamma \leq 0.99$). This correlation pattern of γ was consistent with our previous studies, which demonstrated that the ventral part of the striatum was involved in short-term reward prediction, and that the dorsal part of the striatum was involved in long-term reward prediction (Tanaka et al., 2004; Tanaka 2005, submitted). Activation of the ventral region of the striatum at the decision of immediate reward was also reported in a recent study (McClure et al., 2004). We found activities in the striatum correlated with reward prediction signals $V(t)$ and prediction error signal $\delta(t)$, where the striatum receives both cortical input, representing sensory cues that allow reward prediction, and dopaminergic input from substantia nigra, representing reward prediction error signal for learning. Thus, in an fMRI experiment, both reward prediction and prediction error signals can be detected as BOLD signals.

4.3. Comparison with our previous study

In our previous experiment with a Markov decision task, we changed the rule of reward between two test conditions, immediate small or delayed larger, while using the same rule of state transition (Tanaka et al., 2004). In the present experiment, to test the brain mechanism of the time scale of reward prediction controlled by state transition rule, we changed the rule of state transition while fixing the rule of reward. In the previous study, we found graded maps of time scale of reward prediction in the striatum and insula. In this study, we found gradient maps of time scale of reward prediction in the striatum, insula, and DLPFC. Thus, the specialization of the parallel cortico-basal ganglia loops for reward prediction at different time scales is a robust observation that does not depend on task context. In this study, to make sure the state transition rule visually, we explicitly displayed state by allocating spatially. The DLPFC, where we did not found significant activation in our previous study but found in present block-design analysis, may be involved in the spatial planning for sequential state transition with different time scales, that needs spatial working memory (Owen et al., 1996).

4.4. Different roles in cortico-basal ganglia loops for reward prediction

In accordance with the regression analysis using a fixed time scale, we demonstrated the parallel loop organization in cortico-striatum loops for reward prediction at different time scales. A regression analysis using estimated subjective internal variables revealed that the mPFC, caudate head, and GP represented values for action that subjects actually selected. In the mPFC, we also found significant correlation of the reward prediction signal at broader time scales ($0 \leq \gamma \leq 0.9$), and we did not find any clear topographic map of γ, same as our previous study (Tanaka et al., 2004). This may be because the mPFC included loops for wide range of time scale, and loops with time scales selected for actual action selection were activated. Previous studies suggest that the mPFC plays an important role in action selection for reward-based goal-directed behavior (Balleine and Dickinson, 1998; Matsumoto et al., 2003).

These results suggest the existence of an action selection mechanism in reward prediction at different time scales within cortico-basal ganglia loops: the striatum calculates the reward prediction signal at different time scales in different sub-regions, and the reward prediction signal at the selected time scale is used for action selection through the basal ganglia loops, via the substantia nigra (SNr), GP, thalamus, and cerebral cortex (Doya, 2000). From previous observation that the mPFC is involved in monitoring action, and anatomical findings that the mPFC receives cortical input from multi-modality sensory areas, the mPFC is a candidate for selecting the optimal time scale from information about actual actions and dynamical changes of environment. The reciprocal connection between the mPFC and the dorsal raphe nucleus containing serotonergic neuron (Martin-Ruiz et al., 2001; Celada et al., 2001) suggests that the mPFC can modulate loop organization through close involvement with the serotonergic system.

4.5. Possible mechanism for reward prediction with different time scales

Based on these results, we propose the following mechanism of reward prediction at different time scales. The parallel cortico-basal ganglia loops are responsible for reward prediction at various time scales. The "limbic loop" via the ventral striatum specializes in immediate reward prediction, whereas the "cognitive and motor loop" via the dorsal striatum specializes in future reward prediction. Each loop learns to predict rewards within its own specific time scale. Therefore, to perform an optimal action within a given time scale, the loop with the most appropriate time scale is more strongly modulated in the striatum. In our recent study using dietary tryptophan, we demonstrated serotonergic modulation of striatum activity: the ventral part of the striatum correlated more strongly with short-term reward prediction at low serotonin levels, while the dorsal part correlated mort strongly with long-term reward prediction at high serotonin levels (Tanaka, 2005).

To control the serotonergic projection to the striatum, the mPFC may regulate the activity of the dorsal raphe by reciprocal connection. In fact, representation of the actual reward prediction signal in the mPFC may be useful for estimating an appropriate time scale for present environment.

Chapter 4

Serotonin differentially regulates reward predictive striatal activities in short and long time scales

The evaluation of both delays and amounts of reward is critical for everyday life. Using dietary control of tryptophan, a precursor of serotonin, and functional brain imaging, we tested the serotonergic effects on brain activity when subjects chose small-immediate or large-delayed liquid rewards. Our results showed that while the activities in the ventral part of the striatum strongly correlated with short-term reward prediction at low serotonin levels, those of the dorsal part strongly correlated with long-term reward prediction at high serotonin levels. Thus, serotonin may control the time scale of reward prediction by differentially regulating the activities within the striatum.

1. Introduction

When you are hungry and looking for a restaurant, do you choose a well-reputed restaurant with many people waiting in line, or a fast-food restaurant where you can have a quick but perhaps less palatable meal? In our everyday life, we constantly make choices between actions leading to rewards of various sizes after different delays. “Delay discounting” is a theoretical concept in which the “value” of a reward R after delay D is given by

$$V = R * G(D),$$

where $G(D)$ is a discounting function that decreases with delay D. A steep rate of discounting results in impulsive choice, defined by an abnormally frequent choice of the more immediate reward (Ainslie 1975; Mazur 1987). Serotonin, one of the major ascending neuromodulators, is thought to be involved in temporal discounting; decreased serotonin levels result in impulsive choice (Wogar, Bradshaw et al. 1993; Bizot, Le Bihan et al. 1999; Mobini, Chiang et al. 2000) and increased serotonin levels decrease impulsive choice (Poulos, Parker et al. 1996; Bizot, Le Bihan et al. 1999). Further, lesions of specific parts of cortico-basal ganglia loop, such as the orbitofrontal cortex and the core of the nucleus accumbens, result in impulsive choice (Cardinal, Pennicott et al. 2001; Mobini, Body et al. 2002). What is the neural network mechanism that links the level of serotonin to future reward evaluation and choice behavior? Our working hypotheses on the serotonergic regulation of delay discounting are as follows. 1) Different sub-regions of the topographically organized cortico-basal ganglia network are specialized for reward prediction at different time scales, and 2) these sub-regions are differentially activated by the ascending serotonergic system (Doya 2002). In our previous brain imaging study (Tanaka, Doya et al. 2004), we demonstrated topographic maps of time scales of reward prediction in the insular cortex and the striatum, in support of the first hypothesis. Here we test the second hypothesis by combining dietary regulation of tryptophan, the precursor of serotonin, and functional magnetic resonance imaging (fMRI) during performance of a choice task with variable delays.

2. Methods

2.1. Subjects

Twelve healthy right-handed males aged 22-25 years gave their informed consent to participate in the experiment, which was conducted with the approval of the Ethics and Safety Committees of Advanced Telecommunication Research Institute International (ATR) and Hiroshima University. On the screening day, a psychiatrist interviewed each volunteer to screen them for psychiatric problems using the Structured Clinical Interview for DSM-IV, and each volunteer had a health examination including blood and urine tests, a chest X-ray, and an electrocardiogram. We excluded participants who had health and/or psychiatric problems, or who disliked the isotonic drink used as the reward in the experiment.

2.2. Experimental procedure

Each subject participated for four days: one day for screening and task practice and three days for fMRI experiments under the three different tryptophan conditions (depletion, trp-; loading, trp+; and control). Three days of experiment were scheduled over a minimal interval of one week to completely remove the effects of tryptophan dietary control induced in the preceding experiment. The experiment was a randomized, double-blind, placebo-controlled, within-subjects design in which a controller who was not an experimenter prepared three types of amino acid mixtures, randomly scheduled for each subject. To maximize the dietary effect, subjects were instructed to consume a low-protein diet that we provided (less than 35 g/day total) for 24 hours before the experiment, and to fast overnight before each experimental day (Delgado, Charney et al. 1990; Bjork, Dougherty et al. 1999; Bjork, Dougherty et al. 2000). To motivate subjects for the liquid reward, their water intake was restricted to 500 ml for 24 hours before each experiment.

On each experimental day, subjects consumed one of three amino acid mixtures (trp-, trp+, and control) and underwent two venipunctures to determine their plasma free tryptophan concentration, which is known to correlate with the CSF serotonin level (Young and Gauthier 1981; Young, Smith et al. 1985; Delgado, Charney et al. 1990; Carpenter, Anderson et al. 1998; Williams, Shoaf et al. 1999; Bjork, Dougherty et al. 2000). The first blood samples were obtained before consumption of the amino acid mixture to confirm the tryptophan baseline, and the second blood samples were taken six hours after consumption to determine the effect of dietary manipulations on the plasma tryptophan level. After the second venipuncture, all subjects entered an fMRI scanner and performed the multi-step delayed reward choice task.

2.3. Amino acid mixtures

We prepared amino acid mixtures consisting of the following quantities of 15 amino acids partially dissolved in 350 ml of water: L-tryptophan: 0 g (trp-), 10.3 g (trp+), 2.3 g (control), 5.5 g L-alanine, 4.9 g L-arginine, 3.2 g glycine, 3.2 g L-histidine, 8.0 g L-isoleucine, 13.5 g L-leucine, 11.0 g L-lysine monohydrochloride, 5.7 g L-phenylalanine, 12.2 g L-proline, 6.9 g L-serine, 6.5 g L-threonine, 6.9 g L-tyrosine, and 8.9 g L-valine. This aqueous suspension was flavored with 10 ml of chocolate syrup. In addition, 2.7 g L-cysteine and 3.0 g L-methionine were administered in a little water with each of the trp- , trp+ and control drinks due to their unpalatability in the beverage.

2.4. Experimental task

Under each tryptophan condition, all subjects performed a multi-step delayed reward choice task in an fMRI scanner (Fig. 1). In this task, subjects chose between a white square indicating a small reward and a yellow square indicating a large reward. At the beginning of each trial, the white and yellow squares, occluded by variable numbers of black patches, were displayed side by side on a screen. After the fixation-cross turned red, the subject selected either the white or yellow square by pressing a button on the corresponding side.

The next set of squares was displayed 2.5 seconds after the previous one with a number of black patches removed from the selected square. The position of the squares (left or right) was changed randomly at each step to retain the subject's attention. When either square was completely exposed, a liquid reward (0.8 ml for the white square, or 3.2 ml for the yellow square) was delivered and the trial was completed. As a liquid reward, an isotonic drink (Pocari Sweat, Otsuka Pharmaceutical Co., Ltd.) was delivered through a plastic tube from a computer-controlled pump (Harvard Apparatus, Inc., PHD 2000 Infusion) outside the MRI room. In the next trial, white and yellow squares were displayed with novel mosaic patterns.

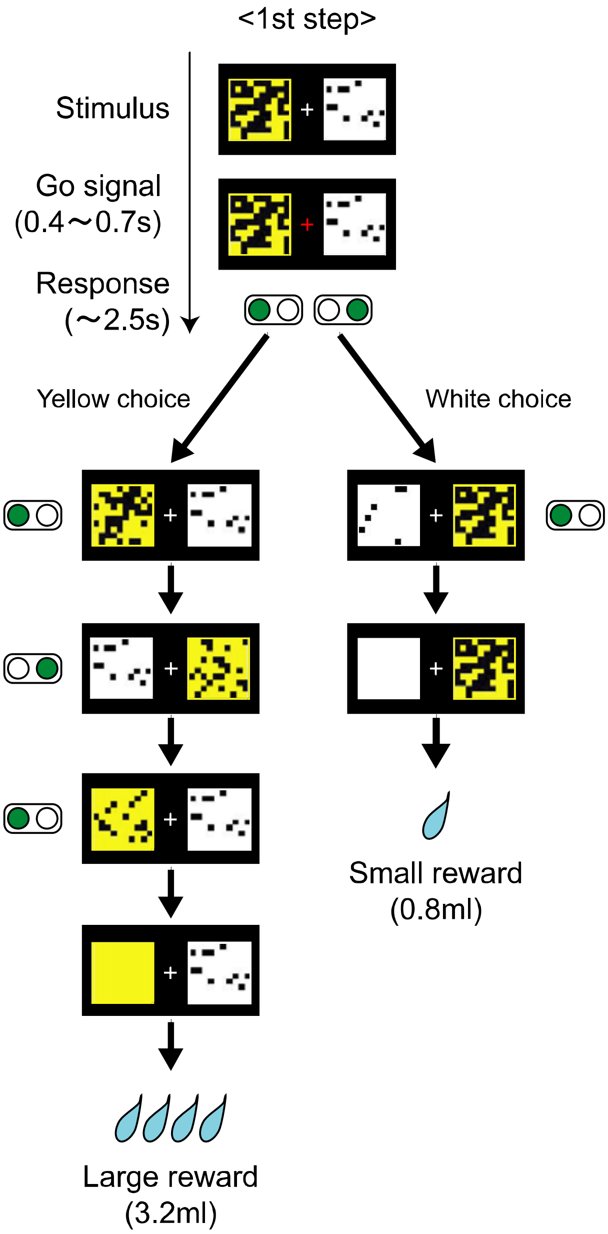

Figure 1: Experimental task. The subject selects either a white or a yellow square. In the example shown here, if the subject chooses a white square at the first step, a small amount of juice (0.8 ml) is delivered in one step. If the subject chooses a yellow square, three yellow choices must be repeated to obtain a larger amount of juice (3.2 ml). The position of the squares (left or right) was changed randomly at each step. For each trial, the initial number of black patches was randomly drawn from a uniform distribution (18±9 patches for white, and 72±24 patches for yellow), and the distribution was fixed throughout the task. The number of patches filled at each step was also drawn randomly from a uniform distribution. Although these distributions were fixed during a three-minute session, they were changed across sessions as follows: (white, yellow) = (6±2, 8±2) in four sessions, (6±2, 16±2) in two sessions and (14±2, 8±2) and (14±2, 16±2) in one session each. The three-minute eight sessions were randomly scheduled.

The initial number of black patches was randomly chosen from uniform distributions (18±9 for white, and 72±24 for yellow). The number of patches filled at each step was also drawn randomly from uniform distributions (*e.g.*, 6±2 for white, and 8±2 for yellow), so that the delay until a small reward (D_s) was usually shorter than the delay until a large reward (D_l). Thus, subjects needed to choose at the beginning of each trial between the more immediate but small reward (white) and the more delayed but large reward (yellow) by comparing the number of black patches on the two squares.

Each experiment consisted of eight sessions lasting three minutes each, with different numbers of patches added to each square, so that the subjects had to remain alert to adapt to the changes in settings. At the beginning of each session, “baseline” blocks (25 sec) showed only the fixation cross; we used these blocks as baselines for brain activity.

All subjects were told that they could obtain a small amount of drink (0.8 ml) when they completely filled the white square by selecting white squares, and they could obtain a larger amount of drink (3.2 ml) when they filled the yellow square by choosing yellow squares. Although subjects could choose either square at any step, they did not usually reverse their choice; reversal accounted for only 8.5% of all choices, and trials with a reversed choice were excluded from analysis.

2.5. Computational model of delay discounting

Exponential discounting of reward *R* at delay *D*

$$V = R\,\gamma^{D}$$

is commonly used in artificial intelligence and economics because it enables an on-line learning algorithm and its optimality under constant rate of reward cancellation (Sutton and Barto 1998). On the other hand, the hyperbolic discounting model,

$$V = R\,/\,(1 + kD)$$

has often been used to explain animal choice behaviors (Ainslie 1975; Mazur 2001). While the exponential model predicts the slope of the indifference line equal to one (from $R_l\,\gamma^{Dl} = R_s\,\gamma^{Ds}$, we have $D_l = D_s + \log\gamma$), the hyperbolic model predict the slope equal to $R_l/R_s = 4$ (from $R_l\,/\,(1 + kD_l) = R_s\,/\,(1 + kD_s)$).Through logistic regression of large and small reward choices, we found that the slopes of the indifference lines of the subjects were intermediate between hyperbolic and exponential model predictions (trp-: 2.815 ± 2.635 (mean ± s.d.), trp+: 2.487 ± 1.913, control: 2.716 ± 2.166). We entered the slopes of the indifference lines of the subjects into a single repeated measures ANOVA with the three tryptophan levels (trp-, trp+, and control). Shifts of the indifference line depending upon the tryptophan condition were observed in a subset of the subjects (Fig. 3b). However, because of large inter-subject variability, we did not find a significant effect of tryptophan level on the slopes or shifts of the indifference lines ($F_{(2,33)} = 0.07$, P = 0.9355, n = 12).

We adopted the exponential model for the analysis of brain activity because 1) there has been no previous report showing hyperbolic growth of neural activity, and 2) hyperbolic discounting can be approximated by a mixture of exponential models (Kurth-Nelson and Redish 2004). Furthermore, we recently showed, in a multi-step delayed reward task similar to that described here, that an exponential model gave a much better fit of human reward choice for monetary reward than a hyperbolic model (N. Schweighofer, et al., submitted (2005)).

We estimated subjects' discount factor γ from the intercept of the indifference line in the D_s-D_l space given by $D_l = D_s + \log(R_s/R_l)/\log\gamma$. Based on the distribution of the estimated discount factor $\gamma = 0.8293 \pm 0.1634$ (mean ± s.d., n = 34, excluded two samples of one subject in two conditions that had negative intercepts), we set γ for the value estimation as 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99.

2.6. Imaging data acquisition and pre-processing

A 1.5 Tesla scanner (Shimadzu-Marconi, MAGNEX ECLIPSE, Japan) was used to acquire both structural T1-weighted images and T2*-weighted echo planar images (TR = 2.5 s, TE = 55 ms, flip angle = 90 deg, 25 transverse slices, matrix = 64 × 64, FoV = 192 mm, thickness = 5 mm, slice gap = 0 mm) with BOLD contrast. We used SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, U.K.) for preprocessing and statistical analyses. The first five volumes of images were discarded to avoid T1 equilibrium effects. The images were realigned to the first image as a reference, spatially normalized with respect to the Montreal Neurological Institute (MNI) EPI template, and spatially smoothed with a Gaussian kernel (8 mm, full width at half maximum).

We checked the brain activities of ventral and dorsal striatum during the "baseline" period, showing only the fixation cross for 25 seconds at the beginning of each session. We performed a single repeated measures analysis of variance (ANOVA) with the three tryptophan levels (trp-, trp+, and control), and did not find significant effects of tryptophan level on the BOLD signal change of ventral and dorsal striatum (ventral: $F_{(2,33)} = 0.69$, $P = 0.509$; dorsal: $F_{(2,33)} = 1.82$, $P = 0.178$). We used ROIs defined at Methods.

2.7. Model-based regression analysis

We estimated the subjective value *V*(*t*) at each step t as follows. We assumed that the subjects knew the mean number of patches filled at each step, s, and its range of variation Δ*s*. The range of possible steps n until the reward from the current number *M*(*t*) of black patches is

$$\frac{M(t)}{s+\Delta s} < n < \frac{M(t)}{s-\Delta s}.$$

We defined the estimated *V*(*t*) as a sum of n-step discounted value $R\gamma^n$ weighed by the probability $P_n(M(t))$ for all possible steps n until the reward,

$$V(t) = \sum_n P_n(M(t)) \cdot R \cdot \gamma^n.$$

The probabilities of reaching the reward in *n* steps is given by

$$P_n(M) = \frac{1}{2\Delta s + 1} \sum_{k=s-\Delta s}^{s+\Delta s} P_{n-1}(M-k).$$

$$P_1(M) = \begin{cases} 1 & \text{if } M \le s-\Delta s \\ \dfrac{M-(s-\Delta s)}{2\Delta s} & \text{if } s-\Delta s < M < s+\Delta s \\ 0 & \text{if } M > s+\Delta s \end{cases}$$

We used the estimated *V*(*t*) as the explanatory variable in a general linear model (GLM) by multiplying the simple event regressor (δ-function) at the timing of stimulus presentation of each step. To remove any effects on factors other than *V*(*t*), we concurrently used other variables in the regression, namely, reward *R* (= 1 for white, 4 for yellow) at reward delivery timing and a box-car function representing the baseline blocks (25 sec) for 8 sessions. All explanatory variables were convolved with a canonical hemodynamic response function (HRF). For each tryptophan condition, images of parameter estimates were created for each subject and entered into a second-level group analysis using a one sample t-test at a threshold of P < 0.001, uncorrected for multiple comparisons (random effect analysis, n = 12). We repeated the same process for each tryptophan condition.

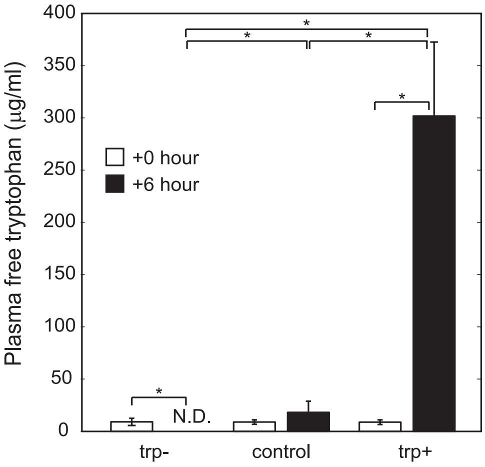

Figure 2: Results of the tryptophan diet. Plasma free tryptophan before consumption (+0 h) and six hours after consumption (+6 h). N.D (not detectable) indicates less than 3.9 g/ml. Error bars indicate standard deviation. *, statistically significant difference (one-tailed two sample t-test, P < 0.0001). In the statistical test for trp-, we used 3.9μg/ml for all subjects because free tryptophan levels were undetectable in all of these subjects.

To compare the results of regression analyses with six different values of γ, we used display software (multi_color: http://www.cns.atr.jp/multi_color/) that can overlay multiple activation maps in different colors on a single brain structure image. When a voxel is significantly activated in multiple values of γ, it is shown by a mosaic of multiple colors, with apparent subdivision of the voxel.

2.8. ROI analysis

We defined ROIs as clustered voxels in which we found significant ($P < 0.001$, uncorrected) correlations with $V(t)$ at $\gamma = 0.6$ in trp- (368 mm^3, union of two clusters peak at (x, y, z) = (26, 0, -4) and (-26, 0, -8); see Supplementary Table), and $\gamma = 0.99$ in trp+ (88 mm^3, union of two clusters peak at (x, y, z) = (24, 2, 22) and (-16, 2, 28)) within an anatomical ROI of the striatum, determined from the normalized T1 image. We again performed GLM within ROIs at the single-subject level, and computed the value of the regression coefficient of $V(t)$ averaged across subjects for each tryptophan condition. We performed a one-tailed two sample t-test between regression coefficients in trp- and trp+ ($\alpha = 0.05$, n = 12 subjects). We used the MarsBaR toolbox (http://marsbar.sourceforge.net/) for ROI analyses.

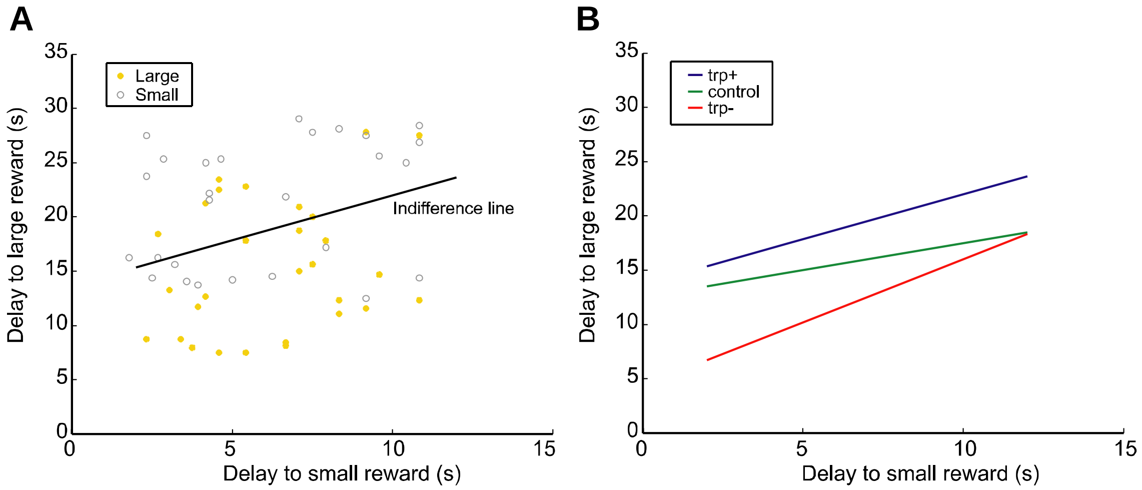

Figure 3: An example of subject's choice. (a) Small and large reward choice on the Ds-Dl space and the indifference line. (b) Shifts of the indifference line depending upon the tryptophan conditions.

3. Results

3.1. Behavioral results

We examined the effects of three different dietary tryptophan levels on the prediction of delayed reward in a double-blind, randomized, within-subject design. Each of the twelve subjects participated in three experimental days, with a minimum interval of one week between days. Each day, a subject consumed one of three amino acid drinks: one containing a standard amount of tryptophan (control; 2.3 g per 100 g amino acid mixture, see Methods for details), one containing excess tryptophan (trp+; 10.3 g), and one without tryptophan (trp-; 0 g). Six hours after consumption, plasma free tryptophan was significantly lower in the trp- subjects (one-tailed, two sample t-test, $P = 6.52\times10^{-5}$ at a significance level < 0.0001), and higher in trp+ subjects ($P = 1.43\times10^{-12}$) compared with control subjects (Fig. 2). Based on previous studies of dietary tryptophan depletion (Young, Smith et al. 1985; Carpenter, Anderson et al. 1998; Williams, Shoaf et al. 1999) and loading (Young and Gauthier 1981; Bjork, Dougherty et al. 2000), we assume significant decreases and increases in central serotonin levels, respectively.

Based on each subject's small and large reward choices at different points on the D_s-D_l space (Fig. 3a), we performed logistic regression analyses of the probability P_l of a large reward (yellow) choice according to the following model,

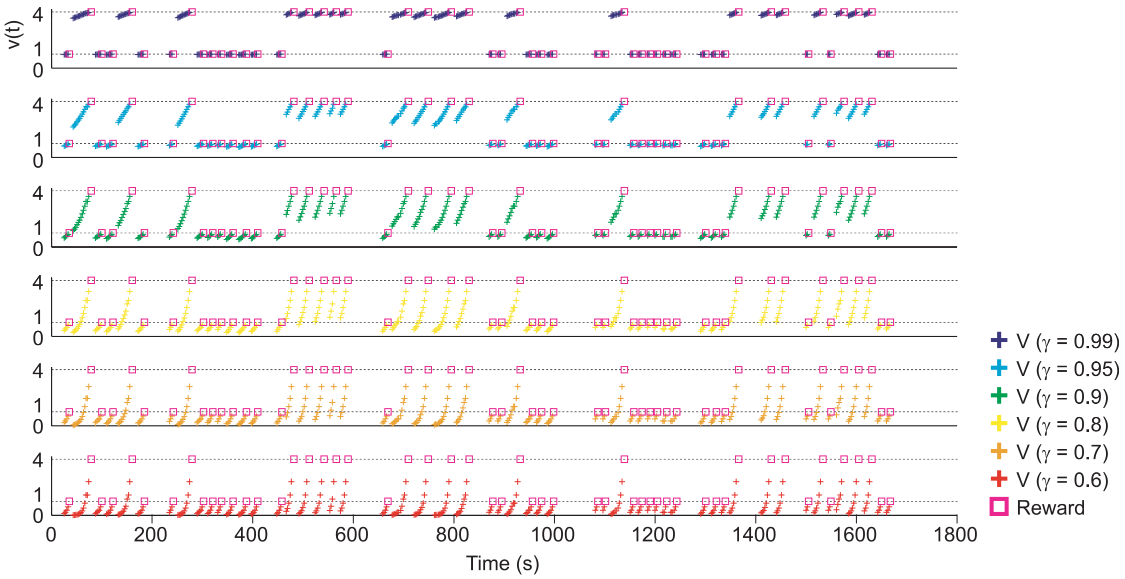

Figure 4: The time course of estimated V(t) (subject S1, control). Each color corresponds to a value of γ used for calculating V(t) (corresponding to the color code used in Figure 2).

$$P_l = 1/(1+\exp[\, -\, (\beta_l\, D_l + \beta_s\, D_s + \beta_0)]).$$

Each subject's choice indifference line was determined by setting $P_l = 0.5$, (*i.e.*, a line given by $D_l = -\ \beta_s/\beta_l\ D_s\ -\ \beta_0/\beta_l$; Fig. S2a). We found no significant differences in the slopes β_s/β_l or shifts β_0/β_l of the indifference lines among the three tryptophan levels for the group of 12 subjects ($F_{(2,33)} = 0.07$, n.s.).

Levels of tryptophan significantly affected subjects' brain activity. We performed model-based fMRI data analyses based on an exponential discounting model (see Methods),

$$V(t) = R\gamma^{\mathrm{T-t}}.$$

Here, the value $V(t)$ represents the discounted future reward R (=1 for white, 4 for yellow) to be acquired at time step T, evaluated at time step t. Because $V(t)$ decreases exponentially with delay $D = T\text{-}t$ it grows exponentially as time t approaches T. The discount factor γ ($0 \leq \gamma < 1$) controls the time scale of reward prediction; smaller γ results in steeper discounting of future rewards, leading to short-term reward prediction. We assumed that subjects estimated the delay $D = T\text{-}t$ until reward delivery from the number of black patches at each step, albeit with some uncertainty (Fig. 4, see Methods).

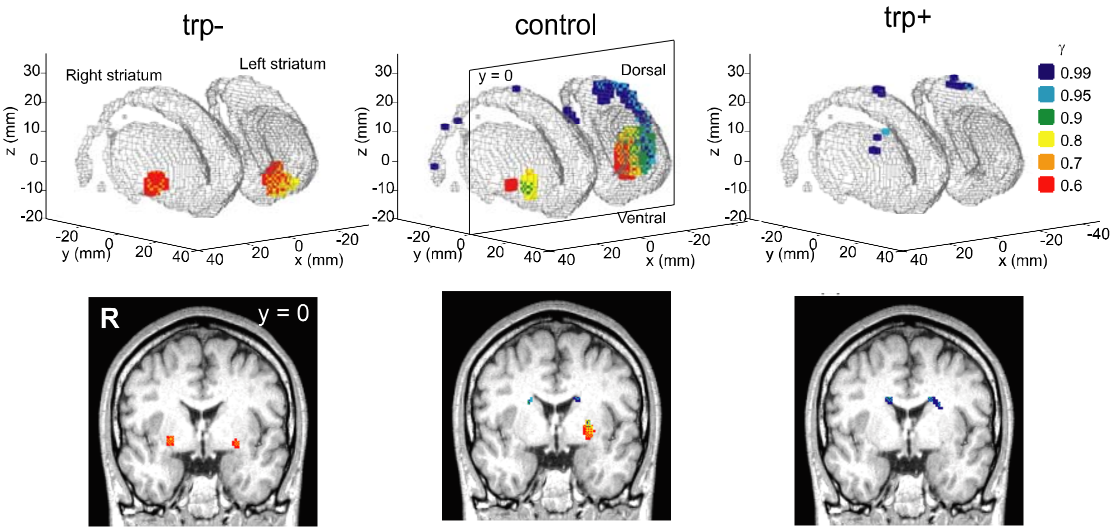

Figure 5: Regression analysis of BOLD signal by expected future reward with different discount rates. Voxels with significant correlation (*P* < 0.001, uncorrected, n = 12 subjects) with *V*(*t*) at different settings of γ with color codes (red: γ = 0.6, orange: 0.7, yellow: 0.8, green: 0.9, cyan: 0.95, blue: 0.99) within the striatum (3D mesh surface). We can see that red to yellow-coded signals are located predominantly in the ventral part of the striatum (ventral putamen and nucleus accumbens), while the green to blue-coded signals are located in the dorsal part of the striatum (dorsal putamen and caudate body).

Based on our hypothesis that different brain areas are involved in reward prediction at different time scales, we estimated $V(t)$ with six different settings of γ (0.6, 0.7, 0.8, 0.9, 0.95, and 0.99), and used each of these as the explanatory variable in a regression analysis (see Methods).

3.2. Brain imaging results: regression analysis of V(t)

We found that blood-oxygen level-dependent (BOLD) signals in the striatum correlated significantly with the estimated $V(t)$ (Fig. 5, Table 1, see Methods). In the control condition (Fig. 5, middle column), we found a significant correlation ($P < 0.001$, uncorrected) with $V(t)$ at all γ values ($0.6 \leq \gamma \leq 0.99$) in the striatum, with a ventral to dorsal gradient ($-4 \leq z \leq 28$) from small to large γ. In the tryptophan depletion condition (trp-; Fig. 5, left column), we found a significant correlation ($P < 0.001$, uncorrected) with $V(t)$ only at smaller γ values (0.6, 0.7, 0.8) in the ventral parts of the striatum ($-12 \leq z \leq -4$). Conversely, in the tryptophan loading condition (trp+; Fig. 5, right column), we found a significant correlation ($P < 0.001$, uncorrected) with $V(t)$ only at larger γ values (0.9, 0.95, 0.99) in the dorsal parts of the striatum ($16 \leq z \leq 28$).

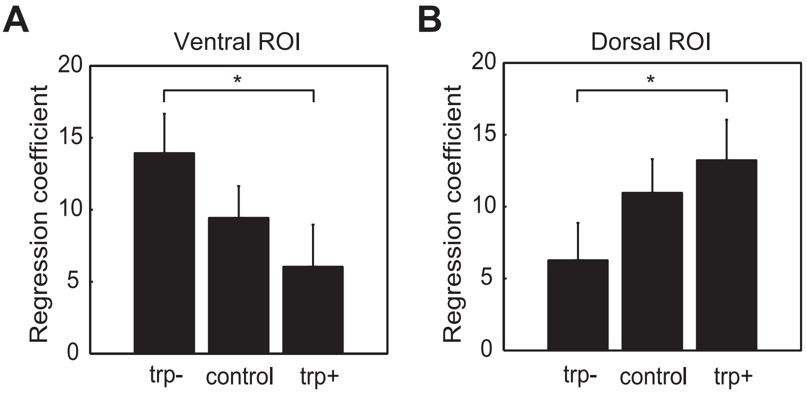

Figure 6: Effects of tryptophan conditions on short- and long-term reward prediction in the ventral and dorsal parts of the striatum. (a) Regression coefficient of the activity in the ventral part of the-striatum with $V(t)$ at $\gamma = 0.6$ was significantly larger in the tryptophan depletion (trp-) than in the loading (trp+) condition. (b) Regression coefficient of the activity of the dorsal part of the striatum with $V(t)$ at $\gamma = 0.99$ was significantly larger in the tryptophan loading (trp+) than in the depletion (trp-) condition. Data shown are the group averages (n = 12 subjects) and error bars represent standard errors. *, statistically significant difference (one-tailed two sample t-test, $P < 0.05$).

3.3. ROI analysis of ventral and dorsal part of the striatum

To quantify the modulation of striatal activity by different tryptophan levels, we compared the regression coefficients of the estimated $V(t)$ and the BOLD signals in the regions of interest (ROI) in the ventral and dorsal parts of the striatum under three tryptophan levels (see Methods). In the ventral part of the striatum, the correlation between $V(t)$ at small γ ($\gamma = 0.6$) and the BOLD signal was significantly stronger in the trp- condition than in the trp+ condition (Fig. 6a; one-tailed two sample t-test, $P =$ 0.0306 at the significance level 0.05), while in the dorsal part of the striatum, the correlation between $V(t)$ at large γ ($\gamma = 0.99$) and the BOLD signal was stronger in the trp+ condition than in the trp- condition (Fig. 6b; $P = 0.0415$).

4. Discussion

4.1. Behavioral and brain imaging results

Our finding presents the first evidence of the relationship between the serotonergic system and the specific localization of brain activity related to reward prediction. Although we did not find significant differences in choice behaviors at different tryptophan levels, as in previous human studies using dietary tryptophan depletion in healthy volunteers (Park, Coull et al. 1994; Salomon, Miller et al. 1997; Crean, Richards et al. 2002), we did observe significant differences in brain activities for reward prediction under the different tryptophan levels. Just as recent studies have revealed differential genotypic effects by brain imaging (Hariri, Mattay et al. 2002; Goldberg and Weinberger 2004), the effects of neuropharmacological regulation may be more sensitively measured by local BOLD signal changes detected by fMRI than by behavioral output, which may be influenced by the entire brain. Thus, our present methods combining dietary tryptophan control and model-based analyses of brain imaging data are effective to evaluate serotonergic effects that may be difficult to detect by behavioral output alone.

4.2. Graded map in the striatum for time scale of reward prediction

In the control tryptophan condition, we found a graded map of delay discount rate in the striatum; activities in the ventral portion correlated with the expected future reward with more rapid discounting, while those in the dorsal part correlated with the expected future reward with slower discounting. This graded map of correlation was consistent with our previous study, which demonstrated that ventral cortico-basal ganglia loops were involved in short-term reward prediction, and that dorsal loops were involved in long-term reward prediction (Tanaka, Doya et al. 2004). Activation of the ventral part of the striatum by an immediate reward choice was also reported in a recent study (McClure, Laibson et al. 2004).

4.3. Discussion of striatal activity

In this study, we found activities in the striatum correlated with reward prediction signals estimated by a computational model. This result is consistent with previous neural recording studies reporting reward expectation-related activities in the striatum

(Kawagoe, Takikawa et al. 1998; Shidara, Aigner et al. 1998; Tremblay, Hollerman et al. 1998; Samejima, Ueda et al. 2005). However, a number of previous functional brain imaging studies have shown striatal activities correlated with reward prediction error (McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003; Seymour, O'Doherty et al. 2004; Tanaka, Doya et al. 2004). The striatum receives both cortical input, representing sensory cues that allow reward prediction, and dopaminergic input from the substantia nigra, representing reward prediction error signal for learning (Schultz, Dayan et al. 1997). Thus, in an fMRI experiment, both reward prediction and prediction error signals can be detected as a BOLD signal. The reason we found correlation with reward prediction in this study could be because the reward prediction error due to the uncertainty of the number of steps until the reward was relatively small compared with the steady build-up of reward expectation.

4.4. Serotonergic modulation of striatal activity for reward prediction

The novel finding of our present study is that the parallel organization for reward prediction at different time scales in the striatum is under differential modulation by the central serotonergic system. We found a graded correlation from the ventral to the dorsal parts of the striatum under control tryptophan conditions, but the correlation in the ventral part was only significant in the depletion condition, and that in the dorsal part only in the loading condition. An ROI analysis confirmed that activity in the ventral striatum, correlated with the expected future reward with more rapid discounting, was enhanced in the tryptophan depletion condition, while activity of the dorsal part, correlated with the expected future reward with slower discounting, was enhanced in the loading condition. These results support our hypothesis that different parts of the striatum are dedicated to reward prediction at different time scales, and that they are differentially enhanced or suppressed by serotonergic modulation. Various subtypes of the serotonin receptor, with different affinities and cellular effects, are differentially distributed in the striatum (Compan, Segu et al. 1998). Such differential distributions of receptor subtypes could allow differential activation in the striatum under different serotonin levels. These mechanisms might be revealed by using particular receptor ligands in a positron emission tomography (PET) experiment (Halldin, Gulyas et al. 2001).

Table 1: Voxels significantly correlated with estimated $V(t)$ ($P < 0.001$, uncorrected for multiple comparisons, $n = 12$).

	trp-			control			trp+		
	Area	T-value	MNI Coordinates (x,y,z)	Area	T-value	MNI Coordinates (x,y,z)	Area	T-value	MNI Coordinates (x,y,z)
$\gamma = 0.6$	Putamen	5.42	(26,0,-4)	Putamen	5.72	(-26,0,-2)			
	Putamen	4.92	(-26,0,-8)	Putamen	4.17	(28,2,-4)			
	Nacc	4.99	(-20,10,-10)	Parietal cortex	5.86	(32,-58,64)			
$\gamma = 0.7$	Putamen	4.94	(26,0,-4)	Putamen	5.84	(-24,2,2)			
	Putamen	4.35	(-26,0,-6)	Putamen	4.67	(24,8,-2)			
	Nacc	5.46	(-20,10,-10)	Parietal cortex	5.72	(32,-58,64)			
$\gamma = 0.8$	Nacc	6.16	(-20,10,-12)	Putamen	5.71	(-24,2,4)			
				Putamen	5.28	(24,8,-4)			
				Parietal cortex	6.01	(24,-72,58)			
				Occipital cortex	8.59	(22,-98,4)			
$\gamma = 0.9$				Putamen	5.02	(-24,2,4)	Putamen	5.38	(24,10,16)
				Putamen	4.41	(26,8,-4)			
				Parietal cortex	6.81	(22,-72,58)			
				Occipital cortex	8.65	(-20,-98,10)			
				Occipital cortex	10.04	(24,-98,4)			
$\gamma = 0.95$				Putamen	5.41	(-30,6,12)	Putamen	4.59	(26,10,16)
				Putamen	4.05	(30,0,10)			
				Caudate	6.22	(24,2,24)			
				Parietal cortex	6.7	(22,-52,44)			
				Occipital cortex	9.99	(-20,-98,10)			
				Occipital cortex	8.33	(24,-96,4)			
$\gamma = 0.99$				Putamen	6.93	(-22,-8,12)	Caudate	6.45	(-16,2,28)
				Caudate	4.96	(-16,-2,24)	Caudate	5.01	(24,2,22)
				Caudate	5.57	(26,2,28)			
				Parietal cortex	6.54	(20,-52,54)			
				Occipital cortex	8.7	(-20,96,12)			
				Occipital cortex	7.36	(24,-92,4)			
				Cerebellum	7.27	(38,-66,-40)			

Chapter 5

Serotonin affects temporal credit assignment in delayed stimulus-outcome association learning

To test the hypothesis that serotonin affects the learning of stimulus-outcome associations based on delayed rewards and punishments, we developed a novel "temporal credit assignment task," and analyzed the resulting behavior under central serotonin manipulation by dietary tryptophan depletion, loading, and control. In this task, to maximize total outcome, subjects needed to learn the possibly delayed association of stimulus-outcome by correctly assigning credit of the present outcome to previously selected stimuli. We found significant differences in the choice rate of delayed small punishment against delayed large punishment under different serotonin conditions. At low serotonin levels, the choice rate in a latter block was lower than under the control condition. In contrast, under high serotonin levels, the choice rate in an early block was significantly higher than under other conditions. Based on reinforcement learning model, we estimated which subjects' learning parameters maximize the likelihood of their actions, and revealed that the estimated trace decay parameter at low serotonin levels was smaller than at high levels. Our findings suggest that serotonin regulates the temporal credit assignment of the present outcome to past stimuli: lower serotonin levels result in impairment of assigning credit to distant stimuli, while higher serotonin levels improve it.

1. Introduction

1.1. Serotonin and impulsivity

In everyday life, we learn and choose a "better" action based on certain policies. Usually, in the real world, we can obtain a reward after variable delays; thus, delay is one of the critical criteria for selecting the optimal action.

Considering action learning based on delayed reward, we take an action by predicting a future reward arising from it (Fig. 1a), and learn an action by associating a reward with a past action (Fig. 1b). In animal experiments, it has been demonstrated that the longer the delay in receiving the reward, the more likely it is to choose an immediate but smaller reward (Ainslie 1975; Cardinal, Winstanley et al. 2004). Dickinson reported that longer delay hindered response-outcome association learning in rats (Dickinson, Watt et al. 1992). Based on these previous studies, "impulsive choice," defined as abnormally frequent choices of immediate-small rewards rather than delayed-large rewards, may occur due to shortsightedness in reward prediction and impairment in associating between reward and distant past action. Thus, excessively short settings of both the time scale of forward view for future reward and backward view for past action may cause impulsive choices.

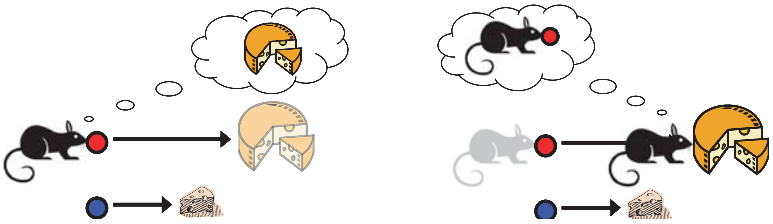

Figure 1: Time scale of forward view and backward view.

1.2. Impulsivity in the reinforcement learning model

These time scales are described in the reinforcement learning model (Sutton and Barto 1998), where the "value" of reward is given by the sum of future outcomes with temporal discounting.

$$V(t) = \mathrm{E}[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \ldots] \qquad (1)$$

The discount factor γ ($0 \leq \gamma \leq 1$) controls the time scale of reward prediction: a very small γ means steeper discounting of the weight of the delayed reward, resulting in immediate choice.

The time scale of the backward view can be explained by the “eligibility trace model.” An eligibility trace is a memory of a past event that is eligible for use in learning. At each step the eligibility traces for all states decay by λ, and the eligibility trace for the one state visited at each step is incremented by 1:

$$\begin{aligned} e_i(t) = {} & \lambda e_i(t-1) + 1 \quad (s(t) = s_i) \\ & \lambda e_i(t-1) \qquad\quad (s(t) \neq s_i) \end{aligned} \tag{2}$$

The traces are said to indicate the degree to which each state is eligible for undergoing learning changes should a reinforcing event occur. The TD error for state-value prediction is

$$V(t+1) = V(t) + \alpha\delta(t)e(t). \tag{3}$$

The TD error signal triggers proportional updates to all recently visited states. The traces record which states have recently been visited, where "recently" is defined in terms of the trace-decay parameter λ ($0 \leq \lambda \leq 1$); an excessively small λ, which means action-outcome association only in the recent past, results in immediate choice.

1.3. Serotonin hypothesis and the present aim

Impulsivity is one of the symptoms of depression, ADHD (attention-deficit hyperactive disorder), and drug abuse (Ainslie 1975; Evenden 1999). Clinical reports and animal experiments suggest that monoamines, such as dopamine and serotonin, are leading causes of impulsivity. Especially, many results from experiments on rats indicated that serotonin was involved in impulsive choice (Wogar, Bradshaw et al. 1993; Evenden and Ryan 1999; Mobini, Chiang et al. 2000). Assuming that impulsive choice can result from excessively too small settings of the time scales for reward prediction and temporal credit assignment, how is serotonin involved in these time scales?

In our previous study, we demonstrated that serotonin can control the time scale of reward prediction by regulating the reward predictive activity in the striatum, in which there are parallel loops involved in reward prediction at different time scales (Chapter 4). In this study, to explore the effect of serotonin on the time scale of temporal credit assignment, we develop a novel task that requires subjects to learn optimal action by associating outcomes with past actions, and examine subjects' behavior under different serotonin levels.

2. Methods

2.1. Subjects and serotonin manipulation

Twenty-two healthy, right-handed males gave their informed consent to participate in the experiment, which was conducted with the approval of the ethics and safety committees of Advanced Telecommunication Research Institute International (ATR) and Hiroshima University. On the screening day, a psychiatrist interviewed each volunteer to assess them for psychiatric problems using the Structured Clinical Interview for DSM-IV, and each volunteer had a health examination including a blood test, urine test, chest X-ray, and an electrocardiogram to screen for health problems. We precluded participants who had health and/or psychiatric problems, and disliked the juice used as the reward in the experiment.

Each subject participated for four days: one day for screening and task training and three days for experiments under the three differential tryptophan conditions (Trp-, Trp+, and Control). These experiments took place over an interval of more than one week to completely remove the effects of tryptophan dietary control on the last experiment day. The experiment was a double-blind, within-subjects design in which the controller prepared a randomized schedule of three tryptophan conditions for each subject. To maximize the pharmacological impact, subjects were instructed to consume only the low-protein diet we provided (less than 35 g/day total) beginning from 24 hours before the experiment and were instructed to fast overnight before each experiment day. Dietary tryptophan depletion is known to reduce the levels of central serotonin metabolite in cerebrospinal fluid (CFS) (Young, Smith et al. 1985; Carpenter, Anderson et al. 1998; Williams, Shoaf et al. 1999), and dietary tryptophan loading increases level of CFS serotonin metabolite (Young and Gauthier 1981; Bjork, Dougherty et al. 2000).

We prepared the amino acid mixtures, comprising the following quantities of 15 amino acids partially dissolved in 350 ml of water: L-tryptophan: 10.3 g (loading), 2.3 g (control), 0 g (depletion), 5.5 g L-alanine, 4.9 g L-arginine, 3.2 g glycine, 3.2 g L-histidine, 8.0 g L-isoleucine, 13.5 g L-leucine, 11.0 g L-lysine monohydrochloride, 5.7 g L-phenylalanine, 12.2 g L-proline, 6.9 g L-serine, 6.5 g L-threonine, 6.9 g L-tyrosine, and 8.9 g L-valine. This aqueous suspension was flavored with 10 ml chocolate syrup. In addition, 2.7 g L-cysteine and 3.0 g L-methionine were administered in a little 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.

On each experiment day, subjects underwent two venipunctures to determine the plasma free tryptophan concentration, which is proved to correlate with the CSF serotonin level. The first blood samples were obtained to confirm the tryptophan baseline, and the second ones were taken six hours after consumption of the amino acid drink to determine the effect of the tryptophan dietary manipulations of the plasma tryptophan level. After the second venipuncture, all subjects performed the temporal credit assignment task.

2.2. Experimental task

In this task, subjects chose one of two fractal figures by pressing a corresponding button. Depending on the selected figure, an outcome was displayed at the screen. If subjects selected the 0 delay figure, the outcome was displayed at the present trial (Fig. 2, trials 2 and 3). On the other hand, if subjects chose the delay 3 figure, the outcome was displayed three trials later (Fig. 2, trial 1). If outcomes from delay 0 and delay 3 appeared in the same trial, the sum of the immediate and delayed outcomes was fed back (Fig. 2, trial 4). Thus, to maximize total the outcome, subjects needed to learn the possibly delayed stimulus-outcome association by correctly assigning credit of the present outcome to previously selected stimuli.

We used eight fractal figures (Fig. 3), each of which had a different setting of outcome (40, 10, -10, -40 yen) and delay (0 trial, 3 trials). At each trial, two fractal figures were chosen from these eight figures in pseudo random order. We prepared sixteen pairs, with consideration given to the number of appearance of figures.

All subjects performed 110 trials in a single session, and performed 6 sessions on each experiment day (about 28 minutes). At the beginning of each session, the session

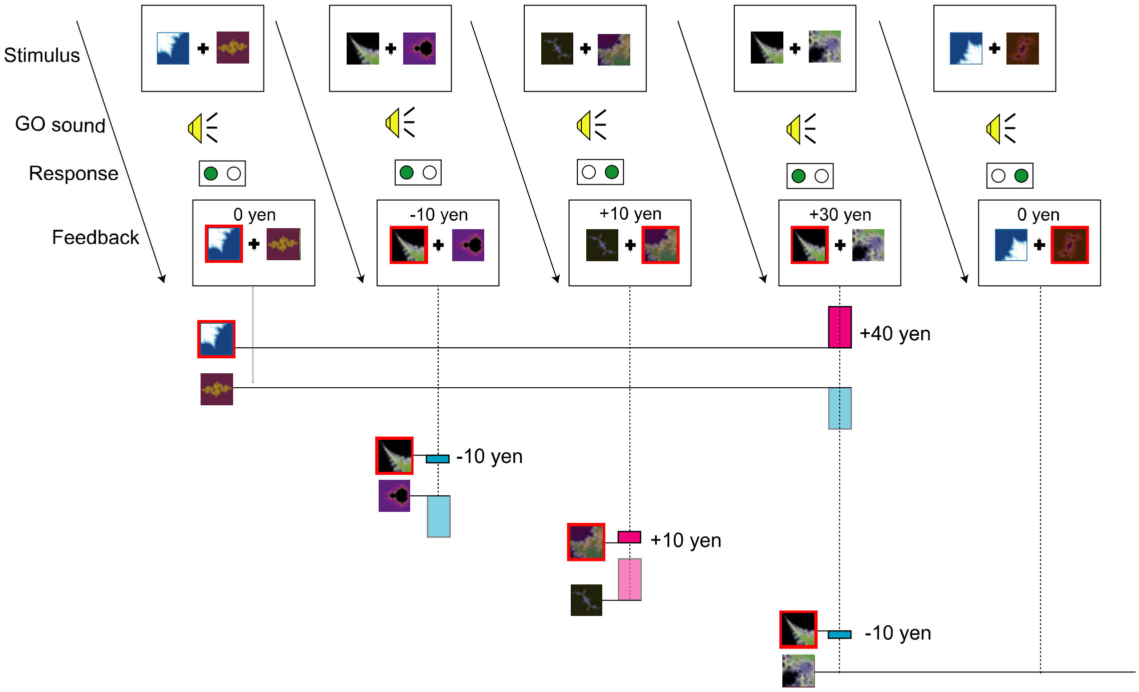

Figure 2: Experimental task. In the "temporal credit assignment task" (Fig. 2), two fractal figures were displayed on the screen at each trial. As subjects heard the beep sound (250 msec), they chose one of two fractal figures by pressing a corresponding button (250 ~ 1,250 msec). Depending on the selected figure, an outcome was displayed either immediately or three trials later. The sum of the immediate and delayed outcomes was fed back (1250 ~ 2,500 msec). A single trial lasted 2.5 seconds. At the next trial, a new pair of fractal figures was displayed.

number was displayed on the screen for 2.5 seconds. None of the subjects were informed about their outcome-delay mapping, although they were instructed that there were different settings: immediate/delayed, small/large, and reward/punishment. We used completely different figures during screening and on experiment days.

3. Results

3.1. Serotonin manipulation

Except for one subject (subject 11), six hours after consumption the plasma free tryptophan had significantly decreased in the trp- condition and increased in the trp+ condition. Therefore, we omitted subject 11 from the following analyses.

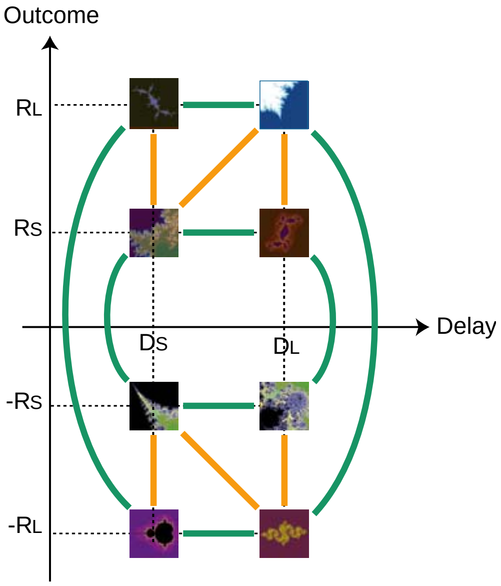

Figure 3: Sixteen pairs of figures were presented pseudo randomly. Each pair was presented as a scheduled trial number; six pairs connected by an orange line were presented on every tenth trial in a single session, and other pairs connected by a green line were presented on every fifth trial during single session.

3.2. Choice rate analysis (three-way ANOVA)

To explore the serotonergic effects on the time scale of temporal credit assignment, we compared the learning performance of immediate pairs (a: immediate small reward vs. large reward, c: immediate small punishment vs. large punishment) and delayed pairs (b: delayed small reward vs. large reward, d: delayed small punishment vs. large punishment) at different tryptophan levels. In these pairs, we could ignore the serotonergic effect on the time scale of reward prediction because of the same delay; even in delayed pairs, their value in the timing of reward prediction was temporally discounted at the same rate.

In shorter temporal credit assignments, subjects can learn the correct choice (larger reward or smaller punishment) in an immediate pair, although they are impaired in learning the correct choice in delayed pairs because they cannot correctly assign the outcome to the action chosen before the three trials. In contrast, in longer temporal credit assignments, subjects can learn correct choices both in immediate and delayed pairs. Our hypothesis is that serotonin controls the time scale of temporal credit assignment. If a low serotonin level might result in short temporal credit assignments, we would expect to observe slow learning performance in delayed pairs (b, d), not in immediate pairs (a, c) in trp-. In contrast, in trp+, we expected to observe good performance even in delayed pairs (b, d).

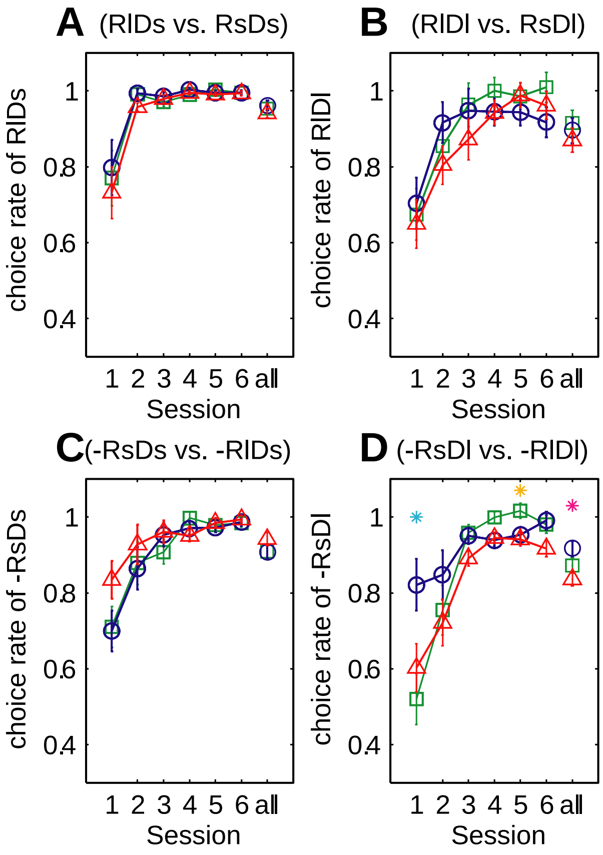

Figure 4: Results of the optimal choice rate in (A) RsDs vs. RlDs, (B) RsDl vs. RlDl, (C) -RsDs vs. -RlDs, and (D) -RsDl vs. –RlDl (trp-: red triangle, control: green square, trp+: blue circle). Optimal choice was determined by a larger reward or smaller punishment for each pair. *: statistically significant difference, p < 0.05 multiple comparison test after three-way ANOVA with tryptophan levels (n = 3), experiment days (n = 3), and subjects (n = 12).

We plotted the choice rate of correct figures for each pair during each block (Fig. 4). In addition, we preformed a three-way analysis of variance (ANOVA) with factors of tryptophan levels, experiment days, and subjects, and a performed multiple comparison test with effects outside our interest removed (experiment days and subjects). We observed a significant difference in the choice rate of delayed small punishment against delayed large punishment in regard to different tryptophan levels (Fig. 4d). In block 1, the choice rate of -DlRs was significantly larger under trp+ than under other conditions, while in block 5, the choice rate of -DlRs was significantly smaller under trp- than under the control condition. For the entire session, the choice rate of -DlRs was significantly smaller in trp- than in trp+ ($p < 0.05$ for the multiple comparison test). This result shows that subjects were slow to learn the association between the selected figure and delayed punishment at low serotonin levels, although learning improved at high serotonin levels. This suggests that serotonin is involved in regulating the time scale of temporal credit assignment.

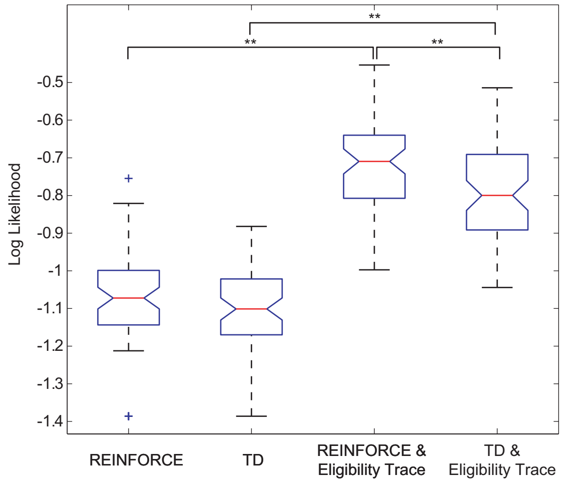

Figure 5: Model comparison. Log likelihood of subjects' action in present model (REINFORCE & Eligibility trace: γ = 0 with eligibility trace), REINFORCE (γ = 0 without eligibility trace model), TD & Eligibility trace (γ = 1 with eligibility trace model), and TD model (γ = 1 without eligibility trace model).

3.3. Estimated parameter from subjects' data using reinforcement learning agents

To clarify the serotonergic effect on association learning, we estimated the subjects' parameter of temporal credit assignment from their action sequence, based on the reinforcement learning model. We computed an eligibility trace (Eq. 2) for all figures, and TD error by

$$\delta(t) = r - V(t). \tag{4}$$

In this task, subjects did not need to predict future reward because there was no state transition rule; thus we used γ = 0. The value function was updated by Eq. 3. In this model, subjects can learn the value function of each figure indirectly by applying the eligibility trace.

3.3.1. Model comparison

To evaluate whether this model can explain the subjects' behavior, we compared log likelihood of subjects' action between other models (Fig. 5). We computed log likelihood by using four possible models: (a) present model (γ = 0 with an eligibility trace), (b) γ = 0 without an eligibility trace, (c) γ = 1 with an eligibility trace, and (d) γ = 1 without an eligibility trace. We found that log likelihood in our present model was larger than in other models. Thus, we used this one for the subjects' learning model in this task.

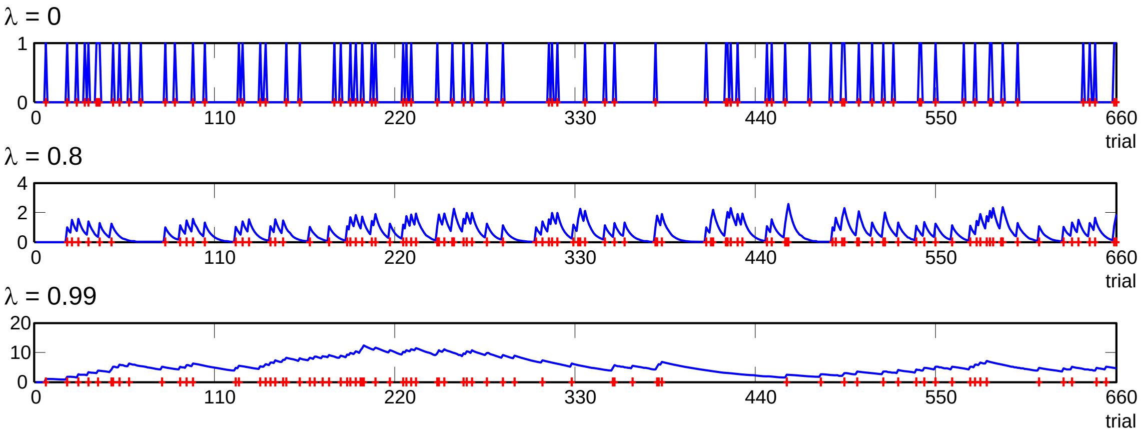

Figure 6: Time course of eligibility trace. This shows an example of the time course of an eligibility trace for one stimulus, and these red lines show visits to this stimulus.

3.3.2. Simulation

Figure 6 shows an example of time course of the eligibility trace of figure DsRs. For λ = 0, the eligibility trace was a spike-like pattern (Fig. 6a), thus TD was used to update V for only the present selected figure. We expected good performance in the immediate pairs but poor learning in delayed pairs for a small λ. For a larger λ, the eligibility trace was sustained over several trials with temporal decay (Fig. 6b), thus TD was used to update V of not only the present figure but also past figures. We expected good performance in both immediate and delayed pairs. For an excessively large λ, the eligibility trace did not discount for a long duration (Fig. 6c), thus TD was used to update the V even of figures that were visited in the long distant past. Since this

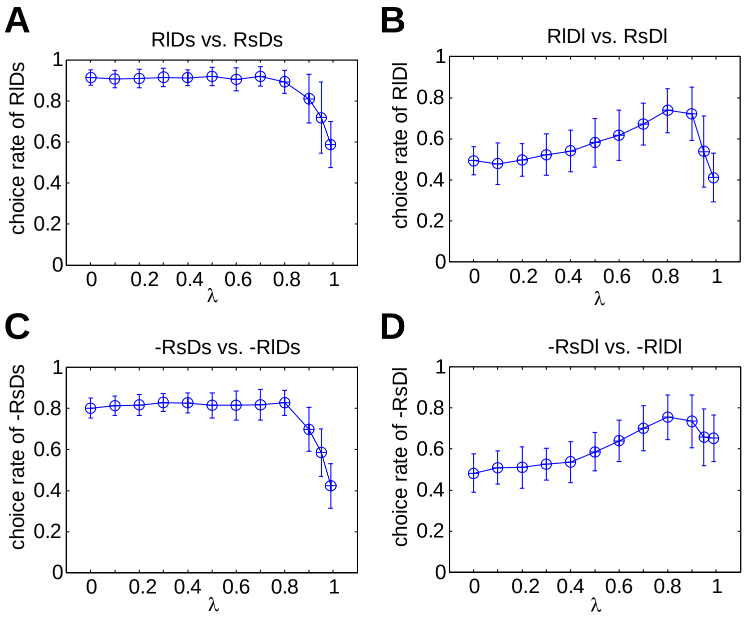

Figure 7: Simulation results for choice rate with the same inputs for each subject. This shows averaged optimal choices against variable λ using the eligibility trace model and Soft Max action selection with parameters α = 0.03 and β = 0.3. Error bars show standard error across subjects (n = 21).

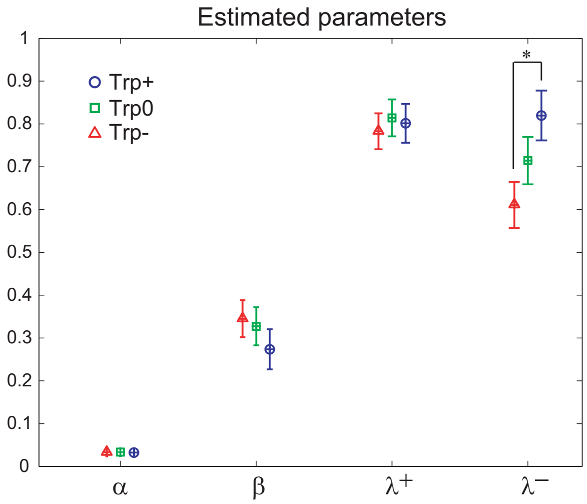

Figure 8: Result of estimated parameters. Error bars show standard error across subjects (n = 21). *: statistically significant difference, p < 0.05 multiple-comparison test after three-way ANOVA with tryptophan levels (n = 3), experiment days (n = 3), and subjects (n = 12).

excessively large eligibility might hinder appropriate association learning, we expected poor learning in both immediate and delayed pairs.

To examine the effects of the trace-decay parameter of an eligibility trace on subjects' behavior, we generated artificial choice data using the eligibility trace model with varying λ. Here. we employed Soft Max as the action selection strategy,

$$P\left(a_{right} \mid s_{t,right}\right) = \frac{\exp\left(\beta V\left(s_{t,right}\right)\right)}{\exp\left(\beta V\left(s_{t,right}\right)\right) + \exp\left(\beta V\left(s_{t,left}\right)\right)}.$$

We varied λ with fixed α = 0.03 and β = 0.3, and plotted the averaged choice rate of correct figures for the whole session (Fig. 7). As for immediate pairs, we found good performance for all λ values less than 0.8 (Fig. 7a, c) while for delayed pairs we found poor performance for small λ values. The optimal λ was 0.8 (Fig. 7b, d). In both immediate and delayed pairs, we found poor performance in λ larger than 0.9. These simulated results were consistent with our expectations, and showed that we can detect the effects on the trace-decay parameter from behavioral data.

3.3.3. Parameter estimation

Using the above model, we estimated subjects' meta-parameters α, β, and λ, maximizing the log likelihood of subjects' action at each tryptophan level. Behavioral results suggested differential involvement in reward and punishment, thus we defined different λ for reward (λ+) and punishment (λ-). Figure 8 shows estimated

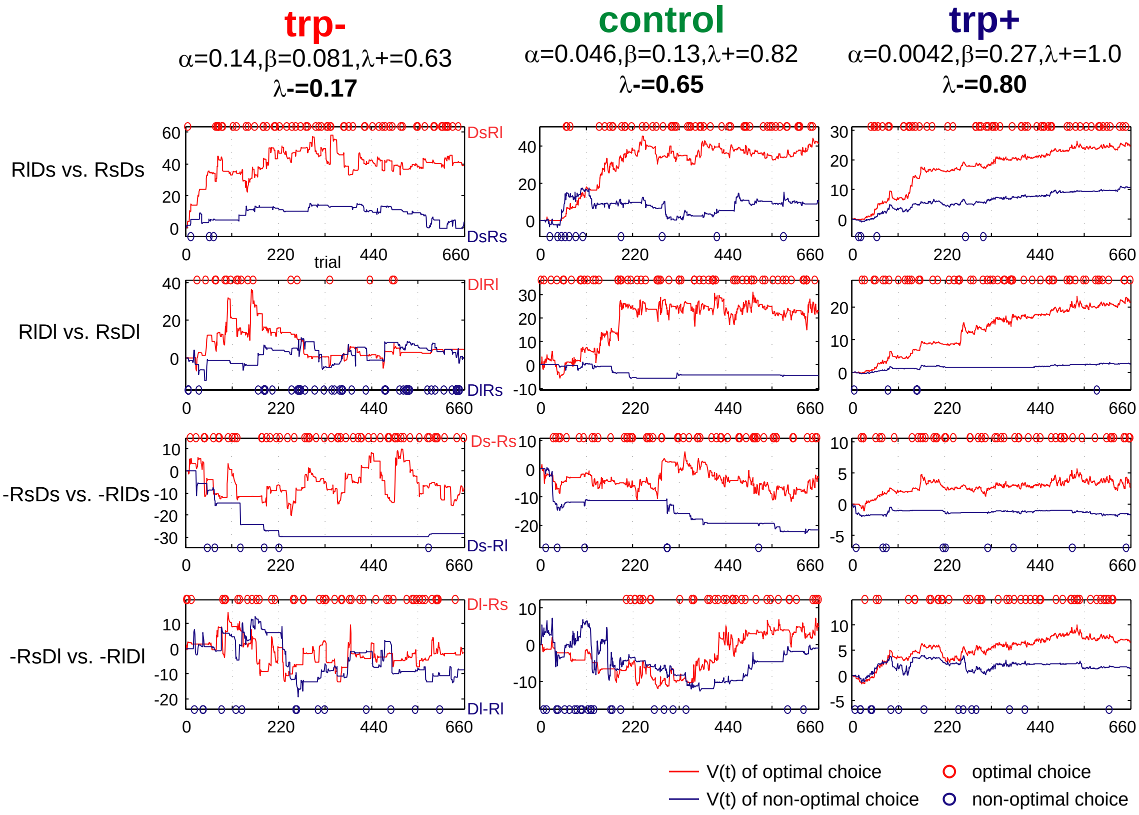

Figure 9: Results of estimated *V* and actual choices. This shows an example of the time course of the value function using estimated meta-parameters in trp- (α = 0.14, β = 0.081, λ+ = 0.63, λ- = 0.17), control (α = 0.046, β = 0.13, λ+ = 0.82, λ- = 0.65), and trp+ (α = 0.0042, β = 0.27, λ+ = 1.0, λ- = 0.80). Lines show estimated value functions, and circles show each subject's actual choices (red: optimal, blue: non-optimal choices).

meta-parameters at each tryptophan level. We found a significant effect of tryptophan levels on estimated l-, where l- was significantly larger in trp+ than trp-. This result was consistent with the behavioral result where we found good learning at high tryptophan levels and poor learning at low serotonin levels. The time course of the estimated value function well explained the subjects' actual choice at each pair (Fig. 9). These results support our hypothesis that serotonin controls time scale of temporal credit assignment.

4. Discussion

4.1. Behavioral changes at different tryptophan levels

Regarding the subjects' choice behavior, we found that the effect of serotonin levels on the acquisition of optimal action was significant in delayed punishment pairs. In the latter session, the choice rate of the delayed small punishment against the delayed large punishment was lower under the tryptophan-depletion condition than the control condition. This indicates that at low serotonin levels, subjects were slow to learn the correct association between delayed punishment and past action. In contrast, in the first session, the choice rate was higher under the tryptophan-loading condition than under other conditions. This shows that at high serotonin levels, subjects successfully learned correct association even in very early sessions. Therefore, these results suggest that serotonin controls the time scale of temporal credit assignment. At low serotonin levels, a small trace parameter results in action-outcome association only in the recent past, thus subjects failed to assign an outcome to the action selected before three trials, although they could learn the association in the present trial. On the contrary, high serotonin levels, a large trace parameter results in action-outcome association also in the distant past, thus subjects could assign an outcome to the action both at present and three trials before.

4.2. Dissociable modulation of serotonin for reward and punishment

Interestingly, we found a significant serotonergic effect on action learning based on delayed punishment, but not reward. Previous studies suggested serotonergic involvement in an aversive system: for example, Dayan and his colleagues proposed that serotonin represents error signals of future punishment prediction (Daw, Kakade et al. 2002). In this study, to test the effects of serotonin manipulation on aversive learning, we used both reward and punishment as reinforcers. If serotonin was involved in only an aversive system, we could expect to find that serotonin levels significantly affect learning from punishment pairs, independent of delay. Our result, where we found that the serotonin level significantly affected the delayed punishment pairs, suggests the multiple effect of serotonin on delay and aversive systems.

To determine the effects of serotonin on delay discounting, we also examined choice behavior in immediate-small vs. delayed-large pairs both in reward (RsDs vs. RlDl) and punishment (-RsDs vs. –RlDl). We did not find any significant differences in choice rates among different serotonin levels (Fig. 10). This result was consistent with our previous study (Chapter 4); although we did not find any significant differences in choice behaviors at different tryptophan levels, as in previous human studies using dietary tryptophan depletion in healthy volunteers, we did observe significant differences in brain activity for reward prediction at different tryptophan levels. These

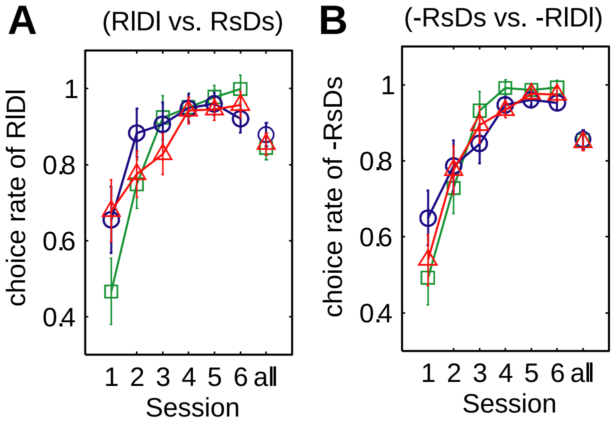

Figure 10: Choice rate of immediate-small vs. delayed-large. These show optimal choice rates of (A) RsDs vs. RlDl, and (B) -RsDs vs. -RlDl. Error bars show standard error across subjects.

different results suggest the different effects of serotonin on the time scale of reward prediction and temporal credit assignment.

Our result was that we found serotonin significantly affects choice behavior in temporal credit assignment for delayed punishment, suggesting that there are distinct systems for delay discounting and temporal credit assignment, or that serotonin differentially affects these systems. In future, we will need to clarify the brain mechanisms and serotonergic effects on delay discounting and temporal credit assignment in the same experimental paradigm.

4.3. Model-based analysis

4.3.1. Eligibility trace model for association learning based on a delayed reinforcer

There have been several studies on instrumental conditioning using a delayed reinforcer, where a longer delay between response and outcome was found to hinder instrumental learning (Dickinson, Watt et al. 1992). Although delay may affect instrumental learning in several ways, such as inhibiting the response outcome contingency, temporally discounting the value of a delayed reinforcer, or inhibiting the of S-R habit process in the instrumental learning model, there is no computational model that can explain this impairment in learning due to a delayed reinforcer. The eligibility trace algorithm, which could well explain the delay effect on action learning in our task, may be one possible candidate for this.

In this model we used $\gamma = 0$ in the TD error equation. In this task, the dynamics of state transition was completely random, so observing the next state s(t+1) was impossible.

Furthermore, the model comparison results indicated that our model with $\gamma = 0$ had a greater likelihood than the model with $\gamma = 1$ and the model without an eligibility trace. Even for $\gamma = 0$, the eligibility trace made it possible to learn the V indirectly.

4.3.2. Parameter estimation

There were three parameters in this model: α, β and λ. We defined the different eligibility trace decay parameter λ for reward and punishment based on our behavioral results that serotonin differentially affected the choice rate for reward and punishment. Because the subjects did not initially know which stimulus corresponded to a reward or a punishment, different trace decay parameters were applied depending on feedback: λ+ for positive feedback and λ- for negative feedback.

We found that the estimated trace decay parameter in negative feedback was significantly lower in the depletion condition than the loading condition. This result indicates that when subjects received negative feedback, they assigned this feedback to only a recent past action when the serotonin level was low, thus they could not learn the correct associations in delayed pairs, with which they needed to assign feedback to the action selected three trials before. In contrast, at high serotonin levels, they also assigned negative feedback to a distant past action, thus they could learn the correct associations in both immediate and delayed pairs. This is consistent with our behavioral results that subjects were slow to learn the correct action in delayed pairs at low serotonin levels, whereas their learning improved with high serotonin levels.

We did not find any significant effects of serotonin on other estimated parameters. To clarify weather learning rate α could cause a difference in the learning of reward and punishment, we estimated α divided into positive and negative feedbacks. The result was that we did not find any significant effect of serotonin on estimated α for both positive and negative feedbacks. Because the Bayesian Information Criteria (BIC) for three models (model 1: α, β, λ+, λ–; model 2: α+, α–, β, λ; model 3: α+, α–, β, λ+, λ–) were not significantly different, we cannot prove that our model is best.

4.4. Brain mechanism of eligibility trace

Previous studies suggested that serotonin systems are involved in learning and memory functions. At low serotonin levels caused by the tryptophan depletion method, healthy human subjects showed impairment in association learning (Park, Coull et al. 1994), delayed recall performance (Riedel, Klaassen et al. 1999), and delayed pattern recognition (Rubinsztein, Rogers et al. 2001). Based on the findings that there was no hindrance to subjects' learning and memory functions, such as working memory and planning, Park and his colleagues suggested that serotonin depletion affects functions involved associated with the hippocampus or the limbic cortex rather than the prefrontal cortex (Park, Coull et al. 1994). Furthermore, lesions of the median raphe nucleus, which sends dense serotonergic projections to the hippocampus, caused a deficit in delayed fear conditioning (Melik, Babar-Melik et al. 2000). Thus, it is possible that serotonin modulates the hippocampus's function. Our results that serotonergic effects were observed in only punishment may reflect the hippocampus functions involved in both aversion and temporal credit assignment.

Chapter 6

General Discussion

In a series of experiments, we demonstrated the hypotheses that 1) there are distinct neural pathways for reward prediction at different time scales, and 2) these pathways are regulated by serotonergic modulation. In this chapter, we discuss the neural implementation of reward prediction at different time scales, and propose a brain network model for this system.

1. Parallel loop organization via the striatum

In this section, we verify the functional differentiation of the striatum for reward prediction at different time scales based on the previous anatomical findings.

1.1. Cortico-striatum connections and reward prediction at different time scales

1.1.1. Graded time-scale maps in the striatum

In our fMRI experiments, we found graded time-scale maps for reward prediction in the striatum. Activities of the ventral parts of the striatum were correlated with short-term reward prediction, and activities of the dorsal parts were correlated with long-term reward prediction. In particular, we found similar gradients in Experiments 1 and 3. In this section (Fig. 1), we verify the graded maps in the striatum that we found in Experiments 1 and 3.

In both maps, the ventral parts are correlated with reward prediction at shorter time scales, denoted by smaller γ values, whereas dorsal parts are correlated with reward prediction at longer time scales, shown as larger γ values. Are both maps graded on the same time scales – i.e. is a particular part of the graded map involved in reward prediction at a particular time scale? If so, a question arises as to whether this map is graded in theoretical time or real time.

To test the ventral-dorsal-directed gradient of both maps, we compared the spatial distribution of both maps along with z-axis with the theoretical time scale, discount factor γ, and the real time scale, decay time constant $\tau = \Delta t/(1-\gamma)$, depending on the duration of a single trial of behavioral task Δt. Figure 2 shows the number of voxels at each z-level that were significantly correlated with reward prediction at each time scale in γ (Fig. 2A) and τ (Fig. 2B) grading. Figure 3 shows graded maps in Exps.1 an 3 for the same slices. Colored lines in Figure 2 indicate the z-level of the median shown in Figure 1. Figure 4 plots the z-level of the median against γ (Fig. 4A) and τ (Fig. 4B). For γ grading, we can see that voxels correlated at the same γ are distributed at about the same z-level, except for γ = 0.99 in Exp. 3. In contrast, for τ grading, there is no consistency in the distribution along the z-axis. This result suggests that the graded maps found in Exps. 1 and 3 are involved in reward prediction at a common “theoretical” time scale.

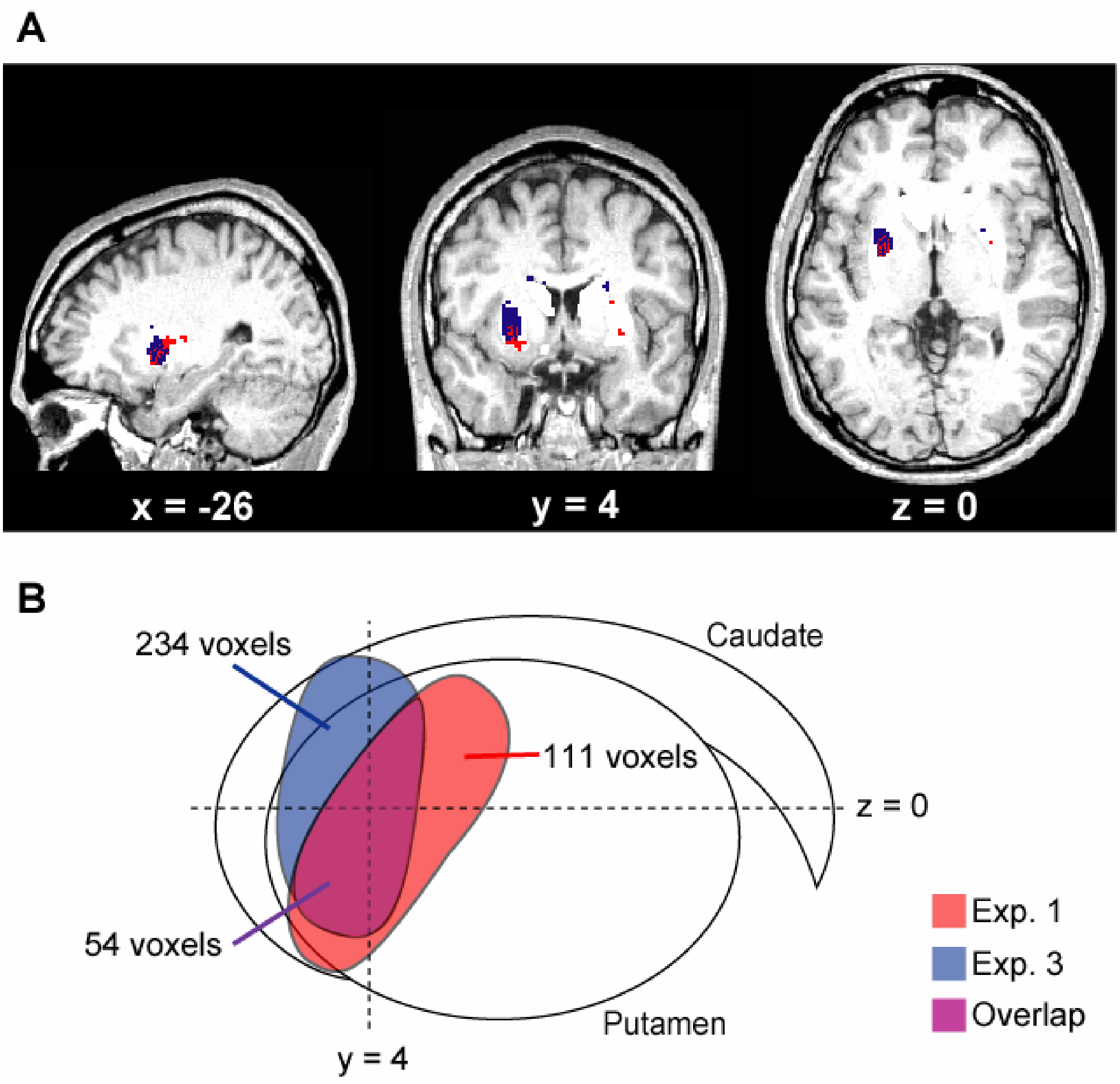

Figure 1: (A) Spatial relationships of graded maps found in Exps. 1 and 3 in the striatum. **(B)** The number of significant voxels with each time scale in the left striatum. The overlapped area is shown by mosaics in **(A)** and violet shading in **(B)**.

These results indicate that particular parts of the striatum are involved in reward prediction at no absolute time scales, one second or one year, but relative time scales, shorter or longer, depending on the task. In the real world, we need to solve problems with variable time scales. At some times we choose an action producing a reward after several seconds or minutes, and at other times, we make decisions that read reward several years later. In this case, the relative grading of a time scale may be effective because the broader region of the striatum can be engaged to compute reward prediction with a limited number of striatal neurons.

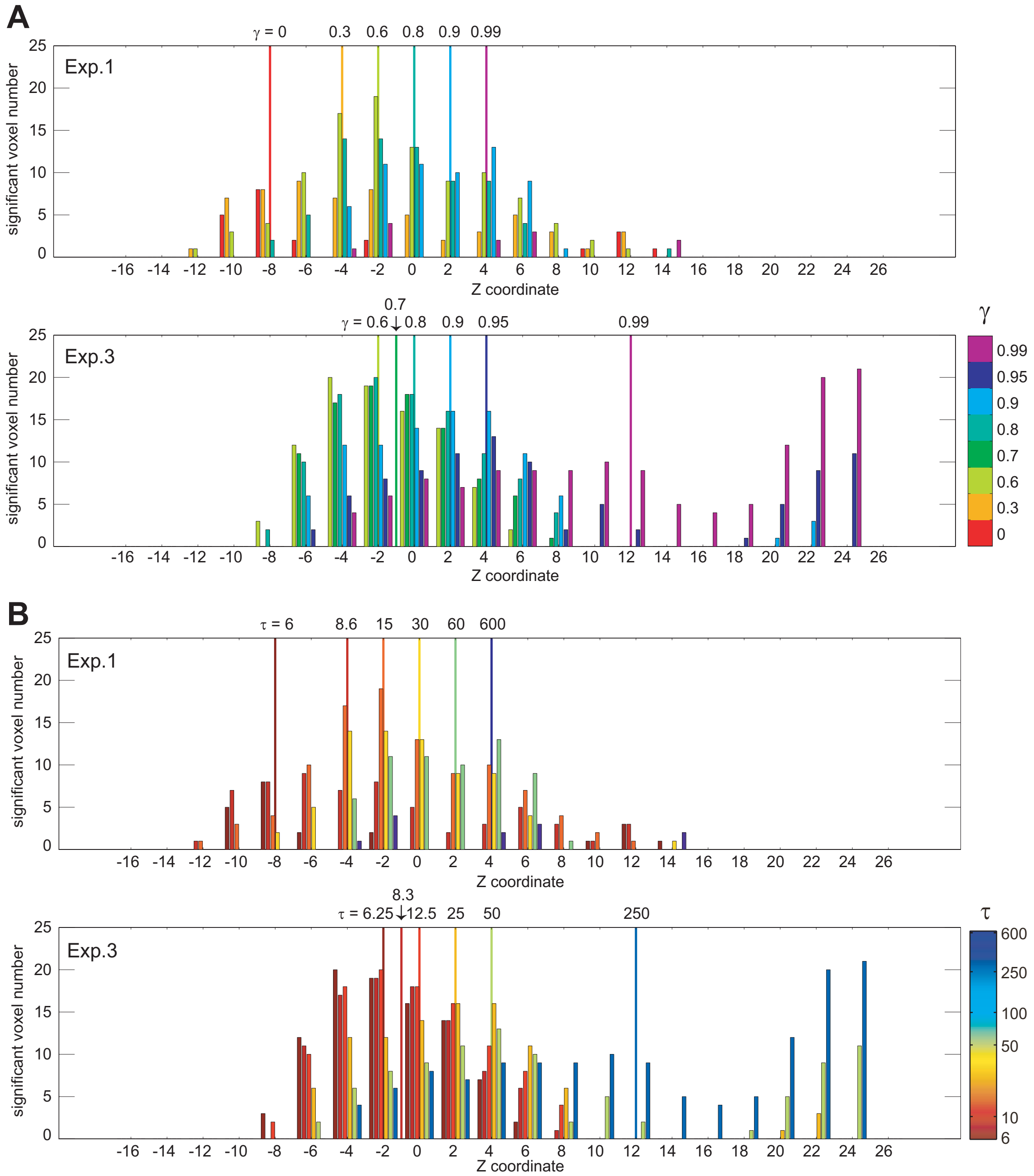

Figure 2: The number of voxels at each z level that were significantly correlated with reward prediction at each time scale in Exps. 1 and 3 in **(A)** γ-grading and **(B)** τ-grading. Colored lines show the median z-coordinate of voxel distribution with each time scale. Although there are gradients of time scales from ventral (low z level) to dorsal (high z level) both in Exps.1 and 3, we can see good consistency of time scales between Exps. 1 and 3 only in g-grading. Note that different color scales are used in γ-grading and τ-grading.

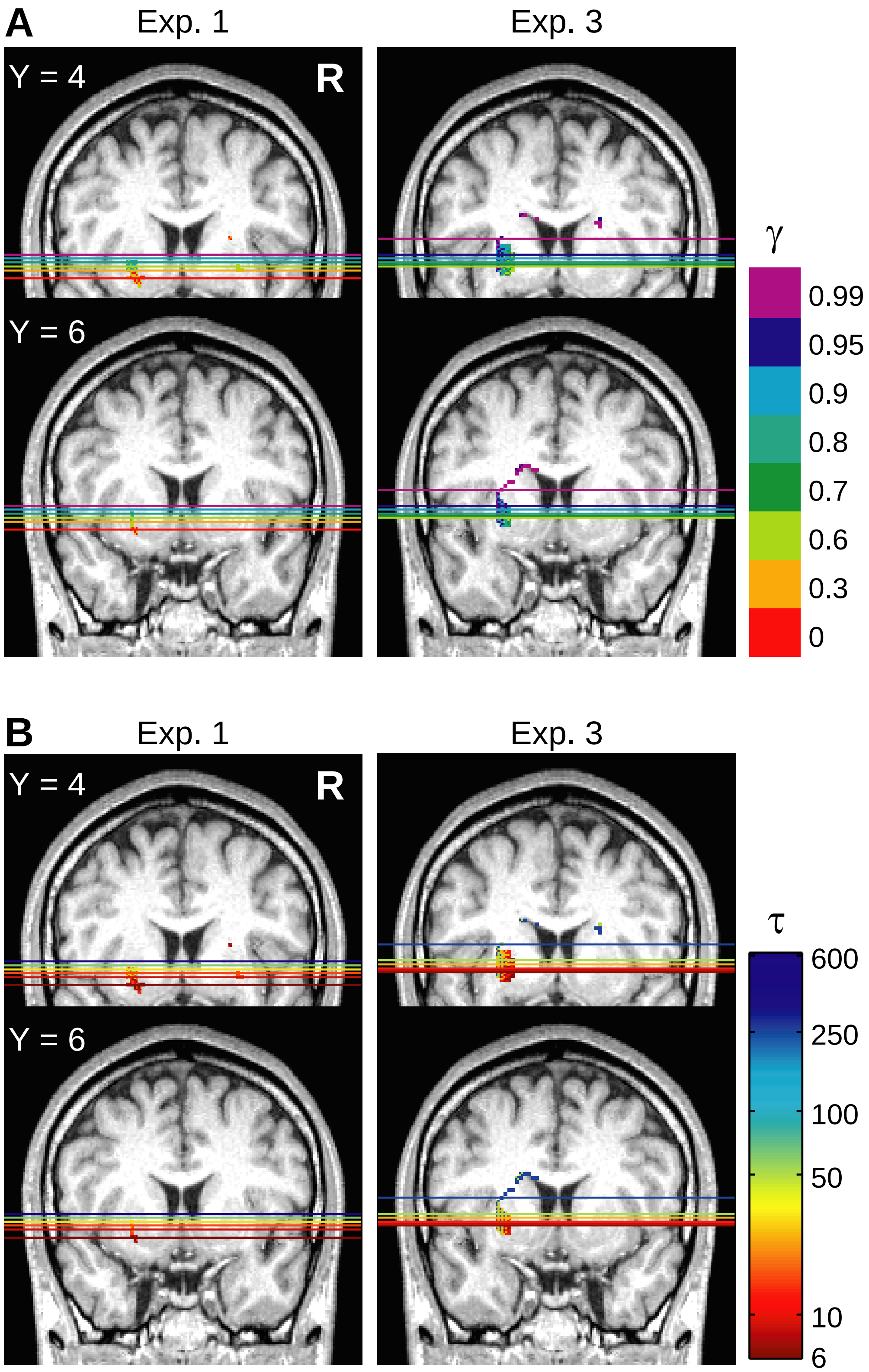

Figure 3: Comparison of graded maps between Exp. 1 and Exp. 3. Colored lines indicate the median z-coordinates shown in Fig. 1.

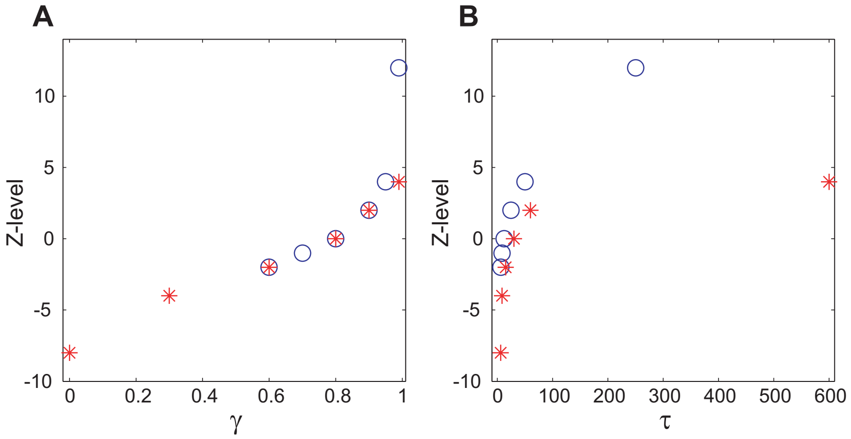

Figure 4: The median z-coordinate of voxel distribution with each time scale. **(A)** In γ-grading, we can see good fitting of data in both Exps. 1 and 3 by the same function. **(B)** In τ-grading, in contrast, this seems difficult to be explained by the same function.

1.1.2. Anatomical evidence of cortico-striatal connections

The graded maps were found in the anterior part of the striatum, mainly the putamen. So, which cortical areas project to this graded map? (In anatomical studies on monkeys, the putamen mainly receives cortical input from motor-related areas such as the primary motor cortex, the premotor cortex, and the supplementary motor area (SMA), and the caudate mainly receives cortical input from the prefrontal cortex.) Recent human studies using the diffusion tensor method succeeded in demonstrating the distinct cortico-striatal connections in humans. They demonstrated that the anterior putamen and caudate received cortical input from the prefrontal and orbitofrontal cortex, and motor-related areas strongly projected to the posterior putamen. Moreover, there was a topographic connection in the anterior putamen and caudate; ventral parts of the anterior putamen and caudate received from the orbitofrontal cortex, and dorsal parts received from the prefrontal cortex (Lehericy, Ducros et al. 2004; Lehericy, Ducros et al. 2004). Fig. 5A, B summarize the cortico-striatum connections in the human striatum. Fig. 5C, D show the spatial relationship between graded maps and cortical connections in the striatum. These figures illustrate that the graded maps are located in the regions that receive input from the OFC and PFC, and that the ventral parts of the map receive from the OFC although the dorsal parts receive from the PFC. These anatomical findings about cortico-striatum connections are quite consistent with results in Exp. 1 that the OFC was more strongly activated in the SHORT condition whereas the DLPFC was more strongly activated in the LONG condition. They also suggest that the loops via the OFC-ventral striatum are involved in reward prediction at shorter time scales while the loops via the PFC-dorsal striatum are involved in reward prediction at longer time scales.

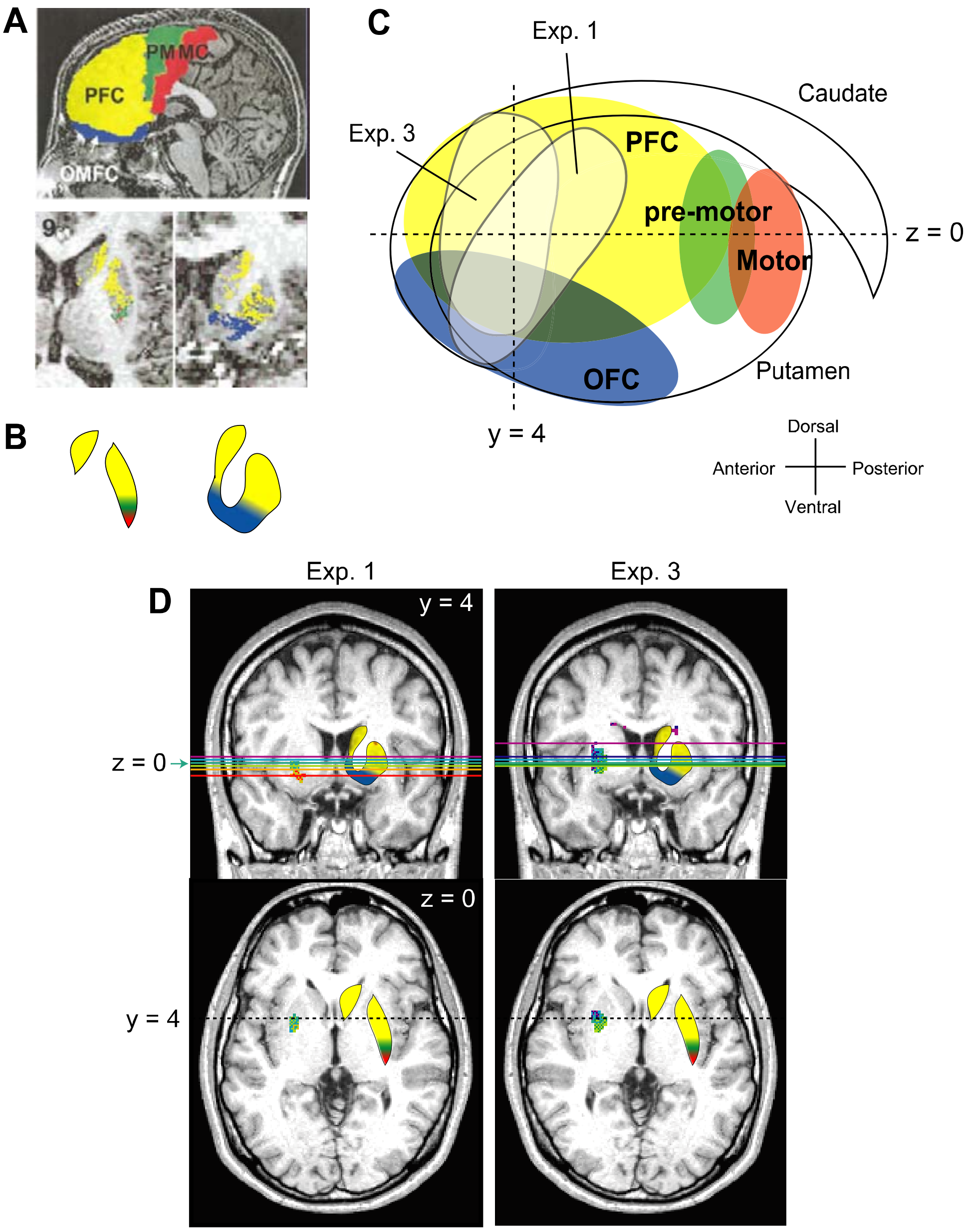

Figure 5: Comparison of graded maps in the striatum and cortico-striatum connection in a human. **(A)** Connectivity maps of the striatum (reprinted from Lehericy et al., 2004; this figure shows the result of a typical single subject). Diffusion tensor tracking was initiated from the orbitofrontal cortex (OFC), prefrontal cortex (PFC), premotor (PM), and motor cortex (MC). Fibers originating from the OFC were directed to the ventral parts of the striatum (blue). Fibers originating from the PFC were occupied most of the anterior and dorsal parts of the striatum (yellow). Fibers originating from the MC were directed to the posterior parts of the putamen (red). Fibers originating from the PM were rostral to the MC fibers although overlapping with these fibers (green). (B) Simplified schemes of the tracking results. (C) Summary diagram, and (D) slices with γ-grading demonstrating the spatial relationship of cortical projections and graded maps in Exps. 1 and 3 in the striatum.

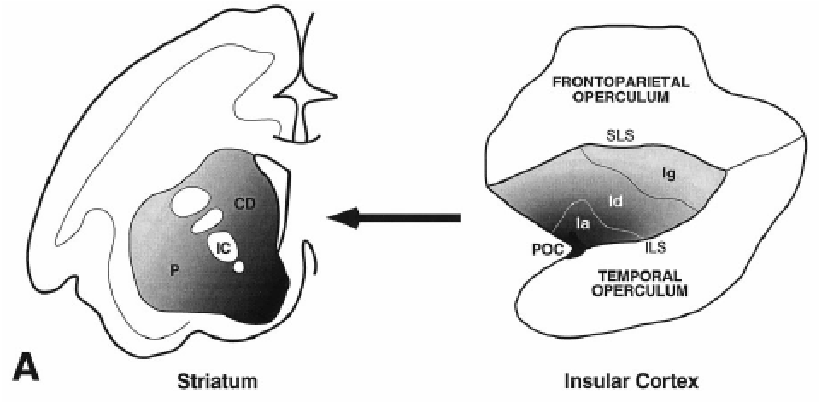

Figure 6: Insula-striatum connection in a monkey (reprinted from Chikama, McFarland et al. 1997)**.** The gray gradients in both the insula and the striatum illustrate the basic organization of insulo-striatal projections from the different cytoarchitectonic regions of the insula.

Based on the reinforcement learning model in the basal ganglia, the cerebral cortex represents the state s needed for reward prediction. The limbic areas involved in emotional processing, e.g. orbitofrontal cortex, may represent state information with a short-range view, for example reward itself. In contrast, the cognitive areas, e.g. the dorsolateral prefrontal cortex, may represent state information needed for long-term reward prediction such as planning and working memory.

There is topographical input from the insula to the striatum (Fig. 6) (Chikama, McFarland et al. 1997). The ventral-anterior parts (agranular area) of the insula send neural fibers to the ventral-anterior parts of the striatum, and the dorsal-posterior part (granular area) of the insula send to the dorsal-posterior part of the striatum. The insula is involved with taste, visceral sensation, and other somatosensory functions, and a lesion of the insula caused deficit in “devaluation” of sated reward (Balleine and Dickinson 2000). Emptiness, sleepiness, satiety, thirst, such “internal states” also strongly affect one’s action selection, and the insula may represent the information on one’s internal state needed for reward prediction.

The striatum receives information on external and internal states represented in different cortical areas with different timescales, and each part of the striatum computes reward prediction at each time scale. The ventral part of the striatum receiving input from the OFC and ventral insula compute reward prediction at shorter time scales, while the dorsal part of the striatum receiving input from the DLPFC and dorsal insula compute reward prediction at longer time scales.

1.1.3. Reinforcement learning model with different time scales of the cortico-basal ganglia loops

Assuming that each part of the striatum computes the value function at each time scale, how are TD errors computed from these different *V*. The projection neurons in the striatum send input to the dopamine neurons in the substantial nigra pars compacta, and also receive a dopaminergic projection from the dopamine neurons in the SNc. This dopaminergic projection acts on presynaptic modulation to the glutamatergic cortico-striatum fibers. These closed striatonigrostriatal pathways have a topography in which input from some particular part of the striatum returns to the same part via some particular part of the SNc (Carpenter, Nakano et al. 1976; Fallon and Moore 1978; Haber, Fudge et al. 2000).

At different settings of γ, different values of δ are generated from different values of *V*, and are used to update *V*. Thus, δ in the SNc and the updating of *V* caused by LTP in the cortico-striatum pathways by dopaminergic projection encoding δ from SNc should be computed for each different γ at each pathway. The topographical striatonigrostriatal pathways allow each loop to process the δ and update the *V* at a particular time scale.

Thus, the graded correlation maps of γ in the striatum may reflect the striatum activities generated by topographic cortical input, representing sensory cues that allow processing *V*, and are generated by topographic dopaminergic projection encoding of δ from the SNc. We found activities in the striatum correlated with reward prediction signals (Exps. 2 and 3) and an error signal (Exp. 1). As mentioned in Chapter 4, in an fMRI experiment both reward prediction and prediction error signals could be detected as a BOLD signal, and which signal is more dominant may depend on the task settings. We found a correlation with reward prediction error in Exp. 1 that might reflect the task setting in which subjects were required to learn the value function from an error signal at each step. On the other hand, the reason we found correlation with reward prediction in Exp. 3 could be because the reward prediction error due to the uncertainty of the number of steps until the reward was relatively small compared with the steady build-up of reward expectation.

In our experiment, we did not find any TD error-related activity in the SNc. This may because the BOLD signal reflects the input and intracortical processing rather than its spiking output itself (Logothetis, Pauls et al. 2001).

From the above discussion, the topographical organization for reward prediction at different time scales can be implemented by the parallel cortico-basal ganglia loops. The problem is to clarify how these loops are modulated by serotonin. In Exp. 3, by combining the tryptophan manipulation and the fMRI measurement, we showed that different parts of the striatum were involved in reward prediction at different time scales, and activities of these parts were differentially affected by serotonin levels. This result suggest that the striatum is the main locus of serotonergic modulation of parallel loop organization. How is serotonin modulated?

1.2. Serotonergic modulation in the striatum

At this point, we would like to discuss the possibility of serotonergic modulation in the striatum by different distributions of 5-HT receptor subtypes. Our result that only ventral part of the striatum was strongly activated at low serotonin level can be explained by the hypothesis that the 5-HT receptors with low affinity for 5-HT are densely distributed in dorsal part of the striatum, because high serotonin level may be needed to activate the dorsal part of the striatum. Our result that only dorsal part of the striatum was strongly activated at high serotonin level can be explained by the hypothesis that the 5-HT autoreceptors with low affinity for 5-HT are densely distributed in ventral part of the striatum, because high serotonin level may be needed to suppress the activities of the ventral part of the striatum.

Actually, autoreceptors 5-HT$_{1B}$ and 5-HT$_{1D}$ are more densely distributed in the ventral part of the striatum than the dorsal part (Compan, Segu et al. 1998), and their affinities are low (5-HT$_{1B}$: 2.290868 ~ 32.380000, 5-HT$_{1D}$: 1.8~11.7; PDSP database, http://kidb.cwru.edu/pdsp.php). Although previous studies showed that 5-HT2A and 2C are more densely distributed in the dorsal parts of the striatum than the ventral parts (Compan, Segu et al. 1998), their affinities have not been investigated in the striatum of primates.

We can clarify this hypothesis by conducting a positron-emission tomography (PET) experiment using radioligands of particular receptor subtypes. For example, we will check the distribution of 5-HT$_1$ and 5-HT$_2$ in the striatum, and check the binding capability of these receptors under different serotonin levels. Recently, a novel method has been developed that visualizes the effects of neuromodulators with finer spatial resolution via enhancing the BOLD signal by injection of some contrast agent

(Mandeville, Marota et al. 1998; Nguyen, Brownell et al. 2000; Chen, Mandeville et al. 2001; Mandeville, Jenkins et al. 2001). This method will enable us to explore serotonergic effects in more detail with finer spatial and temporal information than PET can produce, but we need to conduct more research about this method.

2. Proposal of functional model of serotonin

From the above results and discussions, we propose a possible brain mechanism for reward prediction at different time scales. Based on the reinforcement learning model of reward prediction in the basal ganglia, we attempt to describe parallel mechanisms in detail by examining the topographical organization of the cortico-basal ganglia loops via the striatum.

2.1. Reward prediction processing at different time scales through parallel cortico-basal ganglia loops

The striatum computes reward prediction *V* from cortical input. The loops via the ventral parts of the striatum are involved in reward prediction at small γ, while the loops via the dorsal parts of the striatum are involved in reward prediction at larger γ. TD error δ for each *V* with each γ computed in each part of the striatum is computed in the SNc, and δ is topographically sent to the striatum. **Figure 7** shows a schematic diagram of this parallel mechanism. For simplicity, only two loops, via the ventral and dorsal parts of the striatum, are displayed.

2.2. Serotonergic modulation in the striatum

Serotonin controls the time scale of reward prediction by modulating reward predictive activities in the striatum. The ventral parts of the striatum are strongly activated (or other parts are suppressed) at low serotonin levels, and the ventral loops involved in reward prediction at smaller time scales are selected for actual action selection. In contrast, at high serotonin levels, larger time scales are selected by activating the dorsal parts of the striatum. This serotonergic modulation can be implemented by differential distribution of serotonin receptor subtypes in the striatum; low-affinity receptors are densely distributed in the dorsal part and low-affinity autoreceptors are densely distributed in the ventral part. **Figure 8** shows the scheme of serotonergic modulation in

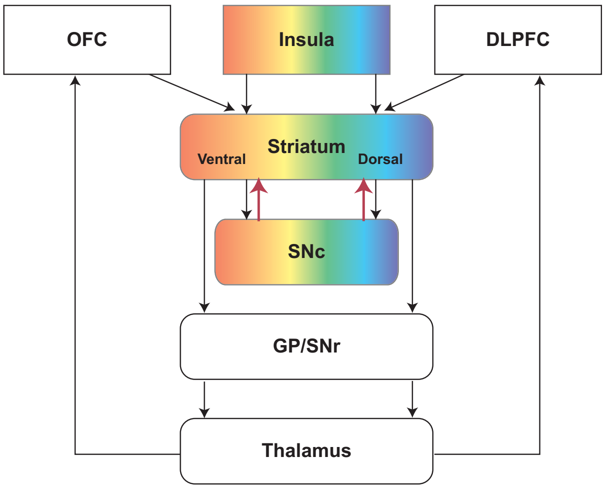

Figure 7: A schematic diagram of the brain areas involved in reward prediction at different time scales. The “limbic loop” (including the lateral OFC and ventral striatum) is involved in short-term reward prediction. The “cognitive and motor loops” (including the DLPFC and dorsal striatum) are involved in long-term reward prediction. Ventroanterior-to-dorsoposterior topographical projections from the insula to the striatum are involved in short-to-long-term reward prediction (rainbow color).

the striatum. In this figure, we simplify the cortical inputs and abbreviate the SNr, GPi, and thalamus. Striatonigrostriatal pathways between the projection neurons (MS neurons) in the striatum and dopamine neurons in the SNc are pictured in detail. Circles located at differently shaded regions of gray in the MS neurons represent different subtypes of serotonin receptors.

In Exps. 3 and 4, we manipulated healthy subjects’ serotonin levels by dietary tryptophan injection. However, in the real world, serotonin levels may be controlled by the external environment or internal state. In a situation that requires long-term reward prediction, loops involved in longer time scales via the dorsal part of the striatum are strongly activated by increasing serotonin levels. On the other hand, in a situation that requires short-term reward prediction, loops involved in shorter time scales via the ventral parts are strongly activated by decreasing serotonin levels. This model needs some mechanism that controls the serotonin level while monitoring environmental changes caused by its own outputs. One candidate is the medial prefrontal cortex, in which we found significant correlation with the value function for actual selected actions. mPFC can modulate serotonergic projection from the dorsal raphe nucleus by reciprocal connection.

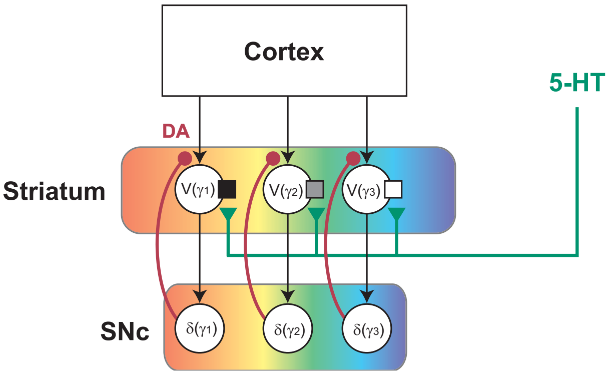

Figure 8: A scheme of serotonergic modulation in the striatum. Different gray-scaled circles in the MS neurons show different subtypes of serotonin receptors. Topographical pathways between the striatum and the substantial nigra pars compacta are involved in short-to-long-term reward prediction error.

2.3. Reinforcement learning module with different time scales

Although we suggest the parallel organization in which each cortico-basal ganglia loop is involved in reward prediction at each time scale, there remains the question how an actual action is generated from these parallel loops.

We propose a modular architecture model in parallel cortico-basal ganglia loops with a modulating system controlled by serotonin receptors **(Fig. 9).** Each loop works as a reinforcement learning module with a particular time scale: loops via the ventral parts of the striatum correspond to modules with smaller γ, while loops via the dorsal parts of the striatum correspond to modules with larger γ. In each module, the value function is computed in the striatum from cortical input, and dopaminergic projection coding TD error signal from SNc updates both the value function and the policy at each time scale. Based on each optimal policy with each time scale, an action candidate is selected through the globus pallidus-thalamus-cortex pathway, and motor command is send to the motor cortex or spinal cord. Since serotonin is involved in module selection, serotonin receptors distributed in modules modulate cortical input to the striatum or striatal activities itself after which a particular module. Serotonin receptors play the role as a gating arrangement for module circuits.

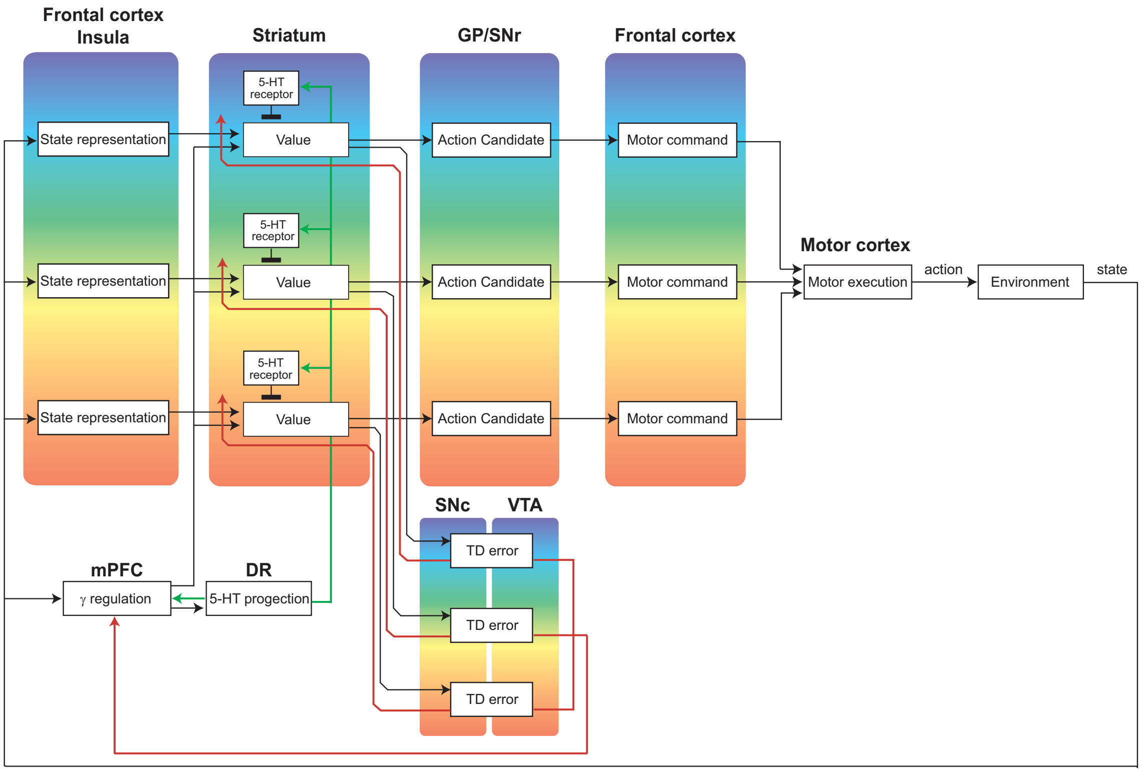

Figure 9: A schematic diagram of the multi-module architecture in cortico-basal ganglia loops. Each loop works as a reinforcement learning module with a particular time scale. There are anatomical topographies between the rainbow-colored areas. Red lines indicate dopaminergic projection, and green lines indicate serotonergic projection.

One particular module is chosen from the parallel modules and this one generates actual action. The mPFC may indirectly choose a particular module by regulating the amount of serotonergic projection from the dorsal raphe to the striatum. How does the mPFC regulate the activity of the dorsal raphe so that the module with the optimal time scale is chosen?

In the reinforcement learning model, it has been shown that a value function learned with a large γ tends to have a large variance, and one learned with a small γ tends to have a small variance (Baxter & Bartlett, 2000; Kakade, 2001). Based on these perspectives, Doya predicted that it would be possible to use the variance of the TD error to regulate the γ. Thus high variability in the dopaminergic activity should have an inhibitory effect on the serotonergic system (Doya 2002). Although these is no direct pathway from the dopamine neurons to the dorsal raphe, another system should mediate this interaction.

The mPFC receives dense dopaminergic input from VTA (Conde, Maire-Lepoivre et al. 1995). Thus, it would be possible that the mPFC receives the TD error signal from the VTA dopamine neurons, controls the activity of the dorsal raphe depending on the variance in the TD error, then regulates the amount of serotonergic projection from the dorsal raphe to the striatum. In a large variance of dopaminergic activity, the mPFC decreases serotonergic projection by suppressing the serotonergic activity in the dorsal raphe. In contrast, in the case of a small variance in the dopaminergic activity, the mPFC increases serotonergic projection by enhancing the activity of the dorsal raphe. In the stratum, a particular module is enhanced or inhibited depending on its affinity for the serotonin receptors. Finally, a particular module is selected to generate actual action.

3. Comparison with previous imaging studies

3.1. Functional distinction of ventral and dorsal parts of the striatum

O'Doherty and his colleagues reported that the ventral striatum was corresponding to a "critic" of the actor-critic model (Sutton and Barto 1998), although the dorsal striatum was corresponding to an "actor" in their fMRI experiment (O'Doherty, Dayan et al. 2004). They computed a TD error signal during the instrumental task, in which subjects were needed to choose one of two fractal figures by pressing a button, and the Pavlovian task, in which subjects pressed the instructed button without choice. They checked the correlation between subjects' brain activities and a TD error signal.

They suggested that the ventral striatum (the ventral putamen and the nucleus accumbens), correlated with a TD error in both tasks, played the role of the critic that might be involved in value prediction in the instrumental and Pavlovian tasks. Note that definition of the "time step" in their model was different to our model. Their TD model described the value $V(t)$ or $\delta(t)$ at a particular time step t during a single trial from stimulus (CS) to reward (USC). Thus, in their model, a discount factor γ indicated temporal discounting within a single trial. In contrast, our model treated each state of Markov decision problem as each time step t. Thus, in our model, a discount factor γ indicated temporal discounting over several future trials. Therefore, their model with $\gamma = 0.99$ might correspond to our model with $\gamma = 0$ considering immediate reward within a single trial. Therefore, our results, that the ventral part of the striatum was correlated with δ (Exps. 1, 2) and V (Exp. 3) with $\gamma = 0$, were well consistent with their result.

They also suggested that the dorsal striatum (the caudate head), correlated with a TD error in only instrumental task, was corresponding to the actor involved in action selection in the instrumental task. In our results, we found significant correlation with V or δ mainly in the putamen, not in the caudate head. Thus, our suggestion that the dorsal part of the striatum was involved in long-term reward prediction is not competitive to their suggestion that caudate head was corresponding to the actor. As they suggested, the caudate head was involved in action selection by learning of stimulus-response or stimulus-response-reward association (Haruno, Kuroda et al. 2004). It may be that we did not find any significant correlation with V or δ in the caudate head because we computed V and δ independent of action (see each definitional equation of V and δ). In contrast, in Exp. 2, we estimated the value function by Bayesian estimation with subjects' action as estimation parameter. Our result that we found significant correlation with this action-dependent value function in the caudate head was well consistent with their suggestion about caudate head.

3.2. Neural substrates for short-term and long-term reward prediction

McClure and his colleagues measured subjects' brain activities while subjects made choices between early-smaller and delayed-larger monetary rewards by questionnaire method. They reported that the "limbic" areas including the ventral striatum and the mPFC were involved in short-term choice, although the "cognitive" areas including the DLPFC and the lateral parietal cortex were involved in long-term choice (McClure, Laibson et al. 2004).

They focused on two types of choices, "immediate trial" and "delayed trial". The immediate trials indicated that an early-smaller reward was delivered as soon as they finished experiment, while the delayed trials indicated that both rewards were delivered after 2 weeks or later. They suggested that the activities of the ventral striatum (the ventral putamen) and the mPFC, significantly correlated with regressor of immediate trial = 1 and delayed trial = -0.5, were corresponding to the β area, in which the value of all delayed rewards were strongly discounted by β of "quasi-hyperbolic" model, $V = \beta\delta^t u$ (u: reward). This result was well consistent with our result that the ventral part of the putamen was more strongly activated in SHORT condition in Exp. 1.

They also suggested that the DLPFC and the lateral parietal cortex, significantly correlated with regressor of 1 for all trials, were corresponding to the δ area, in which the value of all rewards were equally slowly discounted by δ. We also found stronger activation in the DLPFC and the lateral parietal cortex in LONG condition in Exp. 1, thus their result was well consistent with our result. However, because they used regressor of 1 for all trials, it may be that these δ areas were just involved in simple choice problem, not only in long-term reward prediction. This possibility supported their stronger activations in difficult choices (difference between smaller and larger reward was 5~25%) than easy choices. It may be that their δ regressors failed to detect the dorsal part of the striatum involved in log-term reward prediction in our study.

4. Serotonergic modulation of the eligibility trace

As described in Chapter 4, there are two possible reasons why serotonin depletion causes impulsive choice. One is that a shorter time scale of reward prediction reduces the value of a delayed reward; the other is that a shorter temporal credit assignment hinders association learning between action and delayed reward. Although there are many studies on impulsivity focusing on the former, to the best of our knowledge no study has yet demonstrated the latter possibility.

We compared the learning effects between different delays under different serotonin levels, and demonstrated that while low serotonin level caused slow learning, high serotonin levels enhanced learning in delayed punishment. This result suggests that serotonin controls temporal credit assignment of present outcome to past action.

4.1. Brain mechanism of the eligibility trace

Because a lesion of the serotonin system causes impairments in memory and learning functions involved in the hippocampus rather than the prefrontal cortex (Park, Coull et al. 1994), we suggested that the hippocampus represents the eligibility trace in Chapter 4. Assuming that the hippocampus is involved in the eligibility trace function, how can the computational processes be implemented in the hippocampus networks? In updating V, dopaminergic projection encoding δ to cortico-striatum pathways from the SNc changes the synaptic plasticity of cortico-striatum connections. In the eligibility trace model, $V(t+1) = V(t) + \alpha*\delta(t)*e(t)$, eligibility trace *e*, a weight of δ, should modulate δ through the anatomical connection extending from the hippocampus to the dopaminergic system. There are topographical projections from the hippocampus to the nucleus accumbens, and the nucleus accumbens sends output to the nigrostriatal pathways (Somogyi, Bolam et al. 1981; Groenewegen and Russchen 1984). This pathway makes it possible to update *V* with the eligibility trace represented in the hippocampus by modulating δ via the nucleus accumbens and SNc. Furthermore, recent research has demonstrated that a lesion of the nucleus accumbens caused impairment in association learning from delayed reward (Cardinal and Cheung 2005). From these findings, we suggest that the hippocampus-nucleus accumbens-nigra-striatum pathway is involved in learning from a delayed reinforcer using the eligibility trace. The hippocampus receives dense serotonergic projection from the medium raphe, thus the eligibility trace can be modulated by serotonergic system. We will next try to extend a detailed physiology model of the eligibility trace in the hippocampus.

4.2. Delay discounting and eligibility trace

From a series of experiments, we suggest that serotonin can control the time scale of reward prediction and temporal credit assignment. To explore the effect of serotonin on each function, we set the task environments in which subjects performed reward prediction without any learning factors in Exp. 3, and subjects could learn the association without future reward prediction in Exp. 4. In the real world, where we need both reward prediction and association learning, however, these systems are closely linked to each other. Serotonergic modulation of the time scale of these forward and backward views may elicit consistent solutions from both systems and thus facilitate effective action learning.

5. Summary

By using noninvasive measurements of human brain activities and pharmacological manipulation based on the computational model, we demonstrated that reinforcement learning theory could be represented in the brain as a model of reward based decision making and action selection, and that serotonin can control the time scale of reward prediction. To understand the brain function at the system level, non-invasive measurement, in which we can display a direct relationship between a certain region of the brain and some targeted function, is more useful than other methods such as the electrophysiological recording for detecting neuronal phenomena. We should note, however, that the results of non-invasive measurements are generated by neuronal and molecular phenomena such as changes in intracellular calcium concentration or synaptic plasticity. Although we could detect changes in brain activity generated by manipulation of the serotonin level, we did not show how serotonin modulates brain activity. To clarify this problem, we need to perform some invasive experiments, for example using microdialysis, a serotonin receptor agonist/antagonist, or electrophysiological recording. We will try to establish a computational model that can explain human decision behaviors using neuronal or molecular behaviors.

Acknowledgement

I wish to acknowledge the many individual who have contributed in various ways to this thesis, including prof. Mitsuo Kawato (ATR), prof. Shin Ishii, prof. Kotarou Minato (NAIST), assistant prof. Tomohiro Shibata (NAIST), Dr. Kazuyuki Samejima (ATR and Tamagawa univ.), Dr. Nicolas Schweighofer (ATR and USC), and all collaborators at Hiroshima university, including Prof. Shigeto Yamawaki, Dr. Yasumasa Okamoto, Dr. Go Okada, Dr. Kazutaka Ueda, Dr. Shuji Asahi, and Dr. Kazuhiro Shishida. Finally, I thank Dr. Kenji Doya for supervising my doctoral thesis. I learned so much from him, and I'm sincerely grateful to him.

References

Ainslie, G. (1975). "Specious reward: a behavioral theory of impulsiveness and impulse control." Psychol Bull **82**(4): 463-96.

Alexander, G. E., M. R. DeLong, et al. (1986). "Parallel organization of functionally segregated circuits linking basal ganglia and cortex." Annu Rev Neurosci **9**: 357-81.

Baker, S. C., R. D. Rogers, et al. (1996). "Neural systems engaged by planning: a PET study of the Tower of London task." Neuropsychologia **34**(6): 515-26.

Balleine, B. W. and A. Dickinson (2000). "The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory." J Neurosci **20**(23): 8954-64.

Bayer, H. M. and P. W. Glimcher (2005). "Midbrain dopamine neurons encode a quantitative reward prediction error signal." Neuron **47**(1): 129-41.

Bechara, A., H. Damasio, et al. (2000). "Emotion, decision making and the orbitofrontal cortex." Cereb Cortex **10**(3): 295-307.

Berns, G. S., S. M. McClure, et al. (2001). "Predictability modulates human brain response to reward." J Neurosci **21**(8): 2793-8.

Bizot, J., C. Le Bihan, et al. (1999). "Serotonin and tolerance to delay of reward in rats." Psychopharmacology (Berl) **146**(4): 400-12.

Bjork, J. M., D. M. Dougherty, et al. (1999). "The effects of tryptophan depletion and loading on laboratory aggression in men: time course and a food-restricted control." Psychopharmacology (Berl) **142**(1): 24-30.

Bjork, J. M., D. M. Dougherty, et al. (2000). "Differential behavioral effects of plasma tryptophan depletion and loading in aggressive and nonaggressive men." Neuropsychopharmacology **22**(4): 357-69.

Breiter, H. C., I. Aharon, et al. (2001). "Functional imaging of neural responses to expectancy and experience of monetary gains and losses." Neuron **30**(2): 619-39.

Cardinal, R. N. and T. H. Cheung (2005). "Nucleus accumbens core lesions retard instrumental learning and performance with delayed reinforcement in the rat." BMC Neurosci **6**(1): 9.

Cardinal, R. N., D. R. Pennicott, et al. (2001). "Impulsive choice induced in rats by lesions of the nucleus accumbens core." Science **292**(5526): 2499-501.

Cardinal, R. N., C. A. Winstanley, et al. (2004). "Limbic corticostriatal systems and delayed reinforcement." Ann N Y Acad Sci **1021**: 33-50.

Carpenter, L. L., G. M. Anderson, et al. (1998). "Tryptophan depletion during continuous CSF sampling in healthy human subjects." Neuropsychopharmacology **19**(1): 26-35.

Carpenter, M. B., K. Nakano, et al. (1976). "Nigrothalamic projections in the monkey demonstrated by autoradiographic technics." J Comp Neurol **165**(4): 401-15.

Cavada, C., T. Company, et al. (2000). "The anatomical connections of the macaque monkey orbitofrontal cortex. A review." Cereb Cortex **10**(3): 220-42.

Celada, P., M. V. Puig, et al. (2001). "Control of dorsal raphe serotonergic neurons by the medial prefrontal cortex: Involvement of serotonin-1A, GABA(A), and glutamate receptors." J Neurosci **21**(24): 9917-29.

Chen, Y. C., J. B. Mandeville, et al. (2001). "Improved mapping of pharmacologically induced neuronal activation using the IRON technique with superparamagnetic blood pool agents." J Magn Reson Imaging **14**(5): 517-24.

Chikama, M., N. R. McFarland, et al. (1997). "Insular cortical projections to functional regions of the striatum correlate with cortical cytoarchitectonic organization in the primate." J Neurosci **17**(24): 9686-705.

Compan, V., L. Segu, et al. (1998). "Selective increases in serotonin 5-HT1B/1D and 5-HT2A/2C binding sites in adult rat basal ganglia following lesions of serotonergic neurons." Brain Res **793**(1-2): 103-11.

Conde, F., E. Maire-Lepoivre, et al. (1995). "Afferent connections of the medial frontal cortex of the rat. II. Cortical and subcortical afferents." J Comp Neurol **352**(4): 567-93.

Crean, J., J. B. Richards, et al. (2002). "Effect of tryptophan depletion on impulsive behavior in men with or without a family history of alcoholism." Behav Brain Res **136**(2): 349-57.

Critchley, H. D., C. J. Mathias, et al. (2001). "Neural activity in the human brain relating to uncertainty and arousal during anticipation." Neuron **29**(2): 537-45.

Daw, N. D., S. Kakade, et al. (2002). "Opponent interactions between serotonin and dopamine." Neural Netw **15**(4-6): 603-16.

Delgado, P. L., D. S. Charney, et al. (1990). "Serotonin function and the mechanism of antidepressant action. Reversal of antidepressant-induced remission by rapid depletion of plasma tryptophan." Arch Gen Psychiatry **47**(5): 411-8.

Dickinson, A., A. Watt, et al. (1992). "Free-operant Acquisition with Delayed Reinforcement." The Quarterly Journal of Experimental Psychology **45B**(3): 241-258.

Doya, K. (2000). "Complementary roles of basal ganglia and cerebellum in learning and motor control." Curr Opin Neurobiol **10**(6): 732-9.

Doya, K. (2002). "Metalearning and neuromodulation." Neural Netw **15**(4-6): 495-506.

Eagle, D. M., T. Humby, et al. (1999). "Effects of regional striatal lesions on motor, motivational, and executive aspects of progressive-ratio performance in rats." Behav Neurosci **113**(4): 718-31.

Elliott, R., K. J. Friston, et al. (2000). "Dissociable neural responses in human reward systems." J Neurosci **20**(16): 6159-65.

Elliott, R., J. L. Newman, et al. (2003). "Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study." J Neurosci **23**(1): 303-7.

Evenden, J. L. (1999). "Varieties of impulsivity." Psychopharmacology (Berl) **146**(4): 348-61.

Evenden, J. L. and C. N. Ryan (1996). "The pharmacology of impulsive behaviour in rats: the effects of drugs on response choice with varying delays of reinforcement." Psychopharmacology (Berl) **128**(2): 161-70.

Evenden, J. L. and C. N. Ryan (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**(4): 413-21.

Fallon, J. H. and R. Y. Moore (1978). "Catecholamine innervation of the basal forebrain. IV. Topography of the dopamine projection to the basal forebrain and neostriatum." J Comp Neurol **180**(3): 545-80.

Goldberg, T. E. and D. R. Weinberger (2004). "Genes and the parsing of cognitive processes." Trends Cogn Sci **8**(7): 325-35.

Graybiel, A. M. (1990). "Neurotransmitters and neuromodulators in the basal ganglia." Trends Neurosci **13**(7): 244-54.

Groenewegen, H. J. and F. T. Russchen (1984). "Organization of the efferent projections of the nucleus accumbens to pallidal, hypothalamic, and mesencephalic structures: a tracing and immunohistochemical study in the cat." J Comp Neurol **223**(3): 347-67.

Haber, S. N., J. L. Fudge, et al. (2000). "Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum." J Neurosci **20**(6): 2369-82.

Haber, S. N., K. Kunishio, et al. (1995). "The orbital and medial prefrontal circuit through the primate basal ganglia." J Neurosci **15**(7 Pt 1): 4851-67.

Halldin, C., B. Gulyas, et al. (2001). "Brain radioligands--state of the art and new trends." Q J Nucl Med **45**(2): 139-52.

Hanakawa, T., M. Honda, et al. (2002). "The role of rostral Brodmann area 6 in mental-operation tasks: an integrative neuroimaging approach." Cereb Cortex **12**(11): 1157-70.

Hariri, A. R., V. S. Mattay, et al. (2002). "Serotonin transporter genetic variation and the response of the human amygdala." Science **297**(5580): 400-3.

Haruno, M., T. Kuroda, et al. (2004). "A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task." J Neurosci **24**(7): 1660-5.

Hikosaka, K. and M. Watanabe (2000). "Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards." Cereb Cortex **10**(3): 263-71.

Hikosaka, O., H. Nakahara, et al. (1999). "Parallel neural networks for learning sequential procedures." Trends Neurosci **22**(10): 464-71.

Ho, M. Y., S. Mobini, et al. (1999). "Theory and method in the quantitative analysis of "impulsive choice" behaviour: implications for psychopharmacology." Psychopharmacology (Berl) **146**(4): 362-72.

Houk, J. C., J. L. Adams, et al. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. Models of information processing in the basal ganglia. J. C. Houk, J. L. Davis and D. G. Beiser. Cambridge, Mass., MIT Press**:** 249-270.

Kawagoe, R., Y. Takikawa, et al. (1998). "Expectation of reward modulates cognitive signals in the basal ganglia." Nat Neurosci **1**(5): 411-6.

Knutson, B., C. M. Adams, et al. (2001). "Anticipation of increasing monetary reward selectively recruits nucleus accumbens." J Neurosci **21**(16): RC159.

Knutson, B., G. W. Fong, et al. (2003). "A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI." Neuroimage **18**(2): 263-72.

Koepp, M. J., R. N. Gunn, et al. (1998). "Evidence for striatal dopamine release during a video game." Nature **393**(6682): 266-8.

Kurth-Nelson, Z. and A. D. Redish (2004). micro-agents: action-selection in temporally dependent phenomena using temporal difference learning over a collective belief structure. Society for Neuroscience, San Diego, 2004 Abstract Viewer/Itinerary Planner.

Lehericy, S., M. Ducros, et al. (2004). "3-D diffusion tensor axonal tracking shows distinct SMA and pre-SMA projections to the human striatum." Cereb Cortex **14**(12): 1302-9.

Lehericy, S., M. Ducros, et al. (2004). "Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans." Ann Neurol **55**(4): 522-9.

Logothetis, N. K., J. Pauls, et al. (2001). "Neurophysiological investigation of the basis of the fMRI signal." Nature **412**(6843): 150-7.

Mandeville, J. B., B. G. Jenkins, et al. (2001). "Regional sensitivity and coupling of BOLD and CBV changes during stimulation of rat brain." Magn Reson Med **45**(3): 443-7.

Mandeville, J. B., J. J. Marota, et al. (1998). "Dynamic functional imaging of relative cerebral blood volume during rat forepaw stimulation." Magn Reson Med **39**(4): 615-24.

Martin-Ruiz, R., M. V. Puig, et al. (2001). "Control of serotonergic function in medial prefrontal cortex by serotonin-2A receptors through a glutamate-dependent mechanism." J Neurosci **21**(24): 9856-66.

Matsumoto, K., W. Suzuki, et al. (2003). "Neuronal correlates of goal-based motor selection in the prefrontal cortex." Science **301**(5630): 229-32.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. Quantitative analyses of behavior. M. L. Commons, J. E. Mazur, J. A. Nevin and H. Rachlin. Hillsdale, Erlbaum. **5:** 55-73.

Mazur, J. E. (2001). "Hyperbolic value addition and general models of animal choice." Psychol Rev **108**(1): 96-112.

McClure, S. M., G. S. Berns, et al. (2003). "Temporal prediction errors in a passive learning task activate human striatum." Neuron **38**(2): 339-46.

McClure, S. M., D. I. Laibson, et al. (2004). "Separate neural systems value immediate and delayed monetary rewards." Science **306**(5695): 503-7.

Melik, E., E. Babar-Melik, et al. (2000). "Median raphe nucleus mediates forming long-term but not short-term contextual fear conditioning in rats." Behav Brain Res **112**(1-2): 145-50.

Mesulam, M. M. and E. J. Mufson (1982). "Insula of the old world monkey. III: Efferent cortical output and comments on function." J Comp Neurol **212**(1): 38-52.

Middleton, F. A. and P. L. Strick (2000). "Basal ganglia and cerebellar loops: motor and cognitive circuits." Brain Res Brain Res Rev **31**(2-3): 236-50.

Mijnster, M. J., A. G. Raimundo, et al. (1997). "Regional and cellular distribution of serotonin 5-hydroxytryptamine2a receptor mRNA in the nucleus accumbens, olfactory tubercle, and caudate putamen of the rat." J Comp Neurol **389**(1): 1-11.

Mobini, S., S. Body, et al. (2002). "Effects of lesions of the orbitofrontal cortex on sensitivity to delayed and probabilistic reinforcement." Psychopharmacology

(Berl) **160**(3): 290-8.

Mobini, S., T. J. Chiang, et al. (2000). "Effect of central 5-hydroxytryptamine depletion on inter-temporal choice: a quantitative analysis." Psychopharmacology (Berl) **149**(3): 313-8.

Mobini, S., T. J. Chiang, et al. (2000). "Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement." Psychopharmacology (Berl) **152**(4): 390-7.

Nguyen, T. V., A. L. Brownell, et al. (2000). "Detection of the effects of dopamine receptor supersensitivity using pharmacological MRI and correlations with PET." Synapse **36**(1): 57-65.

O'Doherty, J., H. Critchley, et al. (2003). "Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices." J Neurosci **23**(21): 7931-9.

O'Doherty, J., P. Dayan, et al. (2004). "Dissociable roles of ventral and dorsal striatum in instrumental conditioning." Science **304**(5669): 452-4.

O'Doherty, J. P., P. Dayan, et al. (2003). "Temporal difference models and reward-related learning in the human brain." Neuron **38**(2): 329-37.

O'Doherty, J. P., R. Deichmann, et al. (2002). "Neural responses during anticipation of a primary taste reward." Neuron **33**(5): 815-26.

Owen, A. M., J. Doyon, et al. (1996). "Planning and spatial working memory: a positron emission tomography study in humans." Eur J Neurosci **8**(2): 353-64.

Pagnoni, G., C. F. Zink, et al. (2002). "Activity in human ventral striatum locked to errors of reward prediction." Nat Neurosci **5**(2): 97-8.

Park, S. B., J. T. Coull, et al. (1994). "Tryptophan depletion in normal volunteers produces selective impairments in learning and memory." Neuropharmacology **33**(3-4): 575-88.

Pears, A., J. A. Parkinson, et al. (2003). "Lesions of the orbitofrontal but not medial prefrontal cortex disrupt conditioned reinforcement in primates." J Neurosci **23**(35): 11189-201.

Poulos, C. X., J. L. Parker, et al. (1996). "Dexfenfluramine and 8-OH-DPAT modulate impulsivity in a delay-of-reward paradigm: implications for a correspondence with alcohol consumption." Behav Pharmacol **7**(4): 395-399.

Reynolds, J. N. and J. R. Wickens (2002). "Dopamine-dependent plasticity of corticostriatal synapses." Neural Netw **15**(4-6): 507-21.

Riedel, W. J., T. Klaassen, et al. (1999). "Tryptophan depletion in normal volunteers produces selective impairment in memory consolidation." Psychopharmacology

(Berl) **141**(4): 362-9.

Rogers, R. D., B. J. Everitt, et al. (1999). "Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms." Neuropsychopharmacology **20**(4): 322-39.

Rogers, R. D., A. M. Owen, et al. (1999). "Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex." J Neurosci **19**(20): 9029-38.

Rolls, E. T. (2000). "The orbitofrontal cortex and reward." Cereb Cortex **10**(3): 284-94.

Rubinsztein, J. S., R. D. Rogers, et al. (2001). "Acute dietary tryptophan depletion impairs maintenance of "affective set" and delayed visual recognition in healthy volunteers." Psychopharmacology (Berl) **154**(3): 319-26.

Salomon, R. M., H. L. Miller, et al. (1997). "Lack of behavioral effects of monoamine depletion in healthy subjects." Biol Psychiatry **41**(1): 58-64.

Samejima, K., Y. Ueda, et al. (2005). "Representation of action-specific reward values in the striatum." Science **310**(5752): 1337-40.

Schultz, W., P. Dayan, et al. (1997). "A neural substrate of prediction and reward." Science **275**(5306): 1593-9.

Seymour, B., J. P. O'Doherty, et al. (2004). "Temporal difference models describe higher-order learning in humans." Nature **429**(6992): 664-7.

Shidara, M., T. G. Aigner, et al. (1998). "Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials." J Neurosci **18**(7): 2613-25.

Shidara, M. and B. J. Richmond (2002). "Anterior cingulate: single neuronal signals related to degree of reward expectancy." Science **296**(5573): 1709-11.

Somogyi, P., J. P. Bolam, et al. (1981). "Monosynaptic input from the nucleus accumbens--ventral striatum region to retrogradely labelled nigrostriatal neurones." Brain Res **217**(2): 245-63.

Sutton, R. S. and A. G. Barto (1998). Reinforcement Learning.

Sutton, R. S., Barto, A. G. (1998). Reinforcement learning. Cambridge, MA, MIT press.

Tanaka, S. C., K. Doya, et al. (2004). "Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops." Nature Neuroscience **7**(8): 887-893.

Thorndike, E. L. (1911). Animal intelligence; experimental studies. New York, The Macmillan Company.

Tremblay, L., J. R. Hollerman, et al. (1998). "Modifications of reward expectation-related neuronal activity during learning in primate striatum." J Neurophysiol **80**(2): 964-77.

Tremblay, L. and W. Schultz (2000). "Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex." J Neurophysiol **83**(4): 1864-76.

Ullsperger, M. and D. Y. von Cramon (2003). "Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging." J Neurosci **23**(10): 4308-14.

Williams, W. A., S. E. Shoaf, et al. (1999). "Effects of acute tryptophan depletion on plasma and cerebrospinal fluid tryptophan and 5-hydroxyindoleacetic acid in normal volunteers." J Neurochem **72**(4): 1641-7.

Wogar, M. A., C. M. Bradshaw, et al. (1993). "Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers." Psychopharmacology (Berl) **111**(2): 239-43.

Young, S. N. and S. Gauthier (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**(4): 323-8.

Young, S. N., S. E. Smith, et al. (1985). "Tryptophan depletion causes a rapid lowering of mood in normal males." Psychopharmacology (Berl) **87**(2): 173-7.

業績リスト

学術論文

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. *Nature Neuroscience* Vol.7, No.8, pp. 887 - 893

Tanaka SC, Doya K, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2005). Brain mechanism of reward prediction in predictable and unpredictable environment. *Neural Networks* (*submitted)*

Tanaka SC, Schweighofer N, Asahi S, Okamoto Y, Yamawaki S, and Doya K (2006) Serotonin differentially regulates reward predictive striatal activities in short and long time scales. (*submitted)*

国際会議 **(**査読あり**)**

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004). Different cortico-basal ganglia loops specialize in reward prediction on different time scales. *Advances in neural information processing systems 16 (NIPS)*, pp. 701 - 708, MIT Press

テクニカル・レポート

田中 沙織, 銅谷 賢治, 岡田 剛, 上田 一貴, 岡本 泰昌, 山脇 成人 (2002). 長期と短期の報酬予測に伴う脳活動の fMRI 測定. 電気情報通信学会技術研究報告書, Vol.102, No.157, pp. 37-42

田中 沙織, 銅谷 賢治, 岡田 剛, 上田 一貴, 岡本 泰昌, 山脇 成人 (2003). 空間的情報を含むマルコフ決定課題を用いた長期と短期の報酬予測に伴う脳活動の fMRI 測定. 電気情報通信学会技術研究報告書, Vol.103, No.92, pp. 1-6

受賞

2005 年　日本神経回路学会論文賞
2005 年　日本神経回路学会研究賞
2005 年　日本神経回路学会奨励賞