Available online at www.sciencedirect.com

ScienceDirect

Neural Networks 19 (2006) 1242–1254

Neural Networks

www.elsevier.com/locate/neunet

2006 Special Issue

Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning

Masahiko Haruno*, Mitsuo Kawato

ATR Computational Neuroscience Laboratories, Department of Computational Neurobiology, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Received 18 November 2005; accepted 1 June 2006

Abstract

The brain's most difficult computation in decision-making learning is searching for essential information related to rewards among vast multimodal inputs and then integrating it into beneficial behaviors. Contextual cues consisting of limbic, cognitive, visual, auditory, somatosensory, and motor signals need to be associated with both rewards and actions by utilizing an internal representation such as reward prediction and reward prediction error. Previous studies have suggested that a suitable brain structure for such integration is the neural circuitry associated with multiple cortico-striatal loops. However, computational exploration still remains into how the information in and around these multiple closed loops can be shared and transferred. Here, we propose a "heterarchical reinforcement learning" model, where reward prediction made by more limbic and cognitive loops is propagated to motor loops by spiral projections between the striatum and substantia nigra, assisted by cortical projections to the pedunculopontine tegmental nucleus, which sends excitatory input to the substantia nigra. The model makes several fMRI-testable predictions of brain activity during stimulus-action-reward association learning. The caudate nucleus and the cognitive cortical areas are correlated with reward prediction error, while the putamen and motor-related areas are correlated with stimulus-action-dependent reward prediction. Furthermore, a heterogeneous activity pattern within the striatum is predicted depending on learning difficulty, i.e., the anterior medial caudate nucleus will be correlated more with reward prediction error when learning becomes difficult, while the posterior putamen will be correlated more with stimulus-action-dependent reward prediction in easy learning. Our fMRI results revealed that different cortico-striatal loops are operating, as suggested by the proposed model.

Keywords: Reinforcement learning; Cortico-striatal loop; Information integration; Reward prediction; Dopamine; PPTN; Model-based neuroimaging; Stimulus-action-reward association

1. Introduction

Whenever faced with a decision-making situation, our brain has to integrate multiple sources of information. That is, it needs to explore and choose relevant features among limbic, cognitive, visual, auditory, somatosensory, and motor signals to characterize the given situation and evaluate the relative advantages of possible behaviors in terms of reward to select the appropriate action. This integration is particularly important during learning because contextual cues must be associated with both reward and action to achieve valuable and rapid decision-making.

The striatum has been regarded as a key player in multimodal integration during learning for the following two reasons. First, since it constitutes distinct loop circuits with many cortical areas including prefrontal, medial frontal, cingulate, and premotor and primary motor cortices (Alexander, Crutcher, & Delong, 1990; Gerardin et al., 2003; Parthsarathy, Schall, & Graybiel, 1992; Selemon & Goldman-Rakic, 1985; Takada, Tokuno, Nambu, & Inase, 1998), it therefore has access to multimodal information. Second, the striatum has strong reciprocal connections with the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA), so it can therefore regulate multimodal inputs by the effect of dopamine (Hollerman & Schultz, 1998; Schultz, Apicella, Scarnati, & Ljungberg, 1992; Schultz & Dickinson, 2000; Takikawa, Kawagoe, & Hikosaka, 2004). However, it is

* Corresponding author.
E-mail address: mharuno@atr.jp (M. Haruno).

0893-6080/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2006.06.007

well-known that each cortico-striatal loop is closed; a loop originating from a cortical region goes around the striatum, the globus pallidus, and the thalamus and then projects back to the original location without much blending (Middleton & Strick, 2000). This anatomical specificity poses an intriguingly difficult computational question: how can the information processing conducted by different closed cortico-striatal loops be integrated? In other words, an additional mechanism is needed to link different cortico-striatal loops, except the intracortical connections that are quite remote from the center of reinforcement learning.

In this paper, we first propose a heterarchical reinforcement-learning model in which dopamine projection from the SNc and the VTA to the striatum plays a key role in this integration. Dopamine projections are used as a messenger shared by multiple cortico-striatal loops. It is widely accepted that at the timing of reward delivery, dopamine neurons in the SNc and the VTA encode some error of reward prediction, which is computed by the interaction between the cortex, the striatum, the SNc/VTA, and the pedunculopontine tegmental nucleus (Hollerman & Schultz, 1998; Schultz et al., 1992; Houk, Adams, & Barto, 1995; Kobayashi, Inoue, Yamamoto, Isa, & Aizawa, 2002; Montague, Dayan, & Sejnowski, 1996; Takikawa et al., 2004). Furthermore, in the loop circuit between the striatum and the midbrain (SNc/VTA), the dorsal striatum influences a limited midbrain region, but it is affected by a larger midbrain region (Haber, Fudge, & McFarland, 2000; Haber, 2003). Additionally, the pedunculopontine tegmental nucleus (PPTN) and the laterodorsal tegmental receive cortical excitatory inputs and projects to the SNc and the VTA by excitatory synapses, respectively (Oakman, Faris, Kerr, Cozzari, & Hartman, 1995). Based on these two facts, which indicate spatially heterogeneous distribution of reward prediction and reward prediction error, the heterarchical reinforcement learning model proposes that early-learning-stage reward prediction in the striatum and the SNc/VTA is only obtained from very coarse description of a given situation (mainly limbic and associative information such as that obtained by rough inferences or guesses from rewards), which can be spread as dopamine inputs to motor cortico-striatal loops. Then, detailed and reliable reward predictions can be computed, which incorporate richer and more detailed information including motor commands. This gradual propagation of reward prediction and reward prediction error contributes to increase the efficiency of the reinforcement learning of complex tasks.

The proposed model can make several predictions regarding neural activity in the component brain structures such as the cortico-striatal loops, the SNc, the VTA, and the PPTN. In the experimental section of this paper, we will focus on predictions for the cortico-striatal loops because they are testable within the spatial and temporal resolutions of fMRI. More specifically, we hypothesize that activity in the caudate nucleus and cognitive cortical areas is correlated with reward prediction error from the beginning because the error is mainly comprised of rough inferences from rewards. On the other hand, activity in the putamen and motor-related areas is expected to be correlated with stimulus-action-dependent reward prediction because the association of reward and motor information is required. Furthermore, when we make learning more difficult, it is not only predictable that activity in the caudate nucleus and cognitive cortical areas shows more correlation with reward prediction error because learning is at a more initial phase but also that the locus of activity changes systematically within the striatum. We examined these predictions by using model-based fMRI (Haruno et al., 2004; Haruno & Kawato, 2006; O'Doherty et al., 2004), where a computational model is utilized to estimate such internal subject variables as reward prediction and reward prediction error, which are then used in the correlation analysis of brain activity data (it is straightforward to apply this approach to other modalities including electrophysiology, optical imaging, etc).

In the rest of the paper, we will first provide qualitative explanations of the heterarchical reinforcement learning model and its predictions on fMRI signals during stimulus-action-reward association learning. Next, we will move to more formal details of the heterarchical reinforcement learning model. Finally, we will discuss the experimental setting and results using fMRI in comparison with the model's predictions.

2. Heterarchical reinforcement learning model

2.1. *Temporal-difference learning model*

A reinforcement learning model, which is an appealing theoretical framework that might explain the essential aspects of animal and human learning, is only guided by reward and penalty information. The temporal-difference learning (TD) model (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998) is the most established formulation of reinforcement learning, and it tries to learn a so-called value function $V(s, t)$ that represents the expectation of the discounted sum of future rewards starting from context (state) s at time t. In this equation, $r(t)$ and γ are actual rewards at time t and a discount factor for future rewards:

$$V(s, t) = E\left\{\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right)\right\}.$$

During learning, the TD model updates $V(s, t)$ as follows in proportion to TD error $\delta(t)$:

$$V(s, t) \leftarrow V(s, t) + \alpha\delta(t).$$

$$\delta(t) = \gamma V(s_{t+1}, t + 1) + r(t) - V(s_t, t).$$

The first and second terms of TD error are the estimations of $V(s, t)$ after receiving a reward at time t, while the third term is the same estimation before receiving the reward. Therefore, when the estimation of $V(s, t)$ is complete, TD error $\delta(t)$ should be 0. The TD model changes the estimation of $V(s, t)$ in the direction minimizing TD error $\delta(t)$. An important variant of the TD model is Q-learning, which employs $Q(s, a, t)$ rather than $V(s, t)$, that represents the expectation of the discounted sum of future rewards starting from context (state) s and action a at time t. The TD model is attractive not only because simple computation of TD error can be used to handle the

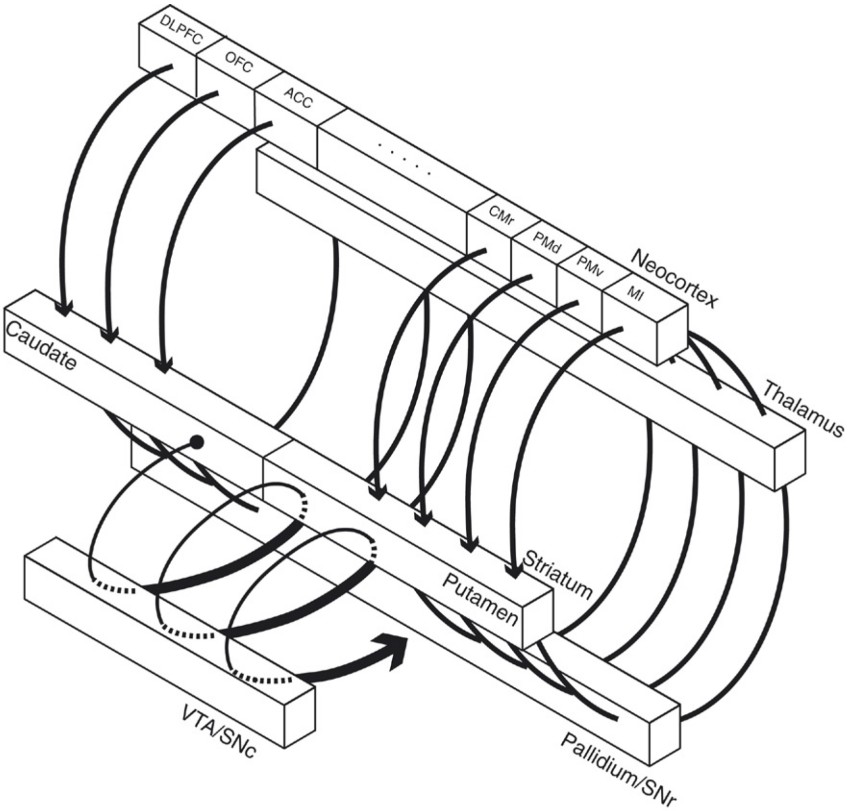

Fig. 1. Neural circuits consisting of cortico-striatal and striato-nigral loops. DLPFC: dorsolateral prefrontal cortex, OFC: orbitofrontal cortex, ACC: anterior cingulate cortex, CMr: rostral cingulate motor area, PMd: dorsal premotor cortex, PMv: ventral premotor cortex, MI: primary motor cortex.

delay between an action and a reward (called the "temporal credit assignment problem") (Sutton & Barto, 1998) but also because some relatively complicated discrete problems such as backgammon can be successfully solved at the same levels as human experts. Furthermore, a recent neurophysiological study (Schultz & Dickinson, 2000) suggests that dopamine neurons represent TD error.

2.2. *Basic ideas and predictions for fMRI experiments*

Applying the standard algorithms to such examples of reinforcement learning as above, even to the medium-sized, real-world sensory-motor problems that animals and humans face daily, reveals them to be so unsatisfactorily slow that they cannot be practically utilized. Only by introducing a modular and hierarchical reinforcement learning architecture could such relatively large-scale problems as a standing robot be solved within a reasonable amount of learning trials (Morimoto & Doya, 1999). Driven by the same necessity from practical applications, studies of hierarchical reinforcement learning have been active (Dayan & Hinton, 1993; Singh, 1992; Sutton, Singh, Precup, & Ravindran, 1999), but they have met with only fairly limited success; in real-world problems it is unrealistic to design the hard hierarchy of modules (subgoals) that strictly controls how to switch reward predictions and subgoals between higher and lower hierarchical layers. This limitation suggests that in actual animal reinforcement learning, heterogeneous reward predictions are softly combined depending on their time-varying significance instead of such a hard hierarchy.

Here, we propose a heterarchical reinforcement learning model as a computational model of animal reinforcement learning. The model's most important aspect is that reward prediction made by limbic and cognitive loops is propagated to motor loops by spiral projections between the striatum and substantia nigra. This mechanism allows the reward prediction error coarsely represented by the cognitive loop to guide the learning of more detailed reward predictions in motor loops, which incorporate more detailed information such as motor commands. Fig. 1 illustrates the anatomical connections of the two key components, the cortico-striatal and striato-nigral loops, based on work from Haber's group (Haber et al., 2000; Haber, 2003). The illustrated spiral striato-nigral connection is one of the two anatomical and physiological pieces of evidence that support our heterarchical reinforcement-learning model. The striatum maintains closed loop circuits with many cortical areas including the medial prefrontal, the orbitofrontal, the dorsolateral prefrontal, the anterior cingulate, the cingulate motor, the dorsal and ventral premotor, and the primary motor cortices. A remarkable difference between the caudate nucleus and the ventral striatum from the putamen is that the caudate and the ventral striatum mainly receive projection from the cognitive areas of the cortex, whereas the putamen is projected more by such motor-related areas as the premotor and cingulate motor areas (Price, Carmichael, & Drevets, 1996). In contrast to the closed nature of the cortico-striatal loop, afferent and efferent projections are asymmetric in the striato-nigral loop (Haber et al., 2000; Haber, 2003). That is, the dorsal striatum influences a limited region of the SNc and the VTA, but is affected by a larger region. On the other hand, the ventral striatum receives limited midbrain input, but projects to a large region. The efference part of these spiral connections enables reward prediction coded by caudate neurons constructed at coarse limbic and cognitive levels to be utilized in initially guiding the acquisition of more detailed and sophisticated representation of reward prediction at the motor level. Simultaneously, because of the afference limb of the spiral connections, the reward prediction error coarsely represented by the dorsomedial midbrain can guide more detailed reinforcement learning at the putamen.

Even within these cortico-striatal and striato-nigral parts of the model, we can still derive predictions of fMRI signal change during stimulus-action-reward association learning because the striatum is voluminous enough to be tested by fMRI. First, the activity of the caudate nucleus, the prefrontal and anterior cingulate cortices are expected to be correlated with reward prediction error because error computed using coarse representation should initially be dominant in the early stage of learning. This tendency should be more prominent when learning is difficult. On the other hand, the putamen, the rostral premotor cortex, and the rostral cingulate motor cortex are expected to show a correlation of activity with stimulus-action-dependent reward prediction because, in the late stage of learning, acquisition of stimulus-action-dependent reward prediction with fine motor command representation has already been achieved while guided by the above cognitive reward prediction error. This tendency should be more prominent when learning is easy. Furthermore, because of internal differences in the midbrain projections within the caudate nucleus and the putamen (Haber et al., 2000; Haber, 2003), correlation with the stimulus-action-dependent reward prediction in the putamen should be posterior in easy tasks but anterior in difficult tasks when we control learning difficulty. In contrast, correlation with reward prediction error in the caudate nucleus is supposed to be located posterior-lateral in easy tasks but anterior-medial in difficult tasks. When the task is very easy, the activity is also likely to be found in the putamen because reward prediction error for fine-grained prediction is dominant during learning.

2.3. Formal explanations of the heterarchical reinforcement learning model

In this section, we introduce the full components of the heterarchical reinforcement learning model and provide formal explanations of how it works. Fig. 2 illustrates the neural circuits of the heterarchical reinforcement-learning model. In this figure, we explain the model's essential features in a very simplified diagram of neural circuits around the multiple cortico-striatal and striato-nigral loops. Here, for simplicity, only two neocortical areas (prefrontal and motor) are assumed to be connected to the two corresponding parts of the striatum (caudate and putamen) by one-to-one connections. We assume that the privacy of these two parallel loops is violated by the following two spread and feedforward connections in computing reward prediction error. One is the spread and efference connection from the caudate nucleus to the ventrolateral part of the SNc, as described above and illustrated in Fig. 1 (Haber, 2003). The other is the spread and feedforward excitatory connection from the prefrontal cortex to PPTN, which then projects to the ventrolateral part of SNc by excitatory synapses. This part of the model is motivated by physiological studies on monkey PPTN neural firings by Kobayashi and his colleagues (Kobayashi et al., 2002; Kobayashi, Okada, Inoue, Yamamoto, & Isa, 2005).

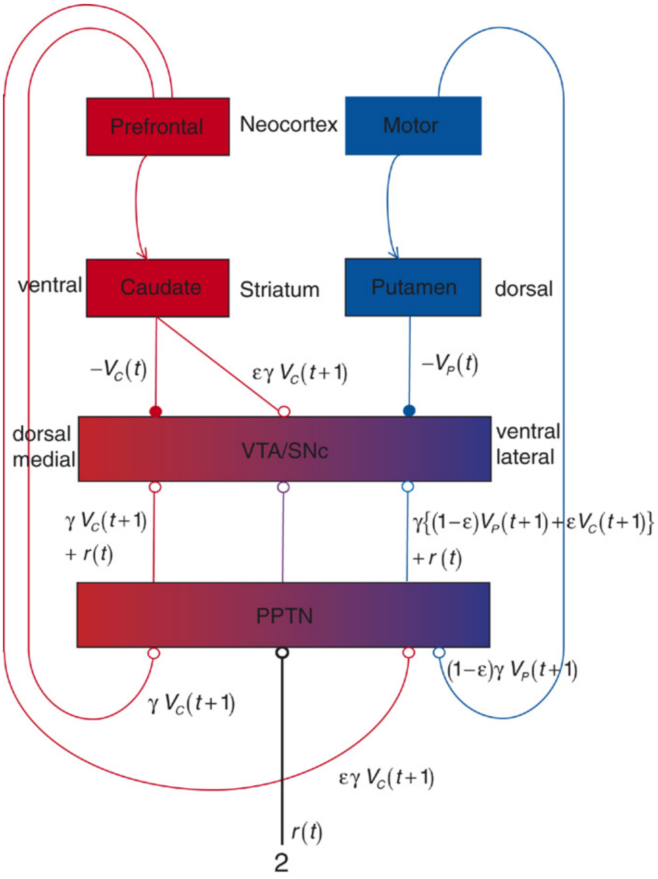
Prefrontal
Neocortex
Motor
ventral
Caudate
Striatum
Putamen
dorsal
$-V_C(t)$
$\varepsilon\gamma V_C(t+1)$
$-V_P(t)$
dorsal medial
VTA/SNc
ventral lateral
$\gamma V_C(t+1)$ $+r(t)$
$\gamma\{(1-\varepsilon)V_P(t+1)+\varepsilon V_C(t+1)\}$ $+r(t)$
PPTN
$\gamma V_C(t+1)$
$(1-\varepsilon)\gamma V_P(t+1)$
$\varepsilon\gamma V_C(t+1)$
$r(t)$
2

Fig. 2. A basic heterarchical reinforcement-learning model consisting of caudate–prefrontal and motor-putamen loops, ventral tegmental area (VTA), substantia nigra pars compacta (SNc), and pedunculopontine tegmental nucleus (PPTN). Open and closed circles show excitatory (disinhibitory) and inhibitory projections, respectively. Vc and Vp represent a coarse and fine value function in the caudate nucleus and putamen, respectively, and r is a reward. γ is a discount factor for future rewards and ε is a time-varying weighting factor between Vc and Vp.

For the dorsomedial part of SNc/VTA, which is inhibited by caudate GABAnergic innervation, temporal-difference reward prediction error is computed as the difference between the excitatory PPTN inputs, that is, the summation of the time-advanced value function (i.e., $t+1$) from the cortex (assumed fast) and the primary reward signal, and the inhibitory caudate input, that is, the value function computed by the caudate as follows. The state variables in value functions are omitted for simplicity.

$$\begin{aligned}\delta_c(t) &= \mathrm{PPTN}_C(t) - \mathrm{Caudate}(t)\\ &= \{\gamma V_C(t+1) + r(t)\} - V_C(t)\\ &= \gamma V_C(t+1) + r(t) - V_C(t).\end{aligned} \tag{1}$$

Here and below, C and P are indices for variables computed by the caudate and putamen, respectively. Computation for the caudate–prefrontal loop is purely private, and the caudate equation (1) is identical to ordinary reinforcement-learning models for temporal-difference (TD) error. However, in contrast to previous basal-ganglia reinforcement-learning models (Brown, Bullock, & Grossberg, 1999; Houk et al., 1995; Montague et al., 1996), we assume that PPTN excitatory inputs originating from the cortex provide the dorsomedial part of the VTA/SNc, the summation of time-advanced value-function terms, and the reward. Thus, the learning equation of the caudate part is identical to the regular TD model, but the neural circuit implementing this differs from previous models.

On the other hand, we assume that temporal-difference reward prediction error for the putamen–motor loop is affected by the caudate–prefrontal loop, as shown below and in the right side of Fig. 2.

$$\begin{aligned}\delta_P(t) &= [\mathrm{PPTN}_P(t) + \mathrm{Caudate}(t)] - \mathrm{Putamen}(t)\\ &= [\{(1-\varepsilon)\gamma V_P(t+1) + r(t)\} + \varepsilon\gamma V_C(t+1)]\\ &\quad - V_P(t)\\ &= \varepsilon\gamma V_C(t+1) + (1-\varepsilon)\gamma V_P(t+1)\\ &\quad + r(t) - V_P(t),\end{aligned} \tag{2}$$

where ε is a positive decreasing function of time, starting from 1 and approaching 0 as time t goes from 0 to infinity. One plausible choice of ε is given as follows:

$$\varepsilon(t) = \int_0^\infty \mathrm{e}^{-s/\tau}|\delta_P(t-s)|\mathrm{d}s \Big/ \int_0^\infty \mathrm{e}^{-s/\tau}|\delta_P(-s)|\mathrm{d}s. \tag{3}$$

Here τ is a time constant for averaging high-frequency fluctuations in $\delta_P(t)$ and should be shorter than the longest time constant of decrease in $\delta_P(t)$ (i.e., learning time constant). As explained below, $\delta_P(t)$ is expected to decrease and approach 0 as learning proceeds. Thus, $\varepsilon(t)$ possesses the above desirable properties. Because we assume that the dopamine neurons of the ventrolateral part of the midbrain encode $\delta_P(t)$, $\varepsilon(t)$ can be realized as a normalized and leaky integrated average of the ventrolateral dopamine level, possibly implemented as a long-term effect of dopamine. By comparing the above Eqs. (1) and (2), notice that the difference between the caudate and putamen equations is in the time-advanced value-function term (first term in (1) and first and second terms in (2)). That is, the temporally-advanced value-function term for putamen is the $\varepsilon(t)$-weighted average of those from caudate and putamen.

At the end of learning when both $\delta_P(t)$ and $\varepsilon(t)$ approach 0, Eq. (2) reverts to the standard TD-reinforcement learning algorithm. On the other hand, at the beginning of learning when $\delta_P(t)$ is large and $\varepsilon(t)$ is close to 1, Eq. (2) can be approximated as follows:

$$\delta_P(t) \cong \gamma V_C(t+1) + r(t) - V_P(t). \tag{4}$$

This approximate equation (4) can be interpreted in at least three different manners. In the first interpretation, the first term represents the subgoal given to the lower level putamen loop provided by the upper level caudate loop, as an addition to the primary reward. This is in the spirit of the hierarchical reinforcement-learning framework and assumes that a more cognitive caudate–prefrontal loop could provide an intermediate subgoal, or highly processed reward information, to the lower putamen–motor loop. The second interpretation is supervised learning. Although discussing reinforcement learning, we could reinterpret Eq. (4) as a definition of error-driven supervised learning. Then we can argue that the summation of the first and second terms is playing the role of teaching signal for the putamen value function. The first term is the coarse-grained approximation of the discounted value function computed by the caudate–prefrontal loop, and it plays a partial role as the teaching signal. Here, we assume that while utilizing more abstract, coarse, limbic, and cognitive representations of the environment, the caudate–prefrontal loop can relatively quickly learn to approximate value function according to (1). On the other hand, learning in the putamen–motor loop is slow because their representations are very detailed, both in space and time, and contain motor commands, and thus are of a much higher dimension. Consequently, if the putamen–motor loop were to learn the value function by itself, it would take an indefinitely long time due to the curse of dimensionality, but with such semi-supervised learning (4) at the beginning, reinforcement learning could be dramatically accelerated. The third interpretation is the coarse-to-fine approximation of the value function. We assume that in the prefrontal cortex, the representation of state space is much coarser than in the motor cortex. Then value function learning that uses a much smaller number of basis functions with lower dimension should proceed much faster than value-function approximation with a much larger number of basis functions with very large dimensions. Thus, caudate–prefrontal loop learning is much faster than putamen–motor loop learning, but its ultimate approximation capability is inferior. If the caudate value-function approximation remains the non-changing subgoal or the teaching signal, the putamen value-function approximation could not become better than that of caudate. Weighting by $\varepsilon(t)$ attenuates the influence from the prefrontal loop to the motor loop as the latter learning catches the former supervision.

Fig. 3 illustrates a similar but different version of the neural circuit model of heterarchical reinforcement learning. Here, the afferent part of the striatal spiral connections with SNc/VTA (Haber, 2003) is taken into account as well as the efferent part. Another point we assume in this version is that the spread projection from the caudate to the ventrolateral part of SNc/VTA is inhibitory and provides the negative of the caudate value approximation and not the positive time-advanced version. Note here that in both models, several weighting factors such as 1/2, γ, and ε are arbitrarily selected to keep equations simple and comparable with the standard TD model:

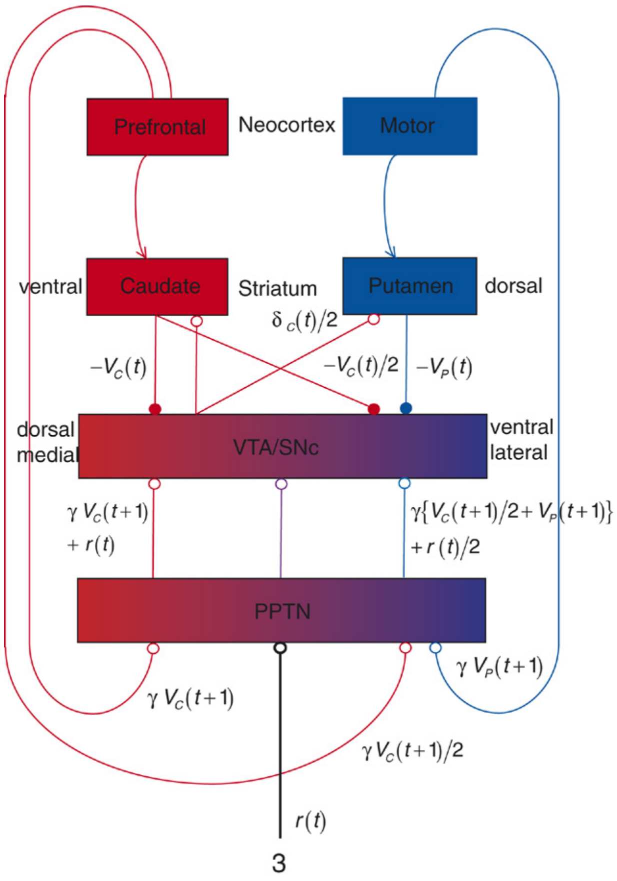

Fig. 3. A different neural implementation of the heterarchical reinforcement-learning model. The key differences from Fig. 2 are the inhibitory projection of a value function (not time-advanced) from the caudate nucleus to VTA/SNc and a spreading afferent dopamine projection ($\delta_C(t)$) from VTA/SNc to the caudate nucleus and putamen.

$$
\begin{aligned}
\delta_P(t) &= \mathrm{PPTN}_P(t) - \mathrm{Caudate}(t)/2 - \mathrm{Putamen}(t) + \delta_C(t)/2 \\
&= \{\gamma V_P(t+1) + \gamma V_C(t+1)/2 + r(t)/2\} \\
&\quad - V_C(t)/2 - V_P(t) \\
&\quad + \{\gamma V_C(t+1) + r(t) - V_C(t)\}/2 \\
&= \gamma\{V_C(t+1) + V_P(t+1)\} \\
&\quad + r(t) - \{V_C(t) + V_P(t)\}. \qquad (5)
\end{aligned}
$$

As before, we assume that prefrontal loop describes the environment with more coarse-grained representations. Then the caudate value-function approximation is expected to be of much lower frequency than putamen in both space and time. At the very early stage of learning when both $V_p(t+1)$ and $V_P(t)$ are close to zero, Eq. (5) is almost equal to standard TD learning. In the early stage of learning when $V_P(t)$ is very small but $V_C(t)$ is significantly different from zero, the $\gamma V_C(t+1) - V_C(t)$ term gives a virtual and additional reward or subgoal, which is an addition to the basic reward $r(t)$. Even when the reward is given at a long future time, and $V_P(t)$ is flat and zero, this extra term can give quite frequent reward clues for the correct direction of changes. In another interpretation of (5) at the early stage of learning, $\delta_P(t)$ is approximated by $\delta_C(t)$ because $V_P(t)$ is very small but $V_C(t)$ is significantly different from zero. Thus, $V_p(t)$ follows $V_C(t)$ in an almost supervised manner. At the late stage of learning, when $\delta_C(t)$ almost vanishes at the coarse spatial and temporal resolution, $\gamma V_P(t+1)+r(t)-V_P(t)$ is well approximated by the fine-grained high-frequency component of the primary reward because the coarse-grained low-frequency component of the reward is cancelled by $\gamma V_C(t+1) - V_C(t)$. Thus, a final brush-up of $V_P(t)$ at the fine-grained resolution might happen at the final stage so that $V_P(t)$ can be a good approximation to the real value function. We again emphasize the coarse-to-fine but parallel nature of the heterarchical reinforcement learning model.

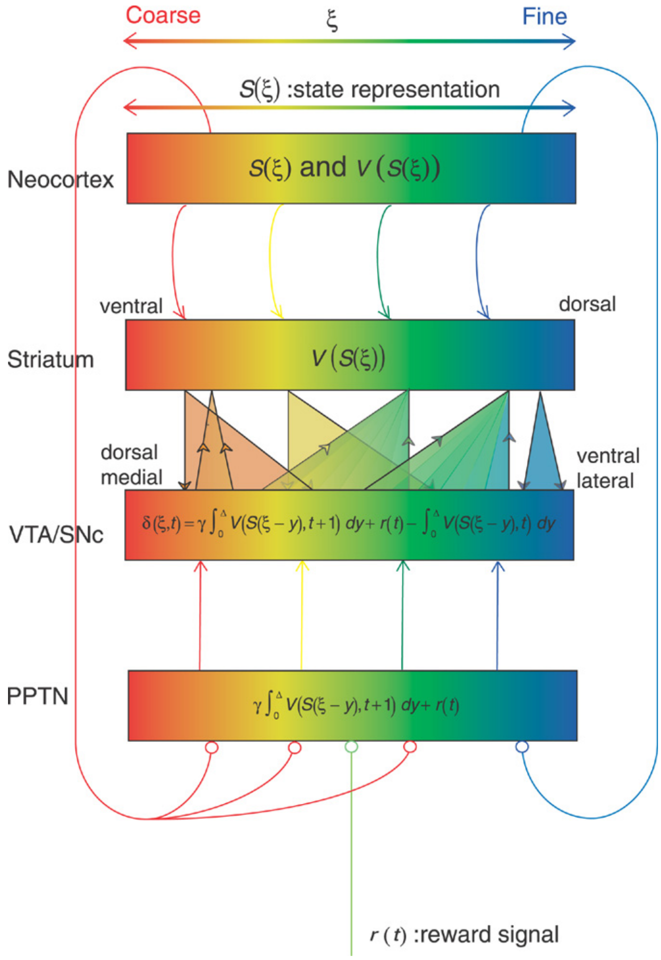
Coarse
ξ
Fine
$S(\xi)$:state representation
Neocortex
$S(\xi)$ and $V(S(\xi))$
ventral
dorsal
Striatum
$V(S(\xi))$
dorsal
medial
ventral
lateral
VTA/SNc
$\delta(\xi,t)=\gamma\int_0^\Delta V(S(\xi-y),t+1)\,dy+r(t)-\int_0^\Delta V(S(\xi-y),t)\,dy$
PPTN
$\gamma\int_0^\Delta V(S(\xi-y),t+1)\,dy+r(t)$
$r(t)$:reward signal

Fig. 4. Generalized scheme of the heterarchical reinforcement-learning model consisting of multiple brain structures. ξ specifies a granularity of state representation that ranges from 0 to 1, and $V(S(\xi))$ represents a value function for different state space and time scale.

Finally, we explain a more realistic multiple cortico-striatal loop model in Fig. 4 while emphasizing coarse-to-fine representations from the DLPFC to the primary motor cortex, as shown in Fig. 1. Just as the simplified models of Figs. 2 and 3, the loop communication between the neocortex, the striatum, the VTA/SNc, and the PPTN is almost private except the spiral connections between the striatum and the VTA/SNc and the cortex to PPTN. We assume that these connections are spread in a direction from coarse-to-fine. Emphasizing the topographic and almost continuous distributions of differentially-grained representations within the neocortical areas and the striatum, as depicted in Fig. 1, we denote the degree of resolution of the representation at each point in the loop by ξ. ξ is 0 for the coarsest representation and 1 for the finest representation. $S(\xi)$ is the state representation at resolution ξ. Independent representations of reward prediction $V(S(\xi), t)$ and reward prediction error $\delta(\xi, t)$ are computed at each resolution ξ by the corresponding parts of the striatum and VTA/SNc, respectively, while influenced by coarser representations. A straightforward extension of Eq. (5) to this continuous case is given as follows:

$$\delta(\xi, t) = \gamma \int_0^{\Delta} V(S(\xi - y), t + 1)\mathrm{d}y + r(t) - \int_0^{\Delta} V(S(\xi - y), t)\mathrm{d}y. \qquad (6)$$

Δ defines the extent of granularity which affects $\delta(\xi, t)$.

3. Experimental methods

To examine the model's predictions, we conducted a model-based fMRI experiment of a stimulus-action-reward association task in which subjects were asked to learn an advantageous button-push (left or right) in response to visual stimuli.

It is ideal for the multiple reward predictions and reward prediction errors, proposed in the heterarchical reinforcement learning model, to be estimated for each subject during learning; unfortunately that is practically intractable. Therefore, we replaced them with the stimulus-action-dependent reward prediction and reward prediction error (hereafter referred to as SADRP and RPE, respectively) in each trial for each subject by using the Q-learning model (Sutton & Barto, 1998) (c.f., temporal difference learning model section) to evaluate the models' predictions as quantitatively as possible. Because each trial is independent in our experimental design, we regarded γ as 0 and focused on within-trial predictions of reward.

3.1. Experimental design

Fig. 5A delineates an experimental trial under TEST (left) and CONTROL (right) conditions, which is the identical experiment reported in (Haruno & Kawato, 2006). In a TEST trial, subjects learned the stochastic association between a visual stimulus, a button-push, and rewards to maximize their total monetary rewards. After one of three fractal stimuli was presented (onset at 0.7 s), subjects either pushed the left or right button triggered by a beep (at 5.2 or 6.2 s). Either to the left or right of the fixation cross, a small green circle appeared to show which button had been selected. All subjects pushed the buttons with their index or middle finger of their right hand. If the trial was successful, the figure frame turned yellow (at 10.2 or 12.2 s), and the subject obtained a 50-yen reward. Otherwise, the frame turned purple, and the subject suffered a 50-yen penalty (not shown in Fig. 5A). The next trial began nine sec after reward feedback.

The actual outcome of each button-push—success or failure—was stochastically determined depending on the fractal stimulus presented and the subject's button-push. As an example of how this stochasticity works, Fig. 5 illustrates experimental session 1 (S1 out of three sessions, S1–S3), which was controlled with a probability of 0.9 (90%). For the yellow fractal figure (FS1), a left button-push yielded +50 yen with a probability of 0.9 and −50 yen with a probability of 0.1. In contrast, a right button-push produced +50 yen with a probability of 0.1 and −50 yen with a probability of 0.9. Therefore, the optimal behavior for FS1 was to push the left button, which the subjects had to learn by trial and error. In S1, the dominant probabilities of the other two fractal figures (FS2 and FS3) were also 0.9, and the advantageous button-push was randomized for left or right (optimal behaviors were FS1: left, FS2: right, and FS3: left). Note that subjects could not develop a stimulus-action-reward association before presentation of the fractal stimulus. Importantly, the subjects were instructed to decide which button to push as soon as the fractal stimulus was presented. Occurrence of the three fractal figures was controlled equally and pseudo-randomly by setting the same random number sequence for all subjects to reduce the variance of learning speed across subjects. Each trial lasted 19 or 21 s, and one TEST block included four repetitions of a trial (Fig. 5B). Accumulated reward was displayed above the figure frame and updated at the moment of reward delivery.

To control learning difficulty, we conducted three experimental sessions, S1, S2, and S3, in which the dominant probability was 0.9, 0.8, and 0.7, respectively. According to stochastic uncertainty, learning was expected to become progressively more difficult. The order of these sessions was counterbalanced across subjects, and the results were analyzed together because no marked differences in learning performance or imaging results were found between the different orderings of tasks. At the start of the experiments, the subjects were told that success or failure depended stochastically on the fractal stimulus presented and the button pushed, but they were not provided with any concrete information on the stochastic parameters. The subjects were encouraged to earn as large a monetary reward as possible, which was actually given to them in addition to their basic compensation (1500 yen).

In a CONTROL trial, subjects were asked to push the same button as in the preceding TEST block. They were signaled which button to push by a small green circle that appeared to the left or right of the fixation point just after fractal stimulus presentation; this reproduced their own button-push in the preceding TEST block in a randomized order. The fractal stimulus and outcome color (yellow or purple) had no influence on subject button selection but simply reproduced the effects of the visual displays in the TEST trials. Thus, aside from the timing of the green circle's presentation, the CONTROL block reproduced all of the physical events of the preceding TEST block and was used to subtract these effects from the TEST trials. No reward or penalty was given in the CONTROL trial. The accumulated reward above the figure box in the CONTROL block remained constant at the value of the preceding TEST trial. As in the TEST block, one trial lasted 19 or 21 s, four repetitions per block, and the TEST and CONTROL blocks were alternated (Fig. 5B). One session included 12 TEST/CONTROL blocks and lasted 32 min (20 s [on average] × 4 trials × 2 [TEST + CONTROL] × 12 blocks.). We prepared five different sets of three fractal stimuli

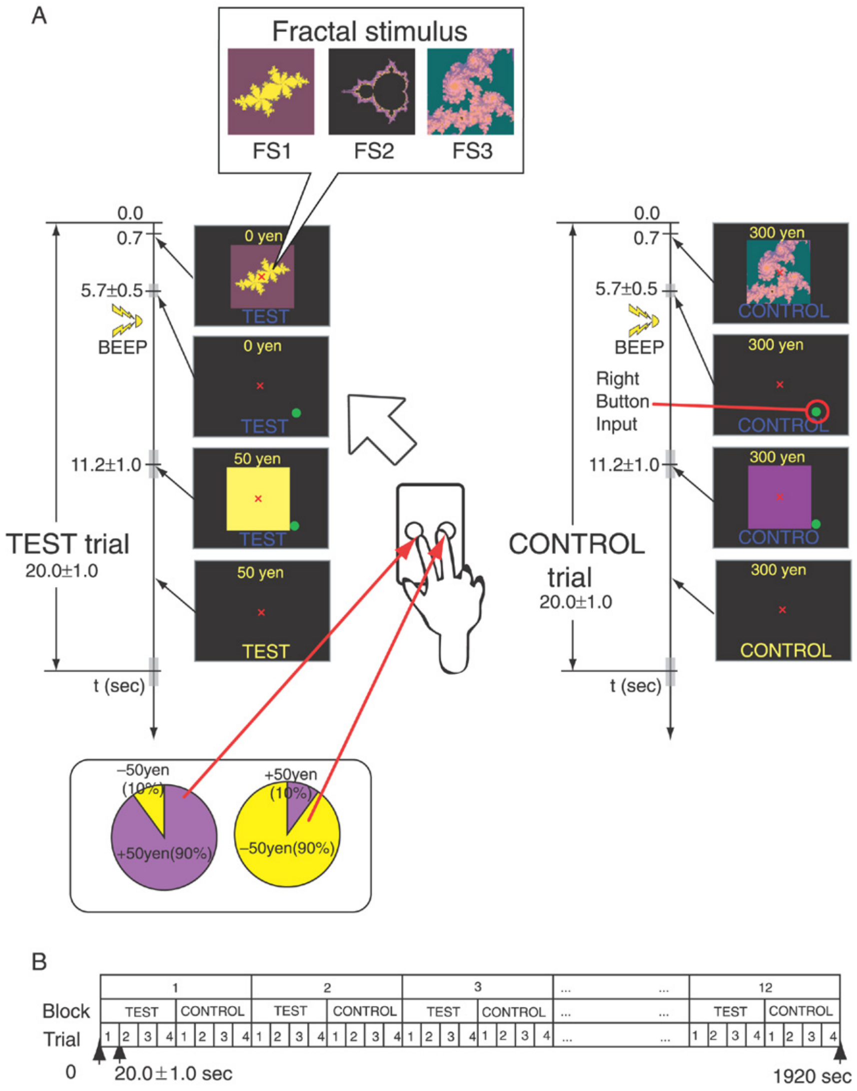

Fig. 5. Experimental design. **A** and **B** illustrate the TEST and CONTROL trials and the overall organization of the experiment. **A left:** In each TEST trial, one of three fractal stimuli (FS) was presented, and the subject was asked to press the left or right button following a beep to obtain a monetary reward. A small green circle appeared showing which button the subject had pushed. In this example (Session 1) the optimal (advantageous) button-push for each FS was set (FS1: left, FS2: right, FS3: left) to yield a reward of 50 yen (yellow frame presented) or a penalty of −50 yen (purple, not shown) with a probability of 0.9 and 0.1, respectively. By contrast, a non-optimal (disadvantageous) button-push (FS1: right, FS2: left, FS3: right) led to a 50-yen reward or penalty with a probability of 0.1 and 0.9, respectively. **right:** In CONTROL, subjects had to reproduce button-pushes in the preceding TEST block for the same set of fractal stimuli, instructed visually by the small green circle's position. The order of the stimulus and button-push was randomized. The fractal stimulus and outcome color (yellow or purple) simply reproduced TEST and was unrelated to the subjects' selection of button-push. The accumulated reward above the figure frame remained constant at the value of the preceding TEST trial (no reward or penalty in CONTROL). **B** Four trials were included in each of the TEST and CONTROL blocks, and they were interleaved twelve times.

and changed the configuration of the stimulus set every session to exclude any brain activity arising from a fixed set of figures.

3.2. Subjects

Twenty healthy adults (23–31 years old, 11 males and 9 females, all right-handed) participated in the experiment. Informed consent of the participants was obtained before the experiment, and the protocol was approved by ATR's ethics committee.

3.3. MRI acquisition and preprocessing

MRI scanning was conducted with a 1.5 T Marconi scanner. For each subject, 768 scans of BOLD images (TR 2.5 s, TE 49 ms, flip angle 80°, FOV 192 mm, resolution 3 × 3 × 5 mm) were acquired over two sessions. In addition to these experimental trials, each session contained two preliminary dummy CONTROL trials (16 scans) to allow for T1 equilibration effects. Then we stopped the MRI scanner and gave the subjects a ten-minute break outside the scanner. After the break, the same procedure was repeated for another (third) session. High-resolution (T1 (1 × 1 × 1 mm) and T2 (0.75 × 0.75 × 5 mm)) structure images were also acquired for each subject. The data were analyzed using standard procedures implemented in Statistical Parametric Mapping (SPM99) (Friston et al., 1995). Prior to the statistical analysis, we conducted motion correction and nonlinear transformation into the standard space of the MNI coordinates, as implemented in SPM99. These normalized EPI images were resliced into $2 \times 2 \times 2$ mm voxels and then smoothed with an 8 mm full width half maximum isotropic Gaussian kernel.

3.4. SADRP and RPE estimations

The Q-learning model was introduced to estimate the subject's SADRP and RPE during learning. More precisely, a subject's SADRP at time t can be represented as a table for $Q(fs, bp, t)$ indicating the predicted amount of reward for button-push bp (right or left) and fractal stimulus fs. Note that the optimal selection of behaviors is trivial once the true SADRP table is acquired; at that point, the button with the larger Q is selected. When the subject receives actual reward $r(t)$, RPE amounts to $r_{(t)} - Q(fs, bp, t)$. Then the model changes the element of the table by the following rule to decrease RPE for the next occurrence of the same combination of stimulus and action:

$$Q(fs, bp, t+1) = Q(fs, bp, t) + \alpha_t^{fs}(r(t) - Q(fs, bp, t)).$$

This procedure only updates the table element corresponding to the subject's selected action bp and the given fractal stimulus fs in proportion to reward prediction error (Sutton & Barto, 1998). It is used here to estimate subjects' SADRP and RPE. Therefore, only the component of SADRP that corresponds to the given stimulus and the selected action in each trial will be shown, updated, and used in the subsequent analysis. In the early stage of learning, when SADRP is inaccurate and RPE has a large value, the change in SADRP is expected to be large, whereas in the late stage of learning when SADRP is accurate and RPE is small, the change in SADRP is expected to be small. Thus, SADRP tends to converge to an asymptotic value.

Learning rate α_t^{fs} controls the amplitude of change and is determined by a standard recursive least-square procedure (Bertsekas & Tsitsiklis, 1996; Young, 1984). In the current situation, α_t^{fs} is reduced to an estimation of the inversed variance for fractal stimulus fs that has a value of 1 when presented and 0 otherwise; then we derive the following update rule:

$$\alpha_t^{fs} = \frac{\alpha_{t-1}^{fs}}{1 + \alpha_{t-1}^{fs}}.$$

Qualitatively, learning rate α_t^{fs} decreases as SADRP becomes reliable. This property of α_t^{fs} is important because SADRP does not necessarily change much after the completion of learning, even if RPE occurs due to the stochastic nature of the task. The update equation indicates that the learning rate sharply decreases below 1, suggesting that the initial value of α_t^{fs} (i.e., α_0^{fs}) has little effect on the estimation of SADRP and RPE. We set a value of 1000 throughout the study.

3.5. Correlation analysis of fMRI data

After preprocessing, we conducted an event-related correlation analysis of fMRI data with SADRP and RPE. We assumed that brain activities related to SADRP and RPE occur at the timing of stimulus presentation and reward delivery, respectively (Haruno & Kawato, 2006). During the CONTROL trials SADRP and RPE were assumed to be 0 for the following reasons. First, there was no monetary reward. Second, the combination of fractal stimuli and button-pushes (left or right) was arbitrary during control trials. Therefore, it was neither necessary nor possible for subjects to predict the number of rewards during CONTROL. Third, the subjects were instructed to passively push the button.

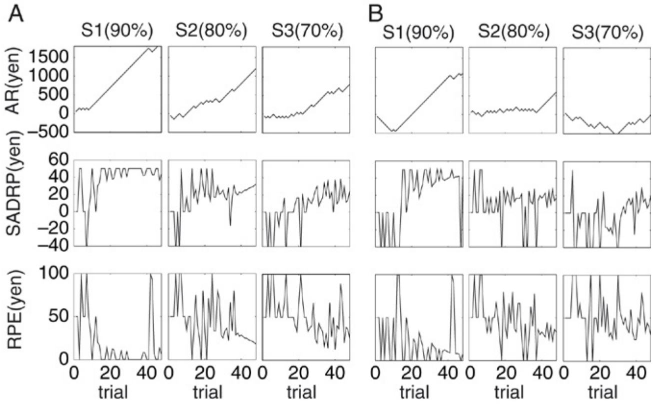

Fig. 6. Behavioral results of learning for the most and least successful subjects in terms of total reward. **A** and **B** show the time courses of the AR, SADRP, and RPE for the most and least successful subjects, respectively. S1, S2, and S3 represent experimental sessions with a dominant probability of 0.9, 0.8, and 0.7, respectively.

4. Results

4.1. Subjects' behaviors and estimation of subjects' SADRP and RPE

Fig. 6 shows how the reward acquisition and button-push behaviors changed during the TEST blocks of the stimulus-action-reward association task for the most successful subject (A) and least successful subject (B) in terms of total monetary reward. Accumulated reward (AR) increases almost monotonically in S1–S3 in A. In contrast, only S1 exhibits a monotonic increase in B, and the flat and decreasing tendencies found in S2 and S3 show that learning was demanding for the subject and that it had not yet been completed within the given number of trials. The averages of all subjects displayed in Fig. 7 show that ARs yielded progressively smaller positive slopes in S1, S2, and S3. Accumulated rewards in the final TEST blocks were significantly larger than zero ($P < 0.0001$; t-test) and ranked in the following order: S1/0.9 > S2/0.8 > S3/0.7 ($P < 0.05$; t-test). These observations are consistent with the hypothesis that learning is progressively more difficult in S1, S2, and S3 in accordance with their stochastic uncertainties.

From their behavior, we used the Q-learning model (Sutton & Barto, 1998) to estimate each subject's stimulus-action-dependent reward prediction (SADRP), which is defined as the amount of reward predicted by a subject based on a given contextual stimulus and an action selected by the subject. RPE simply amounts to the difference between SADRP and actual reward. SADRP is shown in the second rows of Figs. 6 and 7. The horizontal lines in Fig. 7 show theoretical maximum

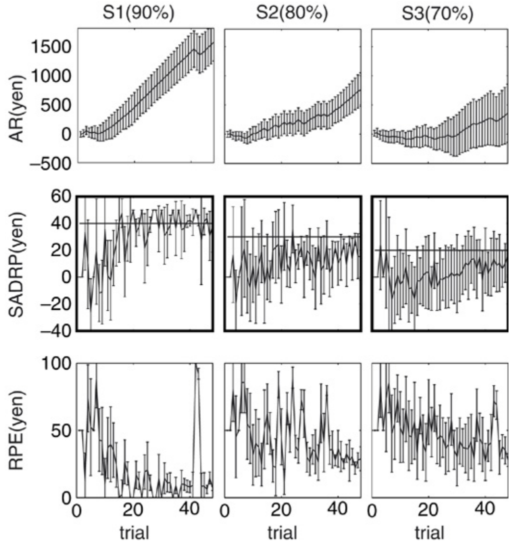

Fig. 7. Behavioral results of learning for the average and standard deviations of all twenty subjects. Similar to Fig. 6, the time courses of AR, SADRP, and RPE averaged over twenty subjects are shown.

values that are expected for optimal button-push (40 yen [$= 50 * (0.9 - 0.1)$], 30 yen [$= 50 * (0.8 - 0.2)$], and 20 yen [$= 50 * (0.7 - 0.3)$] for S1–S3, respectively). In the easiest task (S1), SADRP increased and approached the theoretical maximum (40 yen) within 20 trials for all subjects. In more stochastic tasks (S2 and S3), the increase in SADRP became progressively slower than in S1, and some subjects failed to achieve maximum SADRP even in the final TEST trial. None of the estimated SADRPs of any of the subjects showed a simple monotonically increasing tendency due to the stochasticity of the task. Corresponding to SADRP, the absolute values for RPE shown in the third rows of Figs. 6 and 7 rapidly decreased close to 5 yen within 20 trials in S1, but decreased only slowly in S2 and S3. Again, because of the task's stochasticity, RPEs did not exhibit a monotonically decreasing tendency in time.

4.2. *Correlation analysis of fMRI data*

We carried out an event-related regression analysis of the fMRI data with SADRP and RPE. Here, we focus on an analysis of brain activity in the striatum, the medial and orbital prefrontal, the anterior cingulate, the cingulate motor, and the dorsal premotor cortices because our aim is to examine the predictions of the heterarchical reinforcement learning model. All analyses were conducted with the random-effect model implemented in SPM99 (Friston et al., 1995), and the statistical threshold was set at $P < 0.001$, uncorrected for multiple comparisons, with an additional constraint that at least five contiguous voxels be included. Correlation analyses for the two variables in different sessions (S1–S3) were conducted separately because the scanner was stopped and the subjects took a ten-minute break between their second and third sessions. All illustrations of statistical maps (i.e., Figs. 8–10, average of all subjects) were made using our in-house software 'multi_color,' which is freely available to the research community at http://www.cns.atr.jp/multi_color/.

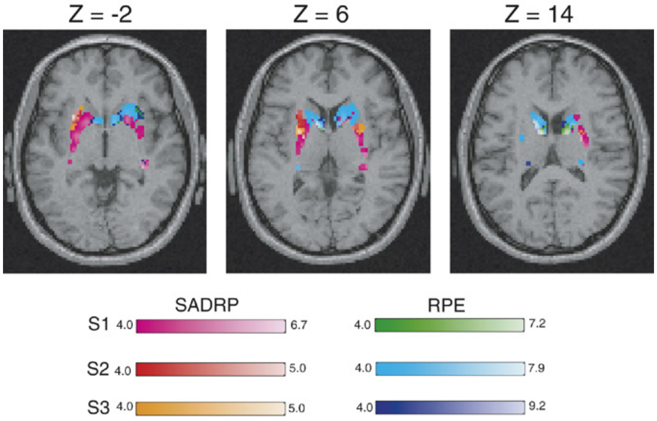

Fig. 8. Activity in striatum correlated with SADRP and RPE for S1–3. Each voxel (2 × 2 × 2 mm) is associated with T-values for SADRP and RPE, represented as the brightness of colors, as shown in color bars. The overlapping voxel activated in the two analyses is represented by a mosaic comprising two corresponding colors. The range of Z in MNI coordinates was −2 to 14, which includes the putamen and caudate nucleus as well as part of the ventral striatum (ventral putamen) (Talairach & Tournoux, 1998).

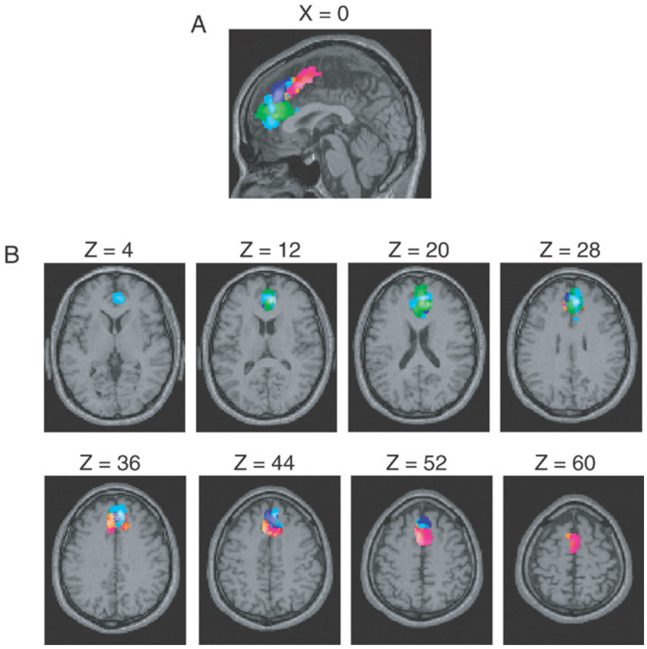

Fig. 9. Activity in medial regions of frontal cortex correlated with SADRP and RPE for S1–3. **A** and **B** show sagital and horizontal views, respectively, in the same format as Fig. 8. The range of Z in MNI coordinates was 4 to 60, which includes the orbitofrontal and medial prefrontal, the anterior cingulate, and the cingulate motor cortices (Talairach & Tournoux, 1998).

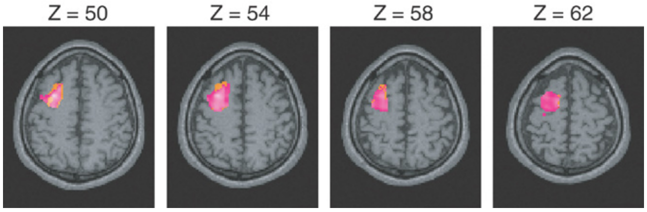

Fig. 10. Activity in left premotor cortex correlated with SADRP and RPE (no voxel) for S1–3, formatted in the same way as Fig. 8. The range of Z in MNI coordinates was 50 to 62.

Fig. 8 illustrates the significant correlation in the striatum (consisting of the putamen and caudate nucleus) with SADRP and RPE for three tasks (S1–S3). Here, the color map associated with each voxel represents its T-values of SPM99 for SADRP and RPE. The most remarkable observation is that SADRP activity for S1–S3 was mainly confined within the putamen, whereas RPE activity was mainly localized within the caudate nucleus and the ventral striatum. These separate distributions of SADRP and RPE activities remained robustly consistent regardless of the differences in task difficulty from S1 to S3. Second, the number of voxels correlated with SADRP and RPE strongly depended on task difficulty in exactly the opposite manner: SADRP activity tended to be more prominent in the less stochastic task (S1) than in the more stochastic tasks (S2 and S3), whereas RPE activity both in the caudate nucleus and ventral striatum tended to exhibit stronger correlations in the more stochastic tasks (S2 and S3). More specifically, the number of voxels that correlated with SADRP in S1, S2, and S3 was 683, 87, and 101, respectively, and the number that correlated with RPE was 399, 864, and 565. Only SADRP activity for S1 significantly overlapped RPE activity (SADRP had only five overlapping voxels with RPE for both S2 and S3). The number of overlapping voxels of SADRP for S1 with RPE for S1, S2, and S3 was 40, 180, and 107, respectively. It is also intriguing to focus on spatial distribution within these SADRP and RPE correlated activities that depended on task difficulty. SADRP activity for S1 was seen in the whole putamen, while activity for S2 and S3 was located in the more anterior-lateral part ($Z = -2$ and 6). In contrast, RPE activity for S1 in the ventral part ($Z = -2$) was almost confined to the putamen, while activity for S2 and S3 was located more medially in the caudate nucleus. Because in S2 and S3, learning is at a comparatively earlier stage than in S1, finding more prediction error in the putamen during S1 and more prediction error in the caudate during S2 and S3 fits well with the model's prediction that reward prediction error gradually shifts from the dorsomedial to the ventrolateral substantia nigra.

Fig. 9 shows the activity in the medial frontal region of the brain that correlated with SADRP and RPE. The SADRP correlation was located in the dorso-caudal regions, mainly in the rostral cingulate motor area, SMA and pre SMA. Activity in simple tasks (S1 and S2) tended to be located dorsally within these areas ($Z = 36$, 44, and 53). In contrast, RPE correlation was located more ronstro-ventrally, mainly in the rostral part of anterior cingulate cortex. The activity in simple tasks (S1 and S2) tended to be located more ventrally within these areas than in difficult tasks (S3) ($Z = 4$, 12, 20, and 28).

Fig. 10 shows the activity in the dorsal premotor cortex that correlated with SADRP. There was no correlated activity found in this area with RPE. In sharp contrast with the striatum and medial frontal regions, task difficulty (S1–S3) did not cause any spatial difference of activity within the region.

In summary, fMRI results are in agreement with the model's predictions, i.e., the activity of the anterior cingulate cortex and the caudate nucleus correlated with reward prediction error, and the rostral premotor, cingulate motor cortices, SMA and pre SMA along with the putamen showed a correlation of activity with stimulus-action-dependent reward prediction. In addition, in the putamen, correlation with SADRP is located posteriorly in easy tasks but anteriorly in difficult tasks. On the other hand, in the caudate nucleus, correlation with reward prediction error is located posterior-laterally in easy tasks but anterior-medially in difficult tasks. The putamen also showed a correlation with RPE in easy tasks.

5. Discussion

We proposed a heterarchical reinforcement learning model that suggests how multimodal information in cortico-striatal loops is integrated. Asymmetric descending and ascending projections between SNc/VTA and the striatum play a key role in propagating reward prediction and its error signal from one cortico-striatal loop to the other through dopamine, enabling gradual refinement or tuning of reward prediction during learning. We conducted an fMRI experiment of stimulus-action-reward association learning to directly examine the model's predictions and obtained results consistent with the model. Specifically, brain activity in the anterior cingulate cortex and the caudate nucleus was correlated with reward prediction error during learning, while the dorsal premotor and rostral cingulate cortices, SMA and the putamen exhibited a strong correlation of activity with the stimulus-action-dependent reward prediction, which is gradually acquired using error signals. We also demonstrated the spatial non-uniformity of both SADRP and RPE activity within a brain structure (particularly, the putamen and caudate nucleus), which depends on the learning difficulty of a task. Due to the temporal and spatial limitations of fMRI resolution, only a few aspects of the model could be investigated. It is also important to compare neural activity in the SNc, the VTA, and the PPTN with the model's behaviors by conducting electrophysiological experiments.

Dynamic association of various contextual cues with action and reward is critical to make effective decisions (Barraclough, Conroy, & Lee, 2004). A crucial question here is how to combine several reward predictions, each of which is based on different information. For example, some reward prediction may only depend on visual cues, but others may utilize not only visual and auditory cues but also the action taken by a subject. Because the accuracy of different reward predictions varies dynamically during the course of learning, the combination of predictions is important (Daw, Niv, & Dayan, 2005). The proposed model takes a coarse-to-fine and continuous approach for blending, which could be naturally implemented by the asymmetric upstream and downstream projections between the SNc/VTA and the striatum. In the model, starting with initial rough reward prediction by guessing or inference, prediction is gradually refined taking motor information into account. In good agreement with the model, our experimental results highlighted the medial prefrontal and the orbitofrontal and anterior cingulate cortices when reward prediction error is huge in the early stage of learning. These areas have been implicated in reward-related cognitive functions (Picard & Strick, 1996; Price et al., 1996), and therefore might be related to reward prediction, which is based solely on visual cues and is called a state value function V in computational literatures (Sutton & Barto, 1998). In contrast, the anterior dorsal premotor and cingulate motor cortices were lit up when stimulus-action-dependent reward prediction is dominant rather than reward prediction error in the late stage of learning. These brain areas have been reported to be involved in motor-related cognitive functions (Picard & Strick, 2001). It is therefore reasonable that these regions encode stimulus-action-dependent reward prediction or state-action value function Q in computational literatures (Sutton & Barto, 1998). Our model and experimental data are also consistent with a pioneering work from Hikosaka's group that focused on the functional roles of parallel cortico-striatal loops during sequential motor learning of monkeys (Hikosaka et al., 1999; Hikosaka, Nakamura, Sakai, & Nakahara, 2002; Miyachi, Hikosaka, Miyashita, Karadi, & Rand, 1997; Miyachi, Hikosaka, & Lu, 2002). They demonstrated that the caudate nucleus, the pre-SMA, and the dorsolateral prefrontal cortex were activated in the early stage of learning, while the putamen and SMA were important in the late stage of learning, suggesting that the observed shift of brain activity centers corresponded to the transformation from external to muscle coordinates. The heterarchical reinforcement learning model may explain how such a shift can be achieved.

It would also be interesting to determine whether the information flow in cortico-striatal loops is fixed in a top-down (coarse-to-fine) direction or regulated both top-down and bottom-up depending on the comparative accuracy of several reward predictions. Such learning and selection of multiple predictions is not a specific topic to the cortico-striatal loops, but rather generic to multiple closed loop structures found in subcortical-cortical structures. We previously proposed a model of cerebro-cerbellar loops known as the MOSAIC model, where multiple internal models are learned and basically selected in a bottom-up manner based on the accuracy of each prediction (Haruno, Wolpert, & Kawato, 2001). Pursuing a unified computational principle that exists behind closed loop circuits in the brain is an important future direction.

Acknowledgements

We would like to thank K. Nakamura and Y. Kobayashi for helpful discussions and comments. We would also like to thank S. Tada and T. Yoshioka for technical assistance. This study was supported by NICT.

References

Alexander, G. E., Crutcher, M. D., & Delong, M. R. (1990). Basal ganglia thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research*, *85*, 119–146.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*, 404–410.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems Man and Cybernetics*, *13*, 835–846.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, *19*, 10502–10511.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal system for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

Dayan, P., & Hinton, G. E. (1993). Feudal reinforcement learning. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems*: *Vol. 5* (pp. 271–278).

Friston, K. J., Holmes, A. P., Worsley, K., Poline, J. B., Frith, C., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional brain imaging: A general linear approach. *Human Brain Mapping*, *2*, 189–210.

Gerardin, E., Lehericy, S., Pochon, J. B., Tezenas du Montcel, S., Mangin, J. F., Poupon, F., et al. (2003). Foot, hand, face and eye representation in the human striatum. *Cerebral Cortex*, *13*, 162–169.

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, *20*, 2369–2382.

Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, *26*, 317–330.

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, *13*, 2201–2220.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., 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. *Journal of Neuroscience*, *24*, 1660–1665.

Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of Neurophysiology*, *92*, 948–959.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., et al. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*, 464–471.

Hikosaka, O., Nakamura, K., Sakai, K., & Nakahara, H. (2002). Central mechanisms of motor skill learning. *Current Opinion in Neurobiology*, *12*, 217–222.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *4*, 304–309.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). MIT Press.

Kobayashi, Y., Inoue, Y., Yamamoto, M., Isa, T., & Aizawa, H. (2002). Contribution of pedunculopontine tegmental nucleus neurons to performance of visually guided saccade tasks in monkeys. *Journal of Neurophysiology*, *88*, 715–731.

Kobayashi, Y., Okada, K., Inoue, Y., Yamamoto, M., & Isa, T. (2005). Reward predicting activity of pedunculopontine tegmental nucleus neurons during visually guided saccade tasks. In *Abstract of 35th annual meeting of society for neuroscience*, 890.5.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Brain Research Reviews*, *31*, 236–250.

Miyachi, S., Hikosaka, O., Miyashita, K., Karadi, Z., & Rand, M. K. (1997). Differential roles of monkey striatum in learning of sequential hand movement. *Experimental Brain Research*, *115*, 1–5.

Miyachi, S., Hikosaka, O., & Lu, X. (2002). Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Experimental Brain Research*, *146*, 122–126.

Montague, P. R., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morimoto, J., & Doya, K. (1999). Hierarchical reinforcement learning for motion learning: learning "stand-up" trajectories. *Advanced Robotics*, *13*, 267–268.

Oakman, S. A., Faris, P. L., Kerr, P. E., Cozzari, C., & Hartman, B. K. (1995). Distribution of pontomesencephalic cholinergic neurons projecting to substantia nigra differs significantly from those projecting to ventral tegmental area. *Journal of Neuroscience*, *15*, 5859–5869.

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

Parthsarathy, H. B., Schall, J. D., & Graybiel, A. M. (1992). Distributed but convergent ordering of corticostriatal projections: Analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *Journal of Neuroscience*, *12*, 4468–4488.

Picard, N., & Strick, P. L. (1996). Motor areas of the medial wall: a review of their location and functional activation. *Cerebral Cortex*, *6*, 342–353.

Picard, N., & Strick, P. L. (2001). Imaging the premotor areas. *Current Opinion in Neurobiology*, *11*, 663–672.

Price, J. L., Carmichael, S. T., & Drevets, W. C. (1996). Networks related to the orbital and medial prefrontal cortex; a substrate for emotional behavior?. *Progress in Brain Research*, *107*, 523–536.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*, 4595–4610.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500.

Selemon, L. D., & Goldman-Rakic, P. S. (1985). Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *Journal of Neuroscience*, *5*, 776–794.

Singh, S. P. (1992) Reinforcement learning with a hierarchy of an abstract models. In *Proceedings of the tenth national conference on artificial intelligence* (pp. 202–207).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. The MIT Press.

Sutton, R. S., Singh, S., Precup, D., & Ravindran, B. (1999). Improved switching among temporally abstract actions. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems*: *Vol. 11* (pp. 1066–1072).

Takada, M., Tokuno, H., Nambu, A., & Inase, M. (1998). Corticostriatal projections from the somatic motor areas of the frontal cortex in the macaque monkey: Segregation versus overlap of input zones from the primary motor cortex, the supplementary motor area, and the premotor cortex. *Experimental Brain Research*, *120*, 114–128.

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*, *92*, 2520–2529.

Talairach, J., & Tournoux, P. (1998). *Co-planar stereotaxic atlas of the human brain*. Thieme.

Young, P. (1984). *Recursive estimation and time series*. Springer-Verlag.