Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Neurocomputing 63 (2005) 253–269

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Model-based reinforcement learning: a computational model and an fMRI study

Wako Yoshida[a,b,*], Shin Ishii[a,b]

[a]Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[b]CREST, Japan Science and Technology Agency, Japan

Received 4 November 2003; received in revised form 27 February 2004; accepted 23 April 2004
Available online 20 August 2004

Abstract

In this paper, we discuss an optimal decision-making problem in an unknown environment on the bases of both machine learning and brain learning. We present a model-based reinforcement learning (RL) in which the environment is directly estimated. Our RL performs action selection according to the detection of environmental changes and the current value function. In a partially-observable situation, in which the environment includes unobservable state variables, our RL incorporates estimation of unobservable variables. We propose a possible functional model of our RL, focusing on the prefrontal cortex and the anterior cingulate cortex. To test the model, we conducted a human imaging study during a sequential learning task, and found significant activations in the dorsolateral prefrontal cortex and the anterior cingulate cortex during RL. From a comparison of the mean activations in the earlier and later learning phases, we suggest that the dorsolateral prefrontal cortex maintains and manipulates the environmental model, while the anterior cingulate cortex is related to the uncertainty of action selection. These experimental results are consistent with our model.

Keywords: Reinforcement learning; Partially-observable Markov decision process (POMDP); Prefrontal cortex; Functional MRI

*Corresponding author. Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan. Tel.: +81-743-72-5985; fax: +81-743-72-5989.

E-mail address: wako-y@is.naist.jp (W. Yoshida).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.04.012

1. Introduction

Although the environment around us is constantly changing, humans can learn the features of their current environment and determine optimal behaviors. Assuming optimality is defined as rewards received from the environment, an adaptation to the environment can be formulated as an optimal decision-making problem with an on-line identification of the current environment. We discuss here a possible reward-based decision-making method that is based on both machine learning and brain learning. To understand brain functions, it is important to integrate findings from various research fields. The aim of our study is to discuss brain functions based on a theoretical model and to evaluate it by means of neuropsychological experiments.

In the machine learning field, an optimal decision-making problem in a known or unknown environment is often formulated as a Markov decision process (MDP). If an MDP includes the direct identification of an unknown environment, the problem can be solved by a model-based reinforcement learning (RL) method [7–9,15,26]. In RL, the objective of an agent is to maximize rewards accumulated for the future, which is achieved by improving the agent's action selection. To improve the strategy for receiving rewards (i.e., policy), a standard RL scheme then estimates expected reward accumulation with respect to current policy, which is the value function. A model-based RL [7–9,15,26] tries to identify current environment directly, and the value function is approximated from the resulting environmental model. However, human environments often include unobservable state variables. If we consider decision-making by a card player, for example, cards held by the other players are not directly observed and hence are unobservable variables. In a previous paper [11], we presented a model-based RL method for a partially-observable Markov decision process (POMDP) to deal with such a realistic problem. The POMDP assumes that the environment contains unobservable variables. In this paper, we briefly re-introduce our model-based RL method, and then propose a functional model for the brain, in which our RL method is realized within a POMDP environment.

Because RL has provided a good model of animal behaviors (conditioning), with recent neurophysiological and neuroimaging studies suggesting that RL algorithms are associated with neural processing in the brain [1,23,30], the components of RL could correspond to brain functions. We assume that the major parts of our RL method (the environmental model, the value function and the estimation of unobservable state variables) are involved in the prefrontal cortex (PFC): the dorsolateral prefrontal cortex (DLPF), the orbitofrontal cortex (OFC) and the anterior prefrontal cortex (APF). Our RL method also requires that action selection is dependent on estimation of the current environmental model and the value function. We consider that this operation occurs in the anterior cingulate cortex (ACC).

To test our functional model, a human imaging study using functional magnetic resonance imaging (fMRI) was conducted in this study. Although POMDP is discussed in the computational section, the experimental study does not assume a partially-observable situation. Our RL method within a POMDP, an extended version of MDP, can be formulated by adding a hidden state estimation process to the scheme in MDP situations. In the experiment, we aimed to clarify fundamental components of our functional brain model, that is, the brain regions realizing MDP. In our MDP task, therefore, the subject is required to determine an optimal control sequence while simultaneously identifying a Markov environment. Our imaging study suggests that the posterior DLPF maintains, and the anterior DLPF manipulates, possible environmental models, and that the ACC represents the uncertainty of the action selection. These results are consistent with an essential part of our model.

2. Model-based RL method

Assuming that an agent is provided by the environment with a scalar reward corresponding to an action for each state, and that the optimality of the agent's action is determined by the amount of rewards, the optimal decision problem can be solved by RL schemes. Since the objective of an RL agent is to maximize rewards accumulated for the future, a standard RL scheme estimates the reward accumulation which is called the value function. To make an appropriate decision in an unknown environment, it is important to understand the dynamics of the environment, i.e., how the current state changes by the agent's own action. Model-based RL methods [7–9,15,26], variations of RL, try to model the environmental dynamics, and the value function is approximated using the model. Model-based RL has an advantage, especially for dealing with partially-observable environments and/or dynamic environments, because the environmental model can explicitly deal with such complexity. A model-based RL also learns faster than a model-free alternative which requires no model of the environmental dynamics.

Since the RL is often formulated as an MDP, the notion of MDP is briefly introduced here. For further explanation, see [27]. We assume discrete Markov environments; $P(s'|s,a)$ gives the probability of reaching state s' by selecting action a at state s. If the state-transition probability is known, the value function for state s, $V(s)$ should satisfy the following optimal Bellman's equation:

$$V(s) = \max_a Q(s,a), \tag{1a}$$

$$Q(s,a) \equiv r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s'), \tag{1b}$$

where $r(s,a)$ denotes an immediate reward for the state-action pair (s,a) and $0 \leqslant \gamma \leqslant 1$ is a discount constant. The action-value function $Q(s,a)$ represents the expected reward accumulation when the agent takes action a at state s and optimal actions at subsequent states. The objective of RL is to obtain the optimal policy, which outputs the action maximizing $Q(s,a)$ for any state s. In many RL problems, the state-transition probability $P(s'|s,a)$ is unknown. The model-based RL tries to model the environment directly by approximating the state-transition probability based on past state transitions. The value function and the environmental model are therefore learned concurrently but independently.

2.1. Partially-observable MDP

In many real-world problems, learning agents can observe only a part of dynamical variables. In the field of machine learning, such a real world is modeled as a Markov environment with unobservable (hidden) state variables, and a decision making problem in the environment is formulated as a POMDP [12]. Although the experiment described in Section 3 is *not* a POMDP, we introduce the POMDP formulation here because our RL model primarily assumes a real-world situation.

Let $s \equiv (y, z)$ be an environmental state, where y and z denote observable and unobservable state variables, respectively. Although the underlying dynamics of the POMDP are still Markov, the environment with respect to the observable variables does not have a Markov property [10], because a learning agent has no direct access to the true state s due to the existence of unobservable variables z. One way to deal with a POMDP is called a belief state MDP, in which the Bellman's equation is modified from that in MDP, (1), by replacing state s with belief state b. A belief state is typically a probability distribution of the observable and unobservable variables. Since there is no probabilistic factor for the observable variables, $b = [y, \hat{P}(z)]$, where $\hat{P}(z)$ is estimated from past observations. We assume that an agent can estimate a new belief state $b' = [y', \hat{P}'(z)]$, using the new observation y'. Even in a finite world, where both state and action spaces are discrete and finite, the belief state MDP is hard to solve, because the belief state value function is defined for the probability distribution of the unobservable variables and is often intractable. Therefore, we need an approximation.

If an RL agent is almost certain of the estimation of the unobservable variables, $[y, \hat{P}(z)]$ is well represented by $u \equiv [y, \hat{z}]$, where $\hat{z}$ denotes the most likely value of z. Using the estimated environmental state u, i.e., a pair of the observable variable y and the estimated unobservable variable $\hat{z}$, the Bellman's equation is approximated as

$$V(u) = \max_a Q(u, a), \tag{2a}$$

$$Q(u, a) = r(u, a) + \gamma \sum_{u'} P(u' | [y, \hat{P}(z)], a) V(u'). \tag{2b}$$

The new Bellman's equation is different from the original one (1), in which $P(z) \approx \delta(z - \hat{z})$[1] is naively applied, because the state transition incorporates the estimated distribution of the unobservable variables $\hat{P}(z)$. Approximation (2) may not be valid, especially when the RL agent is uncertain of the estimation of the unobservable variables $\hat{z}$. In other words, when the uncertainty of the unobservable variables is high, the "best" policy based on the approximated Bellman's equation (2) may not actually be optimal. Our previous RL method [11] therefore introduced an additional reward term, called an exploration bonus, which induces exploratory behaviors for acquiring information from the environment [10].

[1] $\delta(\cdot)$ is the Dirac's delta function. $\delta(t) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon}\{\mathbf{1}(t + \varepsilon) - \mathbf{1}(t - \varepsilon)\}$, where $\mathbf{1}(t)$ is the Heaviside function.

In our RL method, the distribution $\hat{P}(z)$ is estimated by a Bayes inference with a "forgetting" effect on past experiences and a non-informative prior. For detailed formulations, see [11].

2.2. Action selection

Here, we discuss how to select an action depending on the current value function. We define a stochastic policy π by the conditional probability $P^\pi(a|u)$. If the agent knows the correct optimal value function, including the correct estimation of the environmental dynamics, the optimal policy is the one that simply selects a "greedy" action to maximize the value function $\int Q(u,a)P^\pi(a|u)\,\mathrm{d}a$ at each state. In a world where state and action spaces are finite, the greedy policy will assign probability zero to possible actions except one or several actions. Since the agent does not know the correct optimal value function during the process of trial and error, the greedy action is not necessarily optimal. In addition, when the environment changes, the value function approximated by using past experiences may not be optimal. To establish the optimal value function, the agent should execute trial actions, i.e., actions that are not optimal with respect to the current value function. To preserve such adaptability of the policy, free energy is introduced

$$F(P^\pi) = \int Q(u,a)P^\pi(a|u)\,\mathrm{d}a - \frac{1}{\beta}\int P^\pi(a|u)\,\log\,P^\pi(a|u)\,\mathrm{d}a. \tag{3}$$

Maximizing the first and second terms of this formula corresponds, respectively, to exploitation (obtaining a large reward based on the current value function) and exploration (searching for a better policy). Coefficient β is called inverse-temperature; it controls the balance between exploitation and exploration. This method of introducing the exploration is an undirected one which explores the state-action space based on the action randomness, whereas the above-mentioned exploration bonus is a directed exploration technique that uses the statistics obtained through the past experiences in order to execute efficient exploration.

Using the variational method, maximization of $F(P^\pi)$ is achieved by

$$P^\pi(a|u) = \frac{\exp(\beta Q(u,a))}{\int \exp(\beta Q(u,a))\,\mathrm{d}a} \tag{4}$$

which is called the soft-max policy or the Boltzmann policy [27]. When the inverse-temperature is small, the soft-max policy randomly selects one of the possible actions. When it is large, in contrast, the policy selects a greedy action that maximizes the current action-value function.

Since these two strategies, exploitation and exploration, cannot be operated at once, balancing them has long been an important issue in control field. In a dynamic environment, especially, they should be balanced depending on the current situation; for example, exploration should be encouraged when the agent perceives environmental change. In our previous paper [11], we proposed a two-fold control method for inverse-temperature based on the current estimation of the environment: one is based on variation of the action-value function, and the other on detection of environmental changes. Although the details are omitted here, what is important is that the action selection by (4) is modified so as to depend on the current estimation of the environment.

2.3. Working hypothesis

To deal with a POMDP, our model-based RL method needs the following functional components:

(1) estimation of hidden states;
(2) an environmental model describing the whole state space, including hidden states;
(3) prediction of reward accumulation, which depends on the environmental model;
(4) optimal action selection, which enlarges reward accumulation.

Reward-related activations of PFC neurons have been shown by various single neuron recording studies, and this region is known to be crucial for goal-directed behavior. Since RL-like reward-based learning is considered to be plausible in the brain incorporating dopaminergic systems, PFC receiving dense dopaminergic innervations would have an important role in biological RL. Here, we present a possible brain implementation of our RL method by focusing on the PFC regions: DLPF, OFC, APF and ACC. Fig. 1 shows our hypothetical RL diagram within these regions.

The OFC has dense connections with the limbic system involved in emotion and motivation [17], and plays a crucial role in the motivational control of goal-directed behaviors [20]. Recent studies have suggested that the OFC is implicated in processes of desired outcomes, such as reward, including the representation of reward anticipation [22], the magnitude of received rewards [16] and the acquired relationship between stimuli and rewards [29]. Although neurons encoding rewards have been reported in both the OFC and the DLPF, a recent neuro physiological study that simultaneously recorded neuronal activities in both regions found that the OFC is more likely to encode the reward alone while the DLPF is more likely to encode the combination of reward and action [31]. These studies suggest that the

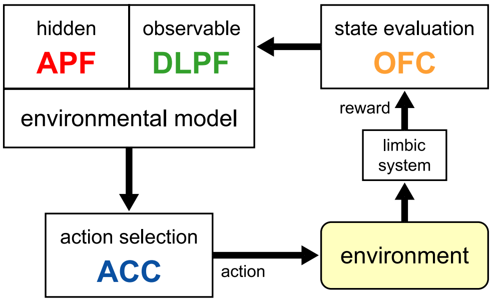

Fig. 1. Hypothesized reinforcement learning diagram within the prefrontal cortex.

OFC is related to rapid stimulus-reward association learning, and we assume that the OFC maintains the evaluation of immediate or short-term accumulated rewards to execute long-term planning.

The DLPF receives sensory inputs processed by other association cortices, and sends outputs to motor systems such as the striatum and the motor association cortex. This region has been studied primarily in relation to higher-order cognitive functions such as attention, working memory and response selection for goal-directed behaviors. Rao et al. [19] reported sustained activities of DLPF neurons depending on state and/or action. In addition, recent recording studies have revealed that DLPF neurons represent the quality and the quantity of future reward [14,32], and it has been suggested that the DLPF is more likely than the OFC to encode the combination of reward and action [31]. Thus, we assume that DLPF represents the long-term estimation of accumulated reward, depending on state and/or action, by integrating various types of information such as reward prediction from the OFC, the history of state observation, and the animal's own action. This function can be regarded as the value function and/or the action-value function in RL.

In model-based RL, the value function is approximated using the environmental model. According to a recent view, based on findings that the DLPF seems to guide behaviors in accordance with environmental requirements, the DLPF constructs automata, cascade networks in which each nodal point generates an action using currently available or memorized information to achieve a behavioral goal [28]. A recent study [18] suggested that the DLPF is involved in preparation for forthcoming sequences of actions based on information stored in working memory. Such behavioral planning requires prediction of environmental changes induced by the animal's own action. We speculate that the environmental models in RL, which predict the next state, are expressed in the DLPF.

The environmental model of our RL method requires estimation of unobservable variables. In a branching task [13], the APF was activated when a subject could not predict whether the forthcoming task would be a primary task or a subtask. An imaging study of an explicit categorization task [25] suggested that a rule change evoked an activation in the APF. From these results, it was proposed that the APF is involved in active switching of behavioral rules without explicit cues. Since such switching reflects the selection of an appropriate subprocess based on unobservable contextual information, we consider that the APF may be related to estimation of unobservable states using past experiences.

In our RL, action selection is based on both the current value function and the detection of environmental change. A primate recording study [24] and a human imaging study [4] revealed that ACC activations are related to voluntary switching of behavioral rules, which is dependent on reward expectation. The ACC is also activated by detection of behavioral error [3] and conflict among incompatible responses [2,5]. We consider that the ACC is associated with the uncertainty of action selection based on the current environmental model maintained in the DLPF. Actually, there are thick innervations from the DLPF to the ACC.

In our hypothetical diagram in Fig. 1, the DLPF maintains and manipulates the environmental model and the reward-based environmental model, i.e., $P(u'|[y, \hat{P}(z)], a)$

and $Q(u, a)$ in (2). The APF estimates unobservable state variables, $\hat{z}$ and $\hat{P}(z)$. These estimations are carried to the ACC, which executes action selection (4).

3. fMRI experiment

3.1. Material and methods

3.1.1. Behavioral task

Sixteen subjects (13 males and 3 females) participated in experiments after giving written informed consent which was reviewed and approved by the ethical committee of Advanced Telecommunications Research Institute International, Japan. All were graduate students in scientific or engineering fields. None of the subjects reported any history of neurological or psychiatric disorders, and all had correct-for-normal vision. Subjects performed sequential learning tasks, and were given a fixed basic monetary payment plus a monetary bonus in proportion to their task scores.

In a sequential learning task, at the start of each trial (Fig. 2(a)), a fixation cross was displayed at the screen center. The cross was surrounded by four squares, and a green trial bar indicating the number of remaining trials was displayed above them. The color of each square was red or gray, and a single task completion was a sequence of eight states, each of which was represented by a color pattern of the four squares. In the initial state of this sequence, all squares were gray; the color of the squares then changed, clockwise, from gray to red in the first round and from red to gray in the second (Fig. 2(b)). For each state, one of a pair of left and right buttons was set as the correct response button and the other was set as the wrong one. In the trial, subjects were required to press either button under their dominant hand within 2 s, but were instructed to respond as quickly and accurately as possible. Immediately after a response, the green trial bar was shortened and the fixation cross disappeared. After that, when the predetermined correct button was pressed, the color of one square changed indicating a successful transition to the next state. If the subject pressed the wrong button, or did not press a button within 2 s, there was no state change, and the same color pattern was displayed as in the previous trial so that the subject was required to try the same state again. Regardless of the response time, the time interval for the state change was 3 s, during which the current color pattern was displayed. The state transition is thus represented by an eight-length automaton, in which the goal state is identical with the initial one. To reach the goal, therefore, subjects had to learn a sequence of eight correct responses, based on feedback indicating whether each response was correct or not.

An experiment consisted of two task conditions, a memory (MEM) condition and an MDP condition, which have different state transition characteristics. In both conditions, since the correct response sequence was not instructed in advance, subjects had to acquire it by trial and error. In the MEM condition with a deterministic state transition, on the one hand, subjects needed simply to memorize the correct response at each state until the first goal was achieved, and thereafter to repeat the memorized sequence of eight responses. In the MDP condition, on the

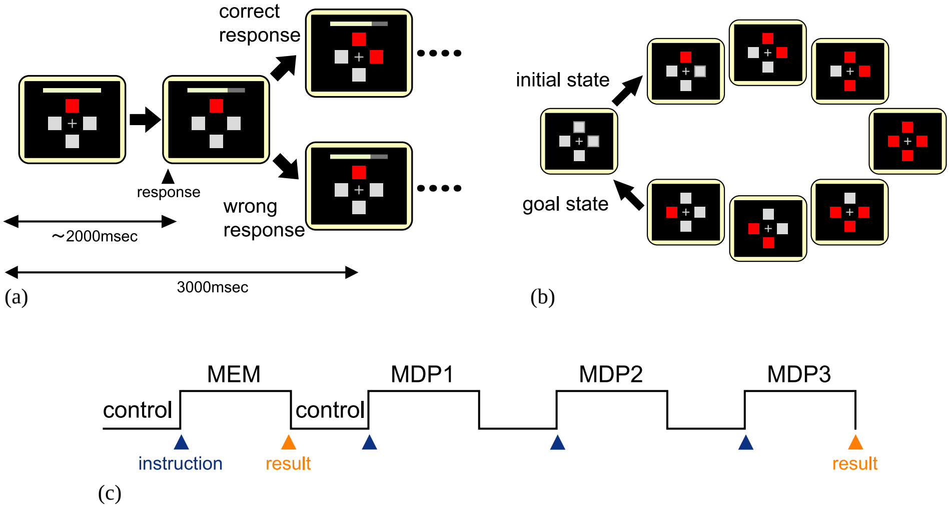

Fig. 2. A sequential learning task using visual stimuli and two response buttons. (a) At the onset of each trial (3 s), four squares surrounding a fixation cross and a trial bar were displayed on the screen. There were eight color patterns of the squares, each of which represented an individual state. A subject was required to press one of the left and right buttons within 2 s, and the response immediately induced the disappearance of the fixation cross and the decrease of the trial bar length. Subsequently, the next stimulus and a fixation cross were presented indicating the next trial. The next stimulus was determined by each response; if the response was correct, the color was changed indicating the next state, while the wrong response caused no state change. (b) A single task completion was a sequence of eight states in which both of the initial and the goal state were represented by four gray squares and the color change of one square indicated a single state transition. The color changed clockwise from gray to red in the first round and from red to gray in the second. (c) An experiment run consisted of one MEM block and three MDP blocks. At the onset of each task block, the visual messages of condition name and block number were displayed. At the end of the MEM block or the third MDP block, the task performance was visually noticed.

other hand, the state transition was first-order Markov, in which a correct response resulted in state transition (success) with 85% or same-state repetition (failure) with 15%; these state transition probabilities did not change. Then, after subjects are certain of the state transition, their action selection becomes almost automatic as in the MEM condition.

Subjects performed two successive 6 min sessions, each of which comprised three MDP blocks of 20 trials and one MEM block of 20 trials (Fig.2(c)). Each behavioral task block was followed by a control condition without the need for either memory or learning. In the control condition, which used the same visual stimuli and buttons as the task conditions, only the left or the right square was turned to red in each trial, and subjects were required to press the corresponding button as a straightforward reaction.

Although the MDP condition was divided into three blocks, the correct response sequence and its probabilities were the same for all three. Thus, subjects needed to retain the learned sequence during the intervening control tasks. Subjects were instructed the condition name by a visual message at the onset of each condition. After each block, the number of bonus points attained, corresponding to the number of achieved goals, was displayed. This indicated the monetary reward added to the basic reward. Even if a subject could not achieve a single goal, indicating zero bonus points, the basic monetary reward was paid. Before entering the scanner, subjects were given details of the task, and performed two training sessions in which the successful state transition probability was set higher than in the scanning MDP condition.

3.1.2. Procedures and analysis

Using a whole-brain 1.5-tesla scanner (Magnetic Eclipse; Shimadzu Marconi, Kyoto, Japan), functional images were obtained with T2*-weighted echo-planar images (EPIs), with blood oxygenation level-dependent (BOLD) contrast (TE, 55 ms; FA, 90 °). Volumes were acquired every 3 s (TR), and contained 28 slices of 5-mm thickness (matrix size, 64×64; FOV, $192 \times 192\,\text{mm}^2$). Stimulus presentation and scanning were synchronized at the beginning of each run. The first six (15 s) EPIs in each session were discarded to avoid T1 equilibrium effects. Each scanning run began with a high-resolution T1-weighted three-dimensional volume acquisition for anatomical localization (voxel size, $1 \times 1 \times 1\,\text{mm}^3$).

The imaging data were analyzed with Statistical Parametric Mapping (SPM99, Wellcome Department of Cognitive Neurology, London, UK), implemented within Matlab 6.5 (Mathworks Inc.). All functional images from each subject were realigned with the first image, and then registered to the individual anatomical image. After that, the co-registered T1 image was normalized into the MNI (Montreal Neurological Institute) template, involving three-dimensional transformations. The parameters from this normalization process were then applied to normalization of each EPI image. The EPI images were reformatted to isometric voxels ($2 \times 2 \times 2\,\text{mm}^3$). The normalized EPIs were spatially smoothed with a Gaussian kernel of 10 mm (FWHM) in the x, y, and z-axes.

Statistical parametric maps of t-statistics were calculated for condition-specific effects within a general linear model. For each MEM or MDP block, sustained activity was modeled as an epoch convolved with a canonical hemodynamic response function. The data were high-pass filtered using low-frequency cosine functions with a cut-off of 168 s. In order to account for inter-subject variability, and to allow statistical inference at the population level, one sample t-test for statistical significance of the group random effect, where the threshold at the voxel level was set to $p < 0.001$ uncorrected and that at the cluster level to $p < 0.05$ corrected, was subsequently applied.

3.2. Results and discussion

3.2.1. Behavioral results

To investigate task performance, we conducted one-way ANOVA using all behavioral data of 32 sessions for 16 subjects. Each subject was required to make prompt and accurate responses in the experiments, and response time (RT), the time

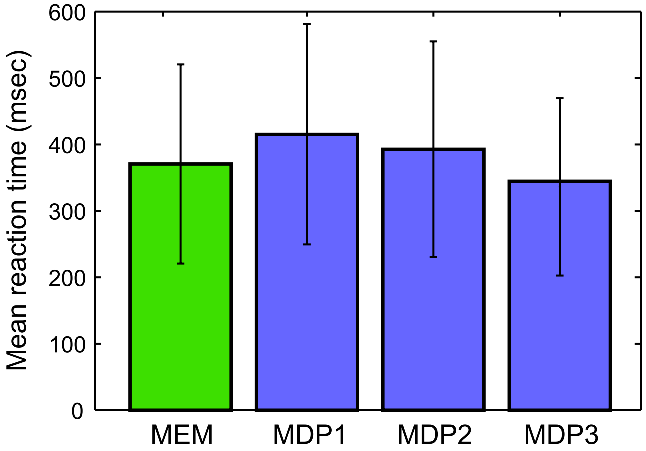

Fig. 3. Mean reaction time (RT) and its standard deviation in each task session. RTs in the MDP condition significantly decreased as the learning blocks proceeded.

interval between the presentation of a stimulus and the initiation of a response was examined. Fig. 3 shows the mean RTs (370.1 ms in the MEM block; 415.2 ms, 392.7 ms and 336.1 ms in the three MDP blocks) and the corresponding standard deviations. Although the RTs were not significantly different between the MEM and MDP conditions, they significantly decreased as the MDP blocks proceeded ($F[2,93] = 3.46, p < 0.05$). All responses in the task conditions can be classified into successful or failed actions, and the error rate (the rate of failed actions) decreased significantly as the three MDP blocks proceeded ($F[2,93] = 27.41, p < 0.001$). These results indicate that the subjects became able to determine their own responses quickly and accurately as the learning proceeded.

We also measured the moving average (window size[2]: 11) both of behavioral correctness by means of overlap with the correct response sequence, and of behavioral variation by means of entropy[3]; both were averaged over all subjects. Figs. 4(a) and (b) show the results for the MEM block and the three MDP blocks, respectively, where the abscissa denotes the number of trials but note that the scale is different in the two figures. The solid and dashed lines correspond to the overlap (right ordinate) and entropy (left ordinate), respectively. Since the entropy decreased, while the overlap increased, with time in both Figs. 4(a) and (b), the behavioral variation decreased as the learning of the correct response sequence proceeded in both conditions. In the MEM condition, subjects almost completely learned the state transition by the end of one block, while the learning in the MDP condition had not been completed at the end of the first MDP block indicated by the

[2] Window size was determined as the average number of trials required to reach the goal state from the initial state.

[3] Behavioral entropy was calculated by $H = -\sum p \log p$, with p being the probability of each action at each state within the sliding window. When only one action occurs at a state, the entropy for the state becomes the minimum value 0.

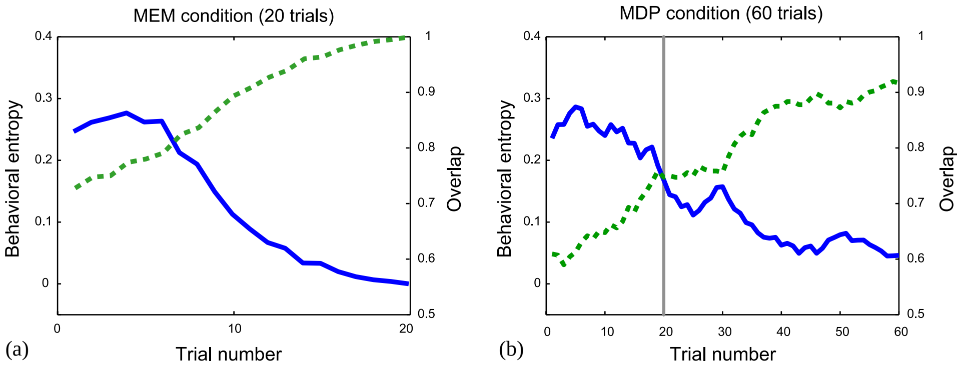

Fig. 4. The moving average of overlap and behavioral entropy in the MEM (a) and MDP (b) conditions. In both conditions, the overlap (solid line) increased and the entropy (dashed line) decreased as learning proceeded. The vertical line in (b) denotes the end of the first session which is equivalent of the maximum trial number in (a).

vertical line. Indeed, in most MDP sessions ($\frac{26}{32}$), exploratory behaviors continued even in the third block. These results suggest that learning of the environmental model (state transition sequence) was continuously conducted throughout the three MDP blocks. In the MEM condition, in contrast, once a subject achieved the goal, his/her behaviors did not fluctuate since there was no longer any need to renew the environmental model.

3.2.2. *Imaging results*

Comparison of the MDP and MEM conditions revealed significant increases in activations, mainly in four regions: the posterior and anterior DLPF, inferior parietal cortex (PCi) and ACC. Figs. 5(a) and (b) show the Z-value map of random-effect analysis for the 16 subjects, rendered onto a three-dimensional template, and a normalized structural MRI image for anatomical visualization, respectively. Table 1 summarizes the statistics of peak voxels in the activated clusters. To investigate activation changes associated with learning progression, we examined the average activation in these four regions in the MEM and MDP blocks. In Fig. 6, the left panel in each subfigure shows the activated region in an anatomical image, and the right bar graph indicates the mean percent signal changes of the peak voxel in the corresponding region.

Although correct action sequences (automata) in both of the MEM and MDP conditions have a length of eight, subjects in the MDP condition were required to retain some possible sequences (environmental models), especially at the early learning stage, due to the stochastic nature of the task. We consider that the significant activation increase in the PFC observed during the MDP blocks is related to the manipulation and maintenance of the automata representing possible environmental models.

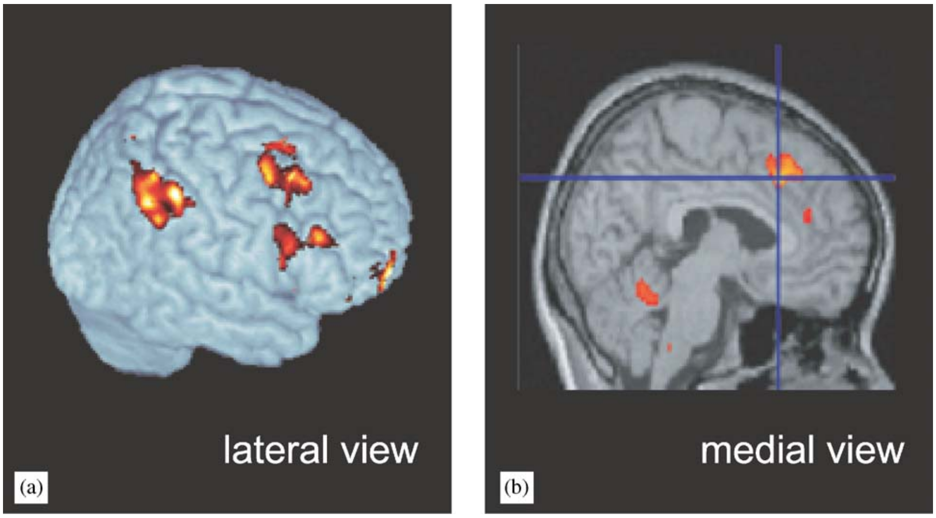

Fig. 5. Significant activation during the MDP condition compared to the MEM condition with a group random-effect model (corrected to $p<0.05$).

Table 1
Statistics of significantly activated regions during the MDP task

Brain region			Z-value	Talairach		
Region	Side	BA		x	y	z
Prefrontal cortex						
Middle frontal gyrus	R	8	4.35	34	25	41
Middle frontal gyrus	R	46/9	3.57	48	32	21
Parietal cortex						
Inferior parietal gyrus	R	40/7	5.34	42	−46	43
Inferior parietal gyrus	L	40/7	5.14	−48	−40	44
Anterior cingulate cortex						
Cingulate gyrus	R	32	4.15	6	22	38
Cingulate gyrus	L	32	3.96	−2	24	44

A recent event-related fMRI study [21] revealed that both the posterior DLPF (BA 8) and the PCi (BA $\frac{40}{7}$) showed sustained activities when a subject maintained spatial information during a delay period, while no activation was associated with the selection of responses from stored information. Because these regions are related to the maintenance of working memory, and not to executive processing, we suggest

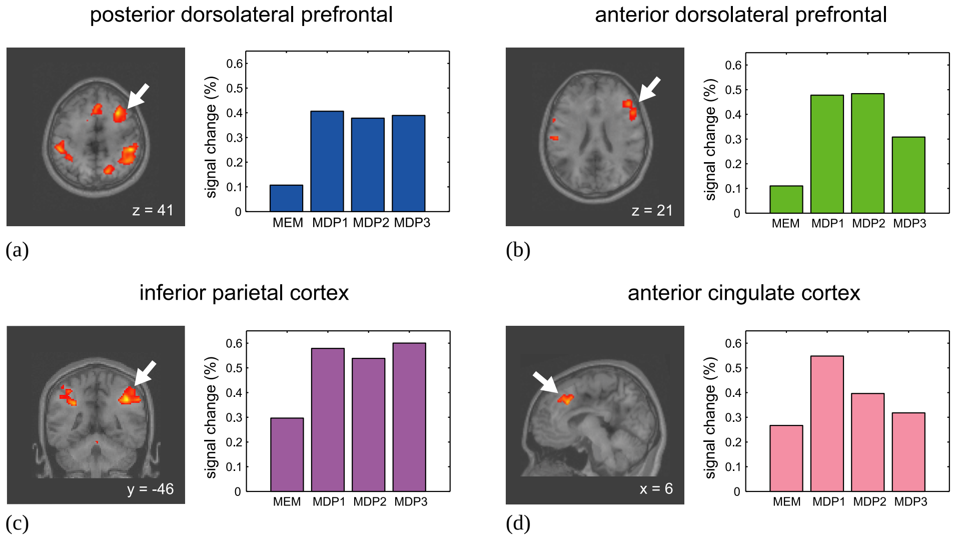

Fig. 6. Major activated brain regions during the MDP condition (left panels) and their signal changes (%) in the MEM and three MDP blocks (right panels, bar graphs). (a) Posterior dorsolateral prefrontal, (b) anterior dorsolateral prefrontal, (c) inferior parietal cortex, and (d) anterior cingulate cortex.

that the activations observed in the posterior DLPF and PCi are involved in the maintenance of environmental models. In our learning tasks, the environmental model must be maintained in working memory even if the correct sequence has already been obtained, and both regions actually showed constant activation ($F[2,93] = 1.12, p = 0.33$[4]) from the first to the last MDP blocks (Fig. 6(a,c)).

A significant activation increase in the anterior DLPF (BA $\frac{46}{9}$) was accompanied by the reproduction of sequentially represented stimuli after a memory delay [18] and the processing of sequence memories [6]. Thus, this region is thought to be involved in the manipulation of sequential information stored in working memory. In the MDP blocks, subjects were required to retain multiple candidates of a response sequence and to manipulate them until the correct sequence was acquired. The anterior DLPF may play a role in temporal processing such as mental simulation of environmental models or updating of stored sequential information. When subjects repeated a learned sequence in the last MDP block, this region showed a significant decrease ($F[1,62] = 8.80, p < 0.005$) in activation, while no such decrease occurred in the posterior DLPF (Fig. 6(b)). This finding supports our interpretation that the posterior and anterior DLPF respectively maintains and manipulates environmental models.

We also found a significant activation increase in the ACC during the MDP condition, and this activation decreased significantly ($F[2,93] = 10.53, p < 0.0001$) as

[4] The difference is not significant.

the MDP blocks proceeded (Fig. 6(d)). The decrease in behavioral variation during learning could be related to the decrease in the activity of the ACC. This interpretation is consistent with our hypothesis, in which the ACC is responsible for the control of action selection based on competition among possible actions using current evaluation of the environment.

In summary, this imaging study suggests that the posterior DLPF maintains, and the anterior DLPF manipulates, possible environmental models, while the ACC represents the uncertainty of the action selection.

4. Concluding remarks

In this study, we have presented a model-based RL method in which the environment is directly estimated. To adapt to changes in the environment, a control method of action selection was introduced.

We proposed a possible functional model of our RL method, in which the DLPF maintains and manipulates the environmental models and the ACC is related to action selection, and conducted an fMRI experiment. The experimental results were consistent with our hypothesis. Although we also suggest that the estimation of unobservable states in RL is expressed in the APF, this possibility has not yet been examined, and is an issue for future study.

References

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

[2] M. Botvinick, L.E. Nystrom, K. Fissell, C.S. Carter, J.D. Cohen, Conflict monitoring versus selection-for-action in anterior cingulate cortex, Nature 402 (1999) 179–181.

[3] T.S. Braver, D.M. Barch, J.R. Gray, D.J. Molfese, A. Snyder, Anterior cingulate cortex and response conflict: effects of frequency inhibition and errors, Cereb. Cortex 11 (2001) 825–836.

[4] G. Bush, B.A. Vogt, J. Holmes, A.M. Dale, D. Greve, M.A. Jenike, B.R. Rosen, Dorsal anterior cingulate cortex: a role in reward-based decision making, Proc. Natl. Acad. Sci. USA 99 (2002) 507–512.

[5] C.S. Carter, T.S. Braver, D.M. Barch, M.M. Botvinick, D. Noll, J.D. Cohen, Anterior cingulate cortex, error detection, and the online monitoring of performance, Science 280 (1998) 747–749.

[6] J.D. Cohen, W.M. Perlstein, T.S. Braver, L.E. Nystrom, D.C. Noll, J. Jonides, E.E. Smith, Temporal dynamics of brain activation during a working memory task, Nature 386 (1997) 604–608.

[7] P. Dayan, T.J. Sejnowski, Exploration bonuses and dual control, Mach. Learn. 25 (1996) 5–22.

[8] R. Dearden, N. Friedman, D. Andre, Model based Bayesian exploration, in: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufman, San Francisco, CA, 1999, pp. 150–159.

[9] K. Doya, Reinforcement learning in continuous time and space, Neural Comput. 12 (2000) 219–245.

[10] R.A. Howard, Information value theory, IEEE Trans. Syst. Sci. Cybernet. SSC-2 (1) (1996) 22–26.

[11] S. Ishii, W. Yoshida, J. Yoshimoto, Control of exploitation-exploration meta-parameter in reinforcement learning, Neural Networks 15 (2002) 665–687.

[12] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artif. Intell. 101 (1998) 99–134.

[13] E. Koechlin, G. Corrado, P. Pietrini, J. Grafman, Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning, Proc. Natl. Acad. Sci. USA 97 (2000) 7651–7656.

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

[15] A.W. Moore, C.G. Atkeson, Prioritized sweeping: reinforcement learning with less data and less real time, Mach. Learn. 13 (1993) 103–130.

[16] J. O'Doherty, M.L. Kringelbach, E.T. Rolls, J. Hornak, C. Andrews, Abstract reward and punishment representations in the human orbitofrontal cortex, Nat. Neurosci. 4 (2001) 95–102.

[17] M. Petrides, D.N. Pandya, Comparative architectonic analysis of the human and macaque frontal cortex, in: J. Graftman, F. Boller (Eds.), Handbook of Neuropsychology, Elsevier, Amsterdam, 1995.

[18] J.B. Pochon, R. Levy, J.B. Poline, S. Crozier, S. Lehericy, B. Pillon, B. Deweer, D. Le Bihan, B. Dubois, The role of dorsolateral prefrontal cortex in the preparation of forthcoming actions: an fMRI study, Cereb. Cortex 11 (2001) 260–266.

[19] S.C. Rao, G. Rainer, E.K. Miller, Integration of what and where in the primate prefrontal cortex, Science 276 (1997) 821–824.

[20] E.T. Rolls, The orbitofrontal cortex, Philos. Trans. Roy. Soc. London. Series B: Biol. Sci. 351 (1996) 1433–1443.

[21] J.B. Rowe, I. Toni, O. Josephs, R.S.J. Frackowiak, R.E. Passingham, The prefrontal cortex: response selection or maintenance within working memory?, Science 288 (2000) 1656–1660.

[22] G. Schoenbaum, A.A. Chiba, M. Gallagher, Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning, Nat. Neurosci. 1 (1998) 155–159.

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

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

[25] B.A. Strange, R.N.A. Henson, K.J. Friston, R.J. Dolan, Anterior prefrontal cortex mediates rule learning in humans, Cereb. Cortex 11 (2001) 1040–1046.

[26] R.S. Sutton, Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, in: Proceeding of the Seventh International Conference on Machine Learning, Morgan Kaufmann, Los Altos, CA, 1990, pp. 216–224.

[27] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, 1998.

[28] J. Tanji, E. Hoshi, Behavioral planning in the prefrontal cortex, Curr. Opin. Neurobiol. 11 (2001) 164–170.

[29] L. Tremblay, W. Schultz, Relative reward preference in primate orbitofrontal cortex, Nature 398 (1999) 704–708.

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

[31] J.D. Wallis, E.K. Miller, Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task, Euro. J. Neurosci. 18 (2003) 2069–2081.

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

Wako Yoshida is a researcher with Graduate School of Information Science at Nara Institute of Science and Technology. She received her B.A. in 1998 from Kobe College, M.E. in 2000 and Ph.D. in 2003 both from Nara Institute of Science and Technology. Her research interest includes theoretical and experimental approach to human's higher order functions such as learning, memory and communication.

Shin Ishii received his B.E. in 1986, M.E. in 1988, and Ph.D. in 1987 from University of Tokyo. He is a professor of Graduate School of Information Science at Nara Institute of Science and Technology. His current research interests are computational neuroscience, systems neurobiology and statistical learning theory.