

# Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops

Saori C Tanaka<sup>1-3</sup>, Kenji Doya<sup>1-3</sup>, Go Okada<sup>3,4</sup>, Kazutaka Ueda<sup>3,4</sup>, Yasumasa Okamoto<sup>3,4</sup> & Shigeto Yamawaki<sup>3,4</sup>

Evaluation of both immediate and future outcomes of one's actions is a critical requirement for intelligent behavior. Using functional magnetic resonance imaging (fMRI), we investigated brain mechanisms for reward prediction at different time scales in a Markov decision task. When human subjects learned actions on the basis of immediate rewards, significant activity was seen in the lateral orbitofrontal cortex and the striatum. When subjects learned to act in order to obtain large future rewards while incurring small immediate losses, the dorsolateral prefrontal cortex, inferior parietal cortex, dorsal raphe nucleus and cerebellum were also activated. Computational model-based regression analysis using the predicted future rewards and prediction errors estimated from subjects' performance data revealed graded maps of time scale within the insula and the striatum: ventroanterior regions were involved in predicting immediate rewards and dorsoposterior regions were involved in predicting future rewards. These results suggest differential involvement of the cortico-basal ganglia loops in reward prediction at different time scales.

In daily life, people make decisions based on the prediction of rewards at different time scales; for example, one might do daily exercise to achieve a future fitness goal, or resist the temptation of sweets to avoid future weight gain. Damage to the prefrontal cortex often impairs daily decision making, which requires assessment of future outcomes<sup>1,2</sup>. Lesions in the core of the nucleus accumbens in rats result in a tendency to choose small immediate rewards over larger future rewards<sup>3</sup>. Low activity of the central serotonergic system is associated with impulsive behavior in humans<sup>4</sup>, and animals with lesions in the ascending serotonergic pathway tend to choose small immediate rewards over larger future rewards<sup>5,6</sup>. A possible mechanism underlying these observations is that different sub-loops of the topographically organized cortico-basal ganglia network are specialized for reward prediction at different time scales and that they are differentially activated by the ascending serotonergic system<sup>7</sup>. To test whether there are distinct neural pathways for reward prediction at different time scales, we developed a 'Markov decision task' in which an action affects not only the immediate reward but also future states and rewards. Using fMRI, we analyzed brain activity in human subjects as they performed this task. Recent functional brain imaging studies have shown the involvement of specific brain areas, such as the orbitofrontal cortex (OFC) and the ventral striatum, in prediction and perception of rewards<sup>8-11</sup>. In these previous studies, however, rewards were given either independent of the subject's actions or as a function of the current action. Our Markov decision task probes decision making in a dynamic context, with small losses followed by a large positive reward. The results of the block-design analysis suggest differential involvement of brain areas in decision making by prediction of rewards at different time scales. By analyzing subjects' performance

data according to a theoretical model of reinforcement learning, we found a gradient of activation within the insula and the striatum for prediction of rewards at different time scales.

## RESULTS

### Behavioral results

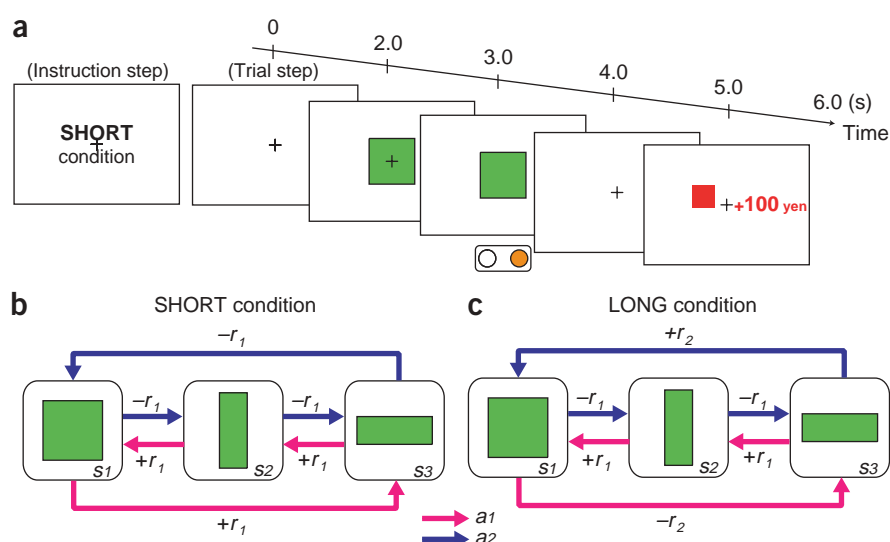
In the Markov decision task, a visual signal (one of three shapes) was presented at the start of each trial to indicate one of three states, and the subject selected one of two actions: pressing the right or left button with the right hand (Fig. 1a; see Methods for details). For each state, the subject's action choice affected not only the immediate reward, but also the state subsequently presented (Fig. 1b,c).

The rule of state transition was fixed during the entire experiment (Fig. 1), but the rules of reward delivery changed according to the task condition. In the SHORT condition, action  $a_1$  gives a small positive reward ( $+r_1 = 20$  yen average; see Methods) and action  $a_2$  gives a small loss ( $-r_1$ ) in all three states (Fig. 1b). The optimal behavior for maximizing total reward in the SHORT condition is to collect small positive rewards by taking action  $a_1$  at each state. In the LONG condition, action  $a_2$  at state  $s_3$  gives a big bonus ( $+r_2 = 100$  yen average; see Methods), and action  $a_1$  at state  $s_1$  results in a big loss ( $-r_2$ ; Fig. 1c). The optimal behavior is to receive small losses at state  $s_1$  and  $s_2$  to obtain a large positive reward at state  $s_3$  by taking action  $a_2$  at each state; this is opposite to the optimal behavior in the SHORT condition. Whereas the optimal strategy in the SHORT condition results in small, immediate rewards at each step, the optimal strategy in the LONG condition results in small immediate losses but a net positive reward by the end of one cycle. Thus, for successful action in the LONG condition, subjects must consider both the

<sup>1</sup>Department of Bioinformatics and Genomics, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan. <sup>2</sup>Department of Computational Neurobiology, ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan. <sup>3</sup>CREST, Japan Science and Technology Agency, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan. <sup>4</sup>Department of Psychiatry and Neurosciences, Hiroshima University, 1-2-3 Kasumi, Minamiku, Hiroshima 734-8551, Japan. Correspondence should be addressed to K.D. (doya@atr.jp).

**Figure 1** Experimental design. (a) Sequences of stimulus and response events in the task. At the beginning of each condition block, the condition is informed by displaying text (6 s), such as 'SHORT condition' (instruction step). In each trial step, a fixation point is presented on the screen, and after 2 s, one of three shapes (square, vertical rectangle or horizontal rectangle) is presented for 1 s. As the fixation point vanishes after 1 s, the subject presses either the right or left button within 1 s. After a short delay (1 s), a reward for that action is presented by a number (indicating yen gained or lost) and the past cumulative reward is shown by a bar graph. Thus, one trial step takes 6 s.

(b,c) The rules of the reward and state transition for action  $a_1$  (magenta arrows) and action  $a_2$  (blue arrows) in the SHORT (b) and LONG (c) conditions. The small reward  $r_1$  was 10, 20 or 30 yen, with equal probability, and the large reward  $r_2$  was 90, 100, or 110 yen. The rule of state transition was the same for all conditions:  $s_3 \rightarrow s_2 \rightarrow s_1 \rightarrow s_3 \dots$  for action  $a_1$ , and  $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_1 \rightarrow \dots$  for action  $a_2$ . Although the optimal behaviors are opposite (SHORT:  $a_1$ ; LONG:  $a_2$ ), the expected cumulative reward during one cycle of the optimal behavior is 60 yen in both the SHORT (+20  $\times$  3) and LONG (−20, −20, +100) conditions.



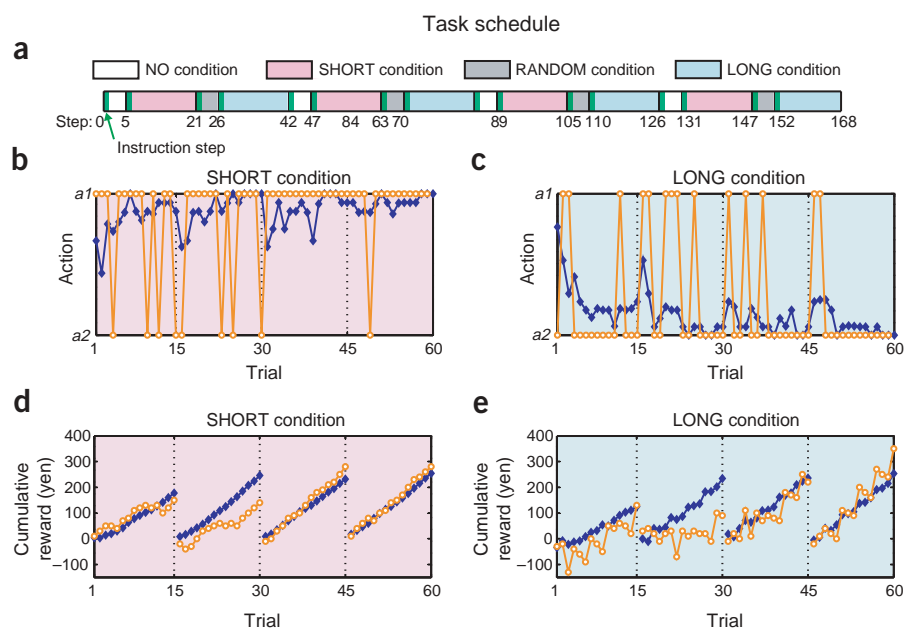
immediate reward and the future reward expected from the subsequent state, and for success in the SHORT condition, subjects need to consider only the immediate outcome of their actions. Subjects performed 15 trials in a SHORT condition block and 15 trials in a LONG condition block. There were also two control conditions, NO (reward was always zero) and RANDOM (reward was + $r_1$  or − $r_1$ , regardless of state or action), so a total of four condition blocks were performed (see Fig. 2a for task schedule).

All subjects successfully learned the optimal behaviors: taking action  $a_1$  in the SHORT condition (Fig. 2b) and action  $a_2$  in the LONG condition (Fig. 2c). Cumulative rewards within each SHORT block (Fig. 2d) and LONG block (Fig. 2e) also indicate successful learning. It can be seen from the single-subject data in the LONG

condition (Fig. 2e, orange) that the subject learned to lose small amounts (− $r_1$ ) twice to get a big bonus (+ $r_2$ ). The average cumulative reward in the last block was 254 yen in the SHORT condition and 257 yen in the LONG condition, which was 84.7% and 85.7%, respectively, of the theoretical optimum of 300 yen.

### Block-design analysis

To find the brain areas that are involved in immediate reward prediction, we compared brain activity during the SHORT condition and the NO condition, in which reward was always zero. In the SHORT versus NO contrast, a significant increase in activity was observed in the lateral OFC (Fig. 3a), the insula and the occipitotemporal area (OTA) (Fig. 3b), as well as in the striatum, the globus pallidus (GP) (Fig. 3c) and the medial cerebellum (Fig. 3d) (threshold of  $P < 0.001$ , uncorrected for multiple comparisons). These areas may be involved in reward prediction based on immediate outcome.



**Figure 2** Task schedule and behavioral results. (a) A set of four condition blocks—NO (4 trials), SHORT (15 trials), RANDOM (4 trials), LONG (15 trials)—was repeated four times. At the beginning of each condition block, the task condition was presented to the subject (instruction step); thus, the entire experiment consisted of 168 steps (152 trial steps and 16 instruction steps). (b,c) The selected action of a representative single subject (orange) and the group average ratio of selecting  $a_1$  (blue) in the (b) SHORT and (c) LONG conditions. (d,e) The accumulated reward in each block of a representative single subject (orange) and the group average (blue) in the (d) SHORT and (e) LONG conditions. To clearly show the learning effects, data from four trial blocks in the SHORT and LONG conditions are concatenated, with the dotted lines indicating the end of each condition block.

To identify areas involved in future reward prediction, we compared the brain activity during LONG and SHORT conditions. In the LONG versus SHORT contrast, a robust increase in activity was observed in the ventrolateral prefrontal cortex (VLPFC), the insula, the dorsolateral prefrontal cortex (DLPFC), the dorsal premotor cortex (PMd), the inferior parietal cortex (IPC) (Fig. 4a), the striatum, GP (Fig. 4b), the dorsal raphe nucleus (Fig. 4c), the lateral cerebellum (Fig. 4d), the posterior cingulate cortex and the subthalamic nucleus ( $P < 0.001$ , uncorrected). Activity in the striatum was highly significant (threshold at  $P < 0.05$ , corrected for a small volume when using an anatomically defined region of interest (ROI) in the striatum; see Methods). These areas are specifically involved in decision making based on the prediction of reward in multiple steps in the future. In the LONG versus NO contrast, the activated areas were approximately the union of the areas activated in the SHORT versus NO and LONG versus SHORT contrasts. These results were consistent with our expectation that both immediate and future reward prediction were required in the LONG condition. The results of block-design analysis, including the LONG versus NO contrast, are summarized in **Supplementary Table 1** online. Activations in both SHORT and LONG conditions were stronger in the first two blocks, when subjects were involved in active trial and error, than in the last two blocks when the subjects' behavior became repetitive.

We compared the activations in the SHORT versus NO contrast and the LONG versus SHORT contrast, and observed that three regions showed significant activity in both contrasts: the lateral prefrontal cortex (lateral OFC and VLPFC), the insula and the anterior striatum (Fig. 5). In the lateral PFC (Fig. 5a), although the activities in lateral OFC for the SHORT versus NO contrast (red) and in the VLPFC for the LONG versus SHORT contrast (blue) were close in location, they were clearly separated on the cortical surface. Activities in the insula were also separated (Fig. 5b). In the anterior striatum (Fig. 5c), we found limited overlaps between the two contrasts (green). In all three areas, activations in the SHORT versus NO contrast were found in the ventral parts, whereas activations in the LONG versus SHORT contrast were found in the dorsal parts.

These results of block-design analysis suggest differential involvement of brain areas in predicting immediate and future rewards.

### Performance-based multiple regression analysis

To further clarify the brain structures specific to reward prediction at different time scales, we estimated how much reward the subjects should have predicted on the basis of their performance data and used their time courses as the explanatory variables of regression

analysis. We took the theoretical framework of temporal difference (TD) learning<sup>12</sup>, which has been successfully used for explaining reward-predictive activations of the midbrain dopaminergic system as well as those of the cortex and the striatum<sup>8,11,13–16</sup>. In TD learning theory, the predicted amount of future reward starting from a state  $s(t)$  is formulated as the 'value function'

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

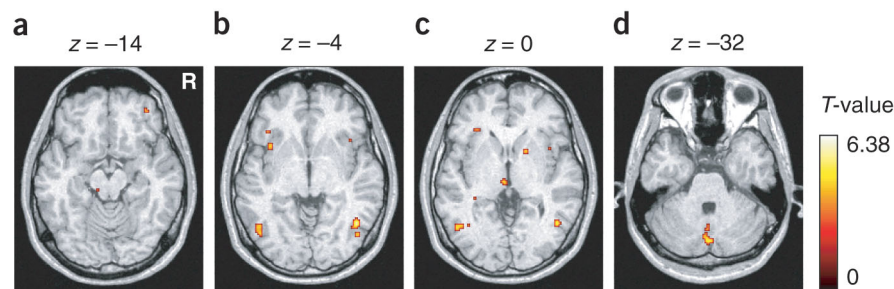
Any deviation from the prediction is given by the TD error

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

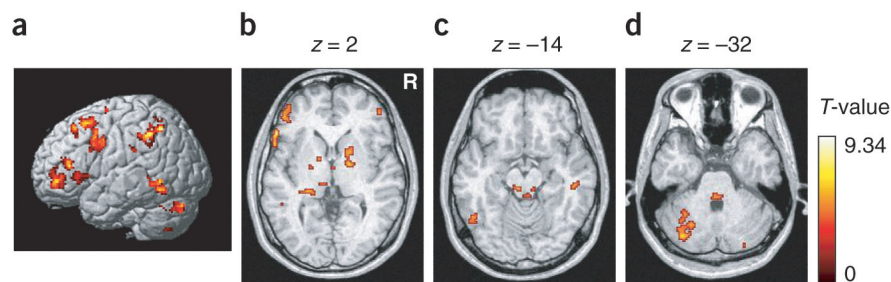
which is a crucial learning signal for reward prediction and action selection. The 'discount factor'  $\gamma$  ( $0 \leq \gamma < 1$ ) controls the time scale of prediction: when  $\gamma = 0$ , only the immediate reward  $r(t+1)$  is considered, but as  $\gamma$  approaches 1, rewards in the further future are taken into account.

We estimated the time courses of reward prediction  $V(t)$  and prediction error  $\delta(t)$  from each subject's performance data and used them as the explanatory variables in multiple regression analysis with fMRI data (see Methods). In our Markov decision task, the minimum value of  $\gamma$  needed to find the optimal action in the LONG condition is 0.36, and any small value of  $\gamma$  is sufficient in the SHORT condition. From the results of our block-design analysis, we assumed that different networks involving the cortex and basal ganglia are specialized for reward prediction at different time scales and that they work in parallel, depending on the requirement of the task. Thus, we varied the discount factor  $\gamma$  as 0, 0.3, 0.6, 0.8, 0.9 and 0.99: small  $\gamma$  for immediate reward prediction and large  $\gamma$  for long future reward prediction. An example of these time courses is shown in **Supplementary Figure 1** online.

We observed a significant correlation with reward prediction  $V(t)$  in the medial prefrontal cortex (mPFC; including the anterior cingulate cortex (ACC) and the medial OFC) (Fig. 6a) and bilateral insula (Fig. 6b), left hippocampus and left temporal pole ( $P < 0.001$ , uncorrected; see **Supplementary Table 2** online). **Figure 6** shows the correlated



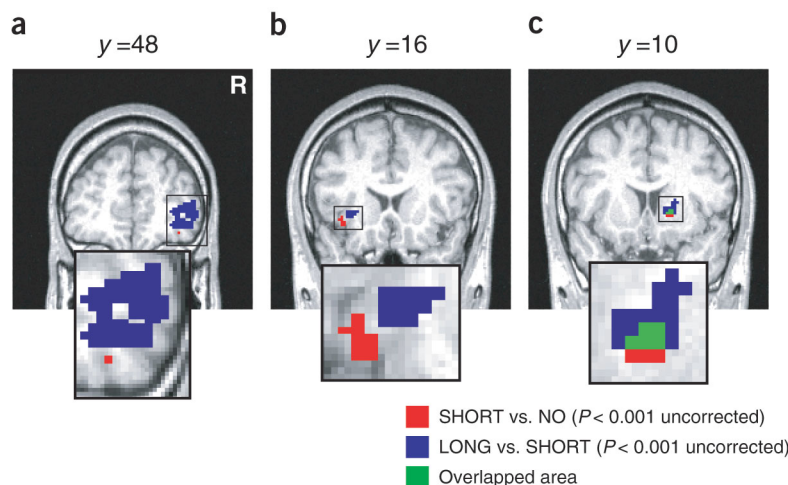
**Figure 3** Brain areas activated in the SHORT versus NO contrast ( $P < 0.001$ , uncorrected; extent threshold of four voxels). (a) Lateral OFC. (b) Insula. (c) Striatum. (d) Medial cerebellum.



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



**Figure 5** Comparison of brain areas activated in the SHORT versus NO contrast (red) and the LONG versus SHORT contrast (blue). (**a–c**) These figures show activation maps focused on (**a**) the lateral OFC (red ( $x, y, z$ ) = (38, 46, -14); blue (46, 47, 3)) (**b**) the insula (red (-36, 13, -4); blue (-30, 18, 1)), and (**c**) the striatum (red (18, 10, 0); blue (18, 12, 3)) where we observed significant activation in both contrasts. The areas where activity overlapped area are shown in green.



voxels within these areas using a gradient of colors for different  $\gamma$  values (red for  $\gamma = 0$ , blue for  $\gamma = 0.99$ ). Activity in the mPFC, temporal pole and hippocampus correlated with reward prediction with a longer time scale ( $\gamma \geq 0.6$ ). Furthermore, in the insula, we found a graded map of activity for reward prediction at different time scales (Fig. 6b). Whereas activity in the ventroanterior region correlated with reward prediction at a shorter time scale, activity in the dorsoposterior region correlated with reward prediction at a longer time scale.

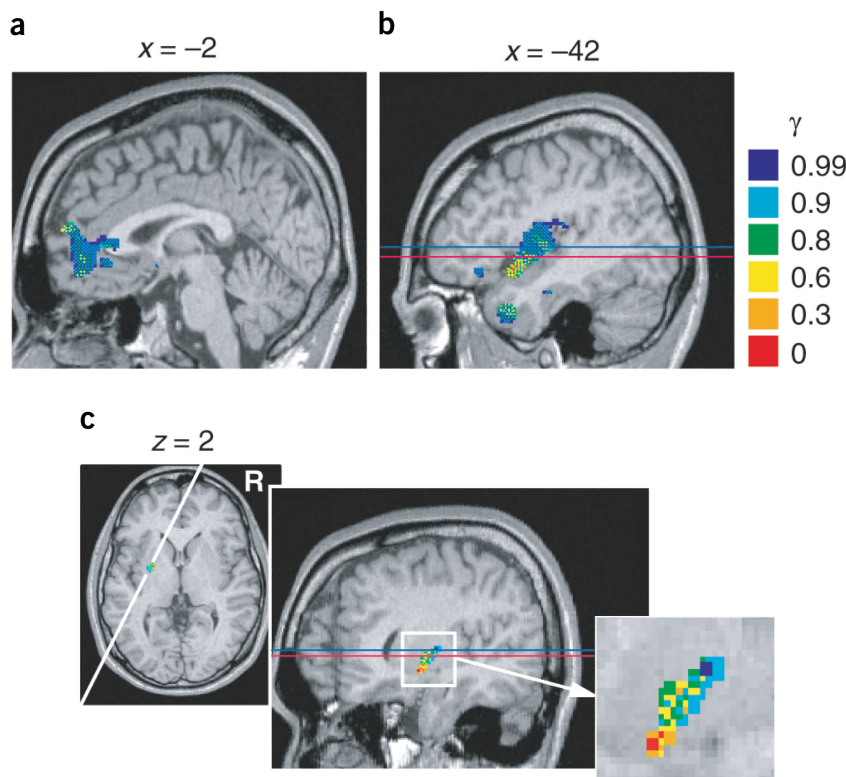
We also found, in the basal ganglia, significant correlation with reward prediction error  $\delta(t)$  using a wide range of time scales (Fig. 6c;  $P < 0.001$ , uncorrected; see Supplementary Table 3 online and Methods). Again, we found a graded map, which had a short time scale in the ventroanterior part and a long time scale in the dorsoposterior part. The coincidence of the ventroanterior-dorsoposterior maps and the ventroanterior-dorsoposterior shifts in activities (Fig. 6b,c) indicate that, while the ventroanterior regions with smaller  $\gamma$  were predominantly active in the SHORT condition, the dorsoposterior regions with larger  $\gamma$  became more active in the LONG condition.

## DISCUSSION

The results of the block-design and performance-based regression analyses suggest differential involvement of brain areas in action learning by prediction of rewards at different time scales. Both block-design and performance-based regression analyses showed activity in the insula and the anterior striatum. Activations of the ventral region in the SHORT versus NO contrast and the dorsal region in the LONG versus SHORT contrast in each area (Fig. 5) are consistent with the ventroanterior-dorsoposterior maps of the discount factor  $\gamma$  found in performance-based regression analysis (Fig. 6).

The insula takes a pivotal position in reward processing by receiving primary taste and visceral sensory input<sup>17</sup> and sending output to the OFC<sup>18</sup> and the striatum<sup>19</sup>. Previous studies showed that the insula is activated with anticipation of primary reward<sup>10</sup> and that insular lesion causes deficits in incentive learning for primary reward<sup>20</sup>. Our

results confirm the role of the insula in prediction of non-primary, monetary reward<sup>21</sup>, and further suggest heterogeneous organization within the insula. Previous imaging studies also showed involvement of the insula, especially the ventroanterior region, in processing aversive outcomes<sup>22,23</sup>. Thus a possible interpretation of the activation of the insula in the LONG condition is that it



**Figure 6** Voxels with a significant correlation (height threshold  $P < 0.001$ , uncorrected; extent threshold of four voxels) with reward prediction  $V(t)$  and prediction error  $\delta(t)$  are shown in different colors for different settings of the discount factor  $\gamma$ . Voxels correlated with two or more regressors are shown by a mosaic of colors. (**a**, **b**) Significant correlation with reward prediction  $V(t)$ . (**a**) mPFC. (**b**) Insula. (**c**) Significant correlation with reward prediction error  $\delta(t)$  restricted to ROI in the striatum (slice at white line in horizontal slice at  $z = 2$  mm). Note the ventroanterior-to-dorsoposterior gradient with the increase in  $\gamma$  both in the insula and the striatum. Red and blue lines correspond to the  $z$ -coordinate levels of activation peaks in the insula and striatum shown in Figure 5b,c (red for the SHORT versus NO and blue for the LONG versus SHORT contrasts).

reflected the losses that subjects acquired before getting a large reward. However, we also ran a regression analysis using losses and found significant correlation in the ventroanterior region of the insula. Anatomical and physiological studies of the insula also showed involvement of its ventroanterior part in perception of aversive stimuli<sup>17</sup>. Thus we argue that the activation of dorsoposterior insula is not simply due to losses in the LONG condition.

Previous brain imaging and neural recording studies suggest a role for the striatum in prediction and processing of reward<sup>9,10,14,21,24–29</sup>. Consistent with previous fMRI studies<sup>8,11,16</sup>, our results showed striatal activity correlated with the error of reward prediction. Reinforcement learning models of the basal ganglia<sup>13–15</sup> posit that the striatum learns reward prediction and action selection based on the reward prediction error  $\delta(t)$  represented by the dopaminergic input. Correlation of striatal activity with reward prediction error  $\delta(t)$  could be due to dopamine-dependent plasticity of cortico-striatal synapses<sup>30</sup>.

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

In reinforcement learning theory, an optimal action selection is realized by taking the action  $a$  that maximizes the 'action value'  $Q(s, a)$  at a given state  $s$ . The action value is defined as  $Q(s, a) = E[r(s, a) + \gamma V(s', a)]$  and represents the expected sum of the immediate reward  $r(s, a)$  and the weighted future rewards  $V(s', a)$ , where  $s'(s, a)$  means the next state reached by taking an action  $a$  at a state  $s$  (refs. 12,15). According to this framework, we can see that prediction of immediate reward  $r(s, a)$  is helpful for action selection based on rewards at either short or long time scales, that is, with any value of the discount factor  $\gamma$ . On the other hand, prediction of state transition  $s'(s, a)$  is helpful only in long-term reward prediction with positive values of  $\gamma$ .

In the lateral OFC, we observed significant activity in both the SHORT versus NO and the LONG versus NO contrasts (Supplementary Table 1 online), but no significant correlation with reward prediction  $V(t)$  or reward prediction error  $\delta(t)$  in regression analysis. This suggests that the lateral OFC takes the role of predicting immediate reward  $r(s, a)$ , which is used for action selection in both SHORT and LONG conditions, but not in the NO condition. This interpretation is consistent with previous studies demonstrating the OFC's role in prediction of rewards, immediately following sensorimotor events<sup>31,32</sup>, and action selection based on reward prediction<sup>23,33,34</sup>.

In the DLPFC, PMd and IPC, there were significant activations in both the LONG versus NO and the LONG versus SHORT contrasts (Supplementary Table 1 online) but no significant correlation with either  $V(t)$  or  $\delta(t)$ . A possible interpretation is that this area is involved in prediction of future state  $s'(s, a)$  in the LONG condition but not in the SHORT or NO conditions. This interpretation is consistent with previous studies showing the role of these cortical areas in imagery<sup>35</sup>, working memory and planning<sup>36,37</sup>.

The dorsal raphe nucleus was activated in the LONG versus SHORT contrast, but was not correlated with  $V(t)$  or  $\delta(t)$ . In consideration of its serotonergic projection to the cortex and the striatum and serotonin's implication with behavioral impulsivity<sup>4–6</sup>, a possible role for the dorsal raphe nucleus is to control the effective time scale of reward prediction<sup>7</sup>. Its higher activity in the LONG condition, where a large setting of  $\gamma$  is necessary, is consistent with this hypothesis.

Let us consider the present experimental results in light of the anatomy of cortico-basal ganglia loops (illustrated in Supplementary

Fig. 2). The cortex and the basal ganglia both have parallel loop organization, with four major loops (limbic, cognitive, motor and oculomotor) and finer, topographic sub-loops within each major loop<sup>38</sup>. Our results suggest that the areas within the limbic loop<sup>39</sup>, namely the lateral OFC and ventral striatum, are involved in immediate reward prediction. On the other hand, areas within the cognitive and motor loops<sup>38</sup>, including the DLPFC, IPC, PMd and dorsal striatum, are involved in future reward prediction. The connections from the insula to the striatum are topographically organized, with the ventral-anterior, agranular cortex projecting to the ventral striatum and the dorsal-posterior, granular cortex projecting to the dorsal striatum<sup>19</sup> (see Supplementary Fig. 2). The graded maps shown in Figure 6b,c are consistent with this topographic cortico-striatal organization and suggest that areas that project to the more dorsoposterior part of the striatum are involved in reward prediction at a longer time scale. These results are consistent with the observations that localized damages within the limbic and cognitive loops manifest as deficits in evaluation of future rewards<sup>1,3,34,40,41</sup> and learning of multi-step behaviors<sup>42</sup>. The parallel learning mechanisms in the cortico-basal ganglia loops used for reward prediction at a variety of time scales may have the merit of enabling flexible selection of a relevant time scale appropriate for the task and the environment at the time of decision making.

A possible mechanism for selection or weighting of different cortico-basal ganglia loops with an appropriate time scale is serotonergic projection from the dorsal raphe nucleus<sup>7</sup> (see Supplementary Fig. 2), which was activated in the LONG versus SHORT contrast. Although serotonergic projection is supposed to be diffuse and global, differential expression of serotonergic receptors in the cortical areas and in the ventral and dorsal striatum<sup>43,44</sup> would result in differential modulation. The mPFC, which had significant correlation with reward prediction  $V(t)$  at long time scales ( $\gamma \geq 0.6$ ), may regulate the activity of the raphe nucleus through reciprocal connection<sup>45,46</sup>. This interpretation is consistent with previous studies using tasks that require long-range prospects for problem solving, such as the gambling problem<sup>1</sup> or delayed reward task<sup>2</sup>, which showed involvement of the medial OFC. Future studies using the Markov decision task under pharmacological manipulation of the serotonergic system should clarify the role of serotonin in regulating the time scale of reward prediction.

Recent brain imaging and neural recording studies report involvement of a variety of cortical areas and the striatum in reward processing<sup>8–11,16,21,23–29,32,33,47–49</sup>. Although some neural recording studies have used experimental tasks that require multiple trial steps for getting rewards<sup>47,48</sup>, none of the previous functional brain imaging studies addressed the issue of reward prediction at different time scales, and considered only rewards immediately following stimuli or actions. We were able to extract specific functions of OFC, DLPFC, mPFC, insula and cortico-basal ganglia loops using our new Markov decision task and a reinforcement learning model-based regression analysis. Our regression analysis not only extracted brain activities specific to reward prediction, but also revealed a topographic organization in reward prediction (Fig. 6). The combination of our Markov decision task with event-related fMRI and magnetoencephalography (MEG) should further clarify the functions used for reward prediction and perception at different time scales, and at finer spatial and temporal resolutions.

## METHODS

**Subjects.** Twenty healthy, right-handed volunteers (18 males and 2 females, ages 22–34 years) gave informed consent to participate in the experiment, which was conducted with the approval of the ethics and safety committees of Advanced Telecommunication Research Institute International (ATR) and Hiroshima University.

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

Subjects completed 4 trials in a NO condition block, 15 trials in a SHORT condition block, 4 trials in a RANDOM condition block and 15 trials in a LONG condition block. A set of four condition blocks (NO, SHORT, RANDOM, LONG) was repeated four times (Fig. 2a). Subjects were informed of the current condition at the beginning of each condition block by text on the screen (first slide in Fig. 1a); thus, the entire experiment consisted of 168 steps (152 trial steps and 16 instruction steps), taking about 17 min. The mappings of the three states to the three figures, and the two buttons to the two actions, were randomly set at the beginning of each experiment, so that subjects were required to learn the amount of reward associated with each figure-button pair in both SHORT and LONG conditions. Furthermore, in the LONG condition, subjects had to learn the subsequent figure for each figure-action pair and take into account the amount of reward expected from the subsequent figure in selecting a button.

**fMRI imaging.** A 1.5-tesla scanner (Shimadzu-Marconi, Magnex Eclipse) was used to acquire both structural T1-weighted images (repetition time, TR = 12 ms, TE = 4.5 ms, flip angle =  $20^\circ$ , matrix =  $256 \times 256$ , FoV = 256 mm, thickness = 1 mm, slice gap = 0 mm) and T2\*-weighted echo planar images (TR = 6 s, TE = 55 ms, flip angle =  $90^\circ$ , 50 transverse slices, matrix =  $64 \times 64$ , FoV = 192 mm, thickness = 3 mm, slice gap = 0 mm) showing blood oxygen level-dependent (BOLD) contrasts.

Because the aim of the present study was to identify brain activity underlying reward prediction over multiple trial steps, we acquired functional images every 6 s (TR = 6 s), in synchrony with single trials. Although shorter TRs and event-related designs are often used in experiments that aim to distinguish brain responses to events within a trial<sup>9,11,21,26</sup>, analysis of those finer events in time were not the focus of the current study. With this longer TR, the BOLD signal in a single scan contained a mixture of responses for a reward-predictive stimulus and reward feedback. However, because of the progress of learning and the stochastic nature of the amount of reward, the time courses of reward prediction  $V(t)$  and prediction error  $\delta(t)$  over the 168 trial steps were markedly different. Thus, we could separate activity corresponding to reward prediction from that corresponding to outcomes by using both reward prediction  $V(t)$  and reward outcome  $r(t)$  in multiple regression analysis, as described below.

**Data analysis.** The data were pre-processed and analyzed with SPM99 ([www.fil.ion.ucl.ac.uk/spm/spm99.html](http://www.fil.ion.ucl.ac.uk/spm/spm99.html)). The first two volumes of images were discarded to avoid T1 equilibrium effects. The images were realigned to the first image as a reference, spatially normalized with respect to the Montreal Neurological Institute EPI template, and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum).

We conducted two types of analysis. One was block-design analysis using four boxcar regressors covering the whole experiment, convolved with a hemodynamic response function as the reference waveform for each condition (NO, SHORT, RANDOM, LONG). We did not find substantial differences between SHORT versus NO and SHORT versus RANDOM contrasts, or between LONG versus NO and LONG versus RANDOM contrasts. Therefore we report here only the results with the NO condition as the control condition. The other method was multivariate regression analysis using explanatory vari-

ables, representing the time course of the reward prediction  $V(t)$  or reward prediction error  $\delta(t)$  at six different timescales  $\gamma$ , estimated from subjects' performance data (described below).

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

**Procedures of performance-based regression analysis.** The time courses of reward prediction  $V(t)$  and reward prediction error  $\delta(t)$  were estimated from each subject's performance data—state  $s(t)$ , action  $a(t)$  and reward  $r(t)$ —as described below.

**Reward prediction.** To estimate how much of a forthcoming reward a subject would have expected at each step during the Markov decision task, we took the definition of the value function (equation 1) and reformulated it based on the recursive structure of the task. Namely, if the subject starts from a state  $s(t)$  and comes back to the same state after  $k$  steps, the expected cumulative reward  $V(t)$  should satisfy the consistency condition  $V(t) = r(t+1) + \gamma r(t+2) + \dots + \gamma^{k-1}r(t+k) + \gamma^k V(t)$ .

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

$$\hat{V}(t) = \frac{r(t+1) + \gamma r(t+2) + \dots + \gamma^{k-1}r(t+k)}{1 - \gamma^k} \quad (1)$$

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

$$V(t) = \frac{1}{L} \sum_{i=1}^L \hat{V}(t_i) \quad (2)$$

where  $\{t_1, \dots, t_L\}$  are the indices of time visiting the same state as  $s(t)$ , that is,  $s(t_1) = \dots = s(t_L) = s(t)$ .

**Reward prediction error.** The TD error (equation 2) was calculated from the difference between the actual reward  $r(t)$  and the temporal difference of the estimated value function  $V(t)$ .

We separately calculated the time courses of  $V(t)$  and  $\delta(t)$  during SHORT and LONG conditions; we concatenated data of four blocks in the SHORT condition, and calculated  $V(t)$  and  $\delta(t)$  as described above. We used the same process for the LONG condition data. During the NO and RANDOM conditions, the values of  $V(t)$  and  $\delta(t)$  were fixed at zero. Finally, we reconstructed the data corresponding to the real time course of the experiment. Examples of the time course of these variables are shown in **Supplementary Figure 1** online. We used either  $V(t)$  or  $\delta(t)$  as the explanatory variable in a regression analysis by SPM. To remove any effects of factors other than reward prediction, we concurrently used other variables in the regression, namely the four box-car functions representing each condition (NO, SHORT, RANDOM, LONG). Because the immediate reward prediction  $V(t)$  with  $\gamma = 0$  can coincide with reward outcome  $r(t)$  if learning is perfect, we included the reward outcome  $r(t)$  in regression analyses with  $V(t)$ . Thus, the significant correlation with  $V(t)$  (Fig. 6a,b) should represent a predictive component rather than a reward outcome.

The amplitude of explanatory variables  $\delta(t)$  with all  $\gamma$  were large in early trials and decreased as subjects learned the task (**Supplementary Fig. 1** online). This decreasing trend causes a risk that areas that are activated early in trials, such as those responsible for general attentiveness or novelty, have correlations with  $\delta(t)$ . Because our aim in regression analysis was to clarify the brain structures involved in reward prediction at specific time scales, we removed the areas that had similar correlation to  $\delta(t)$  at all settings of  $\gamma$  from considerations in **Figure 6** and **Supplementary Table 3** online. To compare the results of regression analysis



with six different values of  $\gamma$ , we used display software that can overlay multiple activation maps in different colors on a single brain structure image. When a voxel is significantly activated in multiple values of  $\gamma$ , it is shown by a mosaic of multiple colors, with apparent subdivision of the voxel (Fig. 6).

Note: Supplementary information is available on the Nature Neuroscience website.

#### ACKNOWLEDGMENTS

We thank K. Samejima, N. Schweighofer, M. Haruno, H. Imamizu, S. Higuchi, T. Yoshioka, T. Chaminade and M. Kawato for helpful discussions and technical advice. This research was funded by 'Creating the Brain,' Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency.

#### COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 5 March; accepted 2 June 2004

Published online at <http://www.nature.com/natureneuroscience/>

- Bechara, A., Damasio, H. & Damasio, A.R. Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* **10**, 295–307 (2000).
- Mobini, S. *et al.* Effects of lesions of the orbitofrontal cortex on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology (Berl.)* **160**, 290–298 (2002).
- Cardinal, R.N., Pennicott, D.R., Sugathapala, C.L., Robbins, T.W. & Everitt, B.J. Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science* **292**, 2499–2501 (2001).
- Rogers, R.D. *et al.* Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms. *Neuropsychopharmacology* **20**, 322–339 (1999).
- Evenden, J.L. & Ryan, C.N. The pharmacology of impulsive behaviour in rats: the effects of drugs on response choice with varying delays of reinforcement. *Psychopharmacology (Berl.)* **128**, 161–170 (1996).
- Mobini, S., Chiang, T.J., Ho, M.Y., Bradshaw, C.M. & Szabadi, E. Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology (Berl.)* **152**, 390–397 (2000).
- Doya, K. Metalearning and neuromodulation. *Neural Net.* **15**, 495–506 (2002).
- Berns, G.S., McClure, S.M., Pagnoni, G. & Montague, P.R. Predictability modulates human brain response to reward. *J. Neurosci.* **21**, 2793–2798 (2001).
- Breiter, H.C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**, 619–639 (2001).
- O'Doherty, J.P., Deichmann, R., Critchley, H.D. & Dolan, R.J. Neural responses during anticipation of a primary taste reward. *Neuron* **33**, 815–826 (2002).
- O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
- Sutton, R.S. & Barto, A.G. *Reinforcement Learning* (MIT Press, Cambridge, Massachusetts, 1998).
- Houk, J.C., Adams, J.L. & Barto, A.G. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J.C., Davis, J.L. & Beiser, D.G.) 249–270 (MIT Press, Cambridge, Massachusetts, 1995).
- Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
- Doya, K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr. Opin. Neurobiol.* **10**, 732–739 (2000).
- McClure, S.M., Berns, G.S. & Montague, P.R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
- Mesulam, M.M. & Mufson, E.J. Insula of the old world monkey. III: Efferent cortical output and comments on function. *J. Comp. Neurol.* **212**, 38–52 (1982).
- Cavada, C., Company, T., Tejedor, J., Cruz-Rizzolo, R.J. & Reinoso-Suarez, F. The anatomical connections of the macaque monkey orbitofrontal cortex. *Cereb. Cortex* **10**, 220–242 (2000).
- Chikama, M., McFarland, N.R., Amaral, D.G. & Haber, S.N. Insular cortical projections to functional regions of the striatum correlate with cortical cytoarchitectonic organization in the primate. *J. Neurosci.* **17**, 9686–9705 (1997).
- Balleine, B.W. & Dickinson, A. The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. *J. Neurosci.* **20**, 8954–8964 (2000).
- Knutson, B., Fong, G.W., Bennett, S.M., Adams, C.M. & Hommer, D. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *Neuroimage* **18**, 263–272 (2003).
- Ullsperger, M. & von Cramon, D.Y. Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *J. Neurosci.* **23**, 4308–4314 (2003).
- O'Doherty, J., Critchley, H., Deichmann, R. & Dolan, R.J. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J. Neurosci.* **23**, 7931–7939 (2003).
- Koepp, M.J. *et al.* Evidence for striatal dopamine release during a video game. *Nature* **393**, 266–268 (1998).
- Elliott, R., Friston, K.J. & Dolan, R.J. Dissociable neural responses in human reward systems. *J. Neurosci.* **20**, 6159–6165 (2000).
- Knutson, B., Adams, C.M., Fong, G.W. & Hommer, D. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**, RC159 (2001).
- Pagnoni, G., Zink, C.F., Montague, P.R. & Berns, G.S. Activity in human ventral striatum locked to errors of reward prediction. *Nat. Neurosci.* **5**, 97–98 (2002).
- Elliott, R., Newman, J.L., Longe, O.A. & Deakin, J.F. Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. *J. Neurosci.* **23**, 303–307 (2003).
- Haruno, M. *et al.* A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.* **24**, 1660–1665 (2004).
- Reynolds, J.N. & Wickens, J.R. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Net.* **15**, 507–521 (2002).
- Tremblay, L. & Schultz, W. Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *J. Neurophysiol.* **83**, 1864–1876 (2000).
- Critchley, H.D., Mathias, C.J. & Dolan, R.J. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29**, 537–545 (2001).
- Rogers, R.D. *et al.* Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *J. Neurosci.* **19**, 9029–9038 (1999).
- Rolls, E.T. The orbitofrontal cortex and reward. *Cereb. Cortex* **10**, 284–294 (2000).
- Hanakawa, T. *et al.* The role of rostral Brodmann area 6 in mental-operation tasks: an integrative neuroimaging approach. *Cereb. Cortex* **12**, 1157–1170 (2002).
- Owen, A.M., Doyon, J., Petrides, M. & Evans, A.C. Planning and spatial working memory: a positron emission tomography study in humans. *Eur. J. Neurosci.* **8**, 353–364 (1996).
- Baker, S.C. *et al.* Neural systems engaged by planning: a PET study of the Tower of London task. *Neuropsychologia* **34**, 515–526 (1996).
- Middleton, F.A. & Strick, P.L. Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Res. Brain Res. Rev.* **31**, 236–250 (2000).
- Haber, S.N., Kunishio, K., Mizobuchi, M. & Lynd-Balta, E. The orbital and medial prefrontal circuit through the primate basal ganglia. *J. Neurosci.* **15**, 4851–4867 (1995).
- Eagle, D.M., Humby, T., Dunnett, S.B. & Robbins, T.W. Effects of regional striatal lesions on motor, motivational, and executive aspects of progressive-ratio performance in rats. *Behav. Neurosci.* **113**, 718–731 (1999).
- Pears, A., Parkinson, J.A., Hopewell, L., Everitt, B.J. & Roberts, A.C. Lesions of the orbitofrontal but not medial prefrontal cortex disrupt conditioned reinforcement in primates. *J. Neurosci.* **23**, 11189–11201 (2003).
- Hikosaka, O. *et al.* Parallel neural networks for learning sequential procedures. *Trends Neurosci.* **22**, 464–471 (1999).
- Mijnster, M.J. *et al.* Regional and cellular distribution of serotonin 5-hydroxytryptamine<sub>2A</sub> receptor mRNA in the nucleus accumbens, olfactory tubercle, and caudate putamen of the rat. *J. Comp. Neurol.* **389**, 1–11 (1997).
- Compan, V., Segu, L., Buhot, M.C. & Daszuta, A. Selective increases in serotonin 5-HT<sub>1B/1D</sub> and 5-HT<sub>2A/2C</sub> binding sites in adult rat basal ganglia following lesions of serotonergic neurons. *Brain Res.* **793**, 103–111 (1998).
- Celada, P., Puig, M.V., Casanovas, J.M., Guillazo, G. & Artigas, F. Control of dorsal raphe serotonergic neurons by the medial prefrontal cortex: involvement of serotonin-1A, GABA(A), and glutamate receptors. *J. Neurosci.* **21**, 9917–9929 (2001).
- Martin-Ruiz, R. *et al.* Control of serotonergic function in medial prefrontal cortex by serotonin-2A receptors through a glutamate-dependent mechanism. *J. Neurosci.* **21**, 9856–9866 (2001).
- Hikosaka, K. & Watanabe, M. Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cereb. Cortex* **10**, 263–271 (2000).
- Shidara, M. & Richmond, B.J. Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science* **296**, 1709–1711 (2002).
- Matsumoto, K., Suzuki, W. & Tanaka, K. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* **301**, 229–232 (2003).