Behavioral/Systems/Cognitive

Temporal Filtering of Reward Signals in the Dorsal Anterior Cingulate Cortex during a Mixed-Strategy Game

Hyojung Seo and Daeyeol Lee
Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut 06510

The process of decision making in humans and other animals is adaptive and can be tuned through experience so as to optimize the outcomes of their choices in a dynamic environment. Previous studies have demonstrated that the anterior cingulate cortex plays an important role in updating the animal's behavioral strategies when the action outcome contingencies change. Moreover, neurons in the anterior cingulate cortex often encode the signals related to expected or actual reward. We investigated whether reward-related activity in the anterior cingulate cortex is affected by the animal's previous reward history. This was tested in rhesus monkeys trained to make binary choices in a computer-simulated competitive zero-sum game. The animal's choice behavior was relatively close to the optimal strategy but also revealed small systematic biases that are consistent with the use of a reinforcement learning algorithm. In addition, the activity of neurons in the dorsal anterior cingulate cortex that was related to the reward received by the animal in a given trial often was modulated by the rewards in the previous trials. Some of these neurons encoded the rate of rewards in previous trials, whereas others displayed activity modulations more closely related to the reward prediction errors. In contrast, signals related to the animal's choices were represented only weakly in this cortical area. These results suggest that neurons in the dorsal anterior cingulate cortex might be involved in the subjective evaluation of choice outcomes based on the animal's reward history.

Key words: **reinforcement learning; game theory; neuroeconomics; decision; dopamine; reward**

Introduction

An action or a sequence of actions that maximizes the desirability or utility of the expected outcome is considered optimal, and in most cases this has to be learned empirically. In reinforcement learning the problem of finding an optimal sequence of actions is formulated by using a set of value functions that express the sum of future rewards expected from a particular state or action (Sutton and Barto, 1998). Specific reinforcement learning algorithms then describe how value functions can be adjusted empirically based on the reward prediction error, namely the discrepancy between the reward predicted from the value functions and the actual reward. Once value functions are estimated accurately, optimal decision-making strategies can be found simply by choosing the actions that maximize the value functions.

Reinforcement learning algorithms can account for the choice behavior of humans and animals during various decision-making tasks (Erev and Roth, 1998; Lee et al., 2004, 2005; Sugrue et al., 2004; Lau and Glimcher, 2005; Samejima et al., 2005; Yechiam and Busemeyer, 2005; Daw et al., 2006; Haruno and Kawato, 2006). In addition, neural activity related to the key variables in reinforcement learning algorithms, such as value functions and reward prediction errors, has been identified in many different cortical and subcortical areas (Daw and Doya, 2006; Lee, 2006; Schultz, 2006). However, the neural mechanisms responsible for computing value functions and reward prediction errors are still poorly understood. For example, signals related to the overall reward rate might be useful for evaluating the overall success of the learning process and therefore can guide the process of so-called meta learning (Aston-Jones and Cohen, 2005; Soltani et al., 2006), but it is not known how such signals are computed in the brain.

In the present study we investigated the role of the dorsal anterior cingulate cortex (ACCd) in evaluating the outcomes of the animal's choices during a simulated competitive zero-sum game (Barraclough et al., 2004; Lee et al., 2004). The results from the behavioral analysis suggested that the animals searched for the optimal strategy during this task, using a reinforcement learning algorithm. We also found that many neurons in the ACCd displayed modulations in their activity related to the rewards in previous trials. These signals displayed diverse time courses and therefore might provide the signals necessary to compute various aspects of rewards resulting from previous actions of the animal. In particular, the reward-related signals in the ACCd might be useful for computing the overall reward rate and how much the actual reward in a given trial deviates from the animal's expectation. These signals then might influence the process of updating the value functions and hence the animal's decision-making strategy.

Received Feb. 18, 2007; revised June 23, 2007; accepted June 25, 2007.

This work was supported by National Institutes of Health Grant MH073246. We are grateful to D. Barraclough and B. McGreevy for their help with this experiment, L. Carr and J. Swan-stone for their technical assistance, and M. Jung for his comments on this manuscript.

Correspondence should be addressed to Dr. Daeyeol Lee, Department of Neurobiology, Yale University School of Medicine, 333 Cedar Street, SHM B404, New Haven, CT 06510. E-mail: daeyeol.lee@yale.edu.

DOI:10.1523/JNEUROSCI.2369-07.2007

 0270-6474/07/278366-12$15.00/0

Materials and Methods

Animal preparations

Two male rhesus monkeys (D and E; body weight, 8–12 kg) were used. Their eye movements were monitored at a sampling rate of 250 Hz with a high-speed eye tracker (ET49; Thomas Recording, Giessen, Germany).

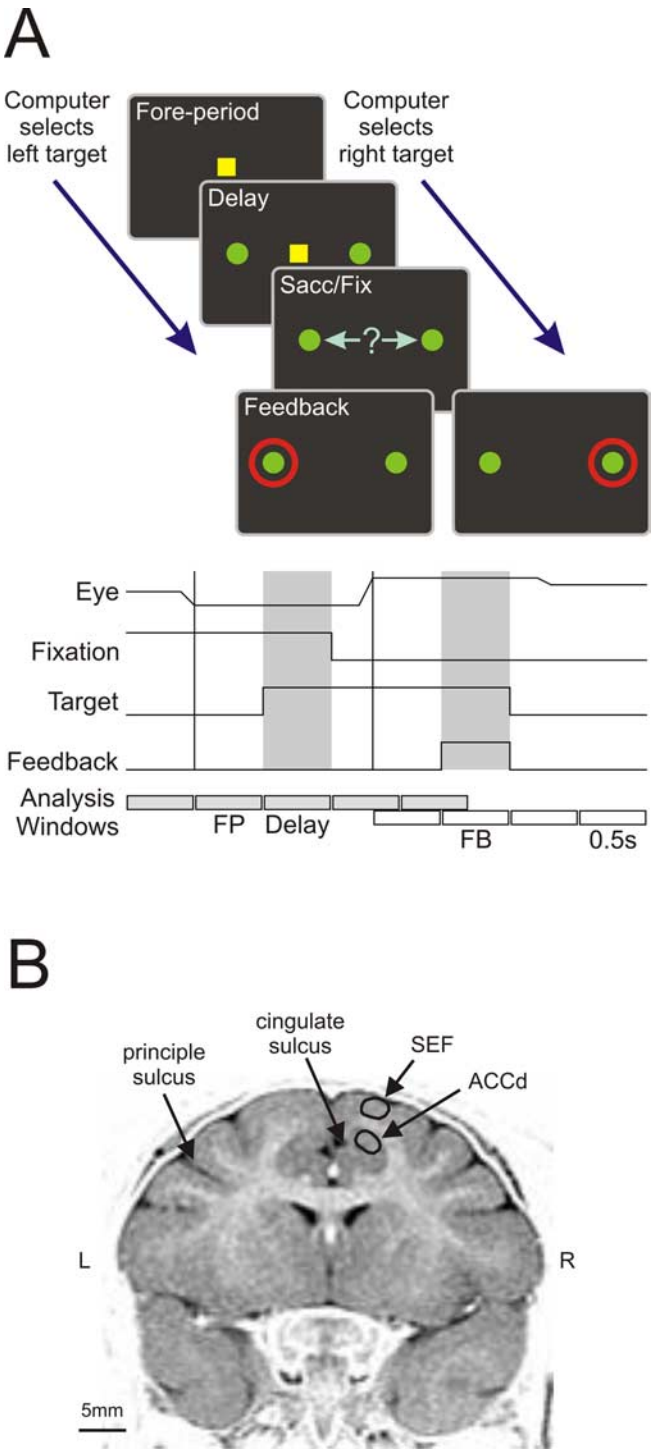

Figure 1. Task and recording sites. ***A***, Oculomotor free choice task. Gray and white rectangles at the bottom correspond to a series of 0.5 s windows used to analyze the neural data that are aligned to the target onset and feedback onset, respectively. FP, Fore period; FB, feedback period. ***B***, An MR image (coronal; anteroposterior, 31 mm; monkey D) showing the recording sites in the ACCd.

All of the procedures used in this study were approved by the University of Rochester Committee on Animal Research and conformed to the *Public Health Service Policy on Humane Care and Use of Laboratory Animals* and the *Guide for the Care and Use of Laboratory Animals*.

Behavioral task

Monkeys were trained to perform an oculomotor free-choice task modeled after a two-player zero-sum game, known as the matching pennies (Fig. 1*A*) (Barraclough et al., 2004; Lee et al., 2004). During a 0.5 s fore period they fixated on a small yellow square [0.9 × 0.9°; Commission Internationale de l'Eclairage (CIE): $x = 0.432$; $y = 0.494$; $Y = 62.9$ cd/m^2] in the center of a computer screen, and then two identical green disks (radius, 0.6°; CIE: $x = 0.286$; $y = 0.606$; $Y = 43.2$ cd/m^2) were presented 5° away in diametrically opposed locations along the horizontal meridian. The central target was extinguished after a 0.5 s delay period, and the animal was required to shift its gaze to one of the targets within 1 s. After the animal maintained the fixation on its chosen peripheral target for 0.5 s, a red ring (radius, 1°; CIE: $x = 0.632$; $y = 0.341$; $Y =$ 17.6 cd/m^2) appeared around the target selected by the computer. If the animal selected the same target as the computer opponent, it was rewarded after maintaining its fixation of the chosen target for 0.5 s from the onset of the red ring. If the animal's choice was different from that of the computer opponent, it was no longer required to maintain the fixation of the chosen target. The central fixation target for the new trial was presented 1 s from the end of the previous trial.

The computer was programmed to exploit statistical biases in the animal's choice behavior. Before each trial the computer made a prediction for the animal's choice by applying the following set of statistical tests to the animal's entire choice and reward history during a given recording session [corresponding to algorithm 2 in Lee et al. (2004)]. First, the conditional probabilities for the animal to choose each target given its choices in the preceding n trials ($n = 0–4$) were estimated (e.g., for $n =$ 1, the probability that the animal would choose the rightward target given that it chose the left target in the previous trial). Second, the conditional probabilities for the animal to choose each target given its choices and rewards in the preceding n trials ($n = 1–4$) also were estimated (e.g., for $n = 1$, the probability that the animal would choose the rightward target given that the animal was rewarded for choosing the leftward target in the previous trial). Next, each of these nine conditional probabilities was tested against the hypothesis that the animal had chosen both targets with equal probabilities. When none of these hypotheses was rejected (binomial test; $p < 0.05$), the computer selected each target randomly with 50% probability. Otherwise, the computer biased its selection according to the conditional probability with the largest deviation from 0.5 that was statistically significant. For example, if the animal chose the rightward target with 80% probability according to this criterion, the computer selected the leftward target with the same probability. To maximize the total reward, therefore, the animal needed to choose both targets equally often and make its choice independently from previous choices and their outcomes. As reported previously (Lee et al., 2004), the choice behavior of monkeys during this matching pennies task was highly stochastic. This is beneficial for the regression analysis of neural data in which the effects of the animal's choices in multiple trials are evaluated simultaneously (see below).

Neurophysiological recording

Single-unit activity was recorded from the neurons in the dorsal anterior cingulate cortex of two monkeys (monkeys D and E), using a five-channel multi-electrode recording system (Thomas Recording). The placement of the recording chamber was guided by magnetic resonance (MR) images (Fig. 1*B*), and this was confirmed by metal pins inserted in known anatomical locations at the end of the experiment. In both animals the supplementary eye field (SEF) was localized, based on eye movements evoked by electrical stimulations with currents <50 μA (100 μA for some sites) during active fixation of a visual target (Goldberg et al., 1986). All neurons were recorded in the dorsal bank of the cingulate sulcus (area 24c) ventral to the SEF (Matelli et al., 1991; Luppino et al., 2003).

Reinforcement learning model

We applied a reinforcement learning model (Sutton and Barto, 1998) to analyze how the animal's choice was influenced by the outcomes of its previous choices in the matching pennies task. In this model the value function for choosing target x_t (R or L for rightward and leftward choices, respectively) in trial t, $Q_t(x_t)$, was updated according to the reward prediction error, as follows:

$$Q_{t+1}(x_t) = Q_t(x_t) + \alpha[r_t - Q_t(x_t)], \quad (1)$$

where r_t denotes the reward received by the animal in trial t (0 and 1 for unrewarded and rewarded trials, respectively), and α denotes the learning rate. The reward prediction error, $[r_t - Q_t(x_t)]$, corresponds to the discrepancy between the actual reward and the expected reward. The probability that the animal would choose the rightward target in trial t, $P_t(R)$, was determined by the SoftMax transformation as follows:

$$P_t(R) = \exp\beta Q_t(R)/\{\exp\beta Q_t(L) + \exp\beta Q_r(R)\}, \quad (2)$$

where β, referred to as the inverse temperature, determines the randomness of the animal's choices. Model parameters were estimated separately for each recording session by using a maximum likelihood procedure (Pawitan, 2001; Lee et al., 2004). For each session the best parameters were selected from 100 independent searches performed with the initial parameters randomly chosen in the interval of [0 1]. The maximum likelihood procedure was implemented by using the fminsearch function in MatLab 7.0 (MathWorks, Natick, MA). For >90% of the sessions the parameters with the maximum likelihood were found consistently in at least 10% of the searches.

As described in Results, the animals displayed a significant tendency to make their choices according to the so-called win–stay–lose–switch strategy (Lee et al., 2004). In other words, they were more likely to choose the same target rewarded in the previous trial and switched to the other target otherwise. If the animal selected its targets based on a fixed probability of adopting the win–stay–lose–switch strategy, p_{WSLS}, the likelihood that the animal would choose the target x would be $P_t(x) = p_{WSLS}$, when the animal was rewarded for choosing x or unrewarded for choosing the other target in the previous trial, and $P_t(x) = (1 - p_{WSLS})$ otherwise.

If we denote the animal's choice in trial t as c_t (= R or L), the likelihood for the animal's choices in a given session is given by the following:

$$L = \Pi_t P_t(c_t) = P_1(c_1) P_2(c_2) \cdots P_N(c_N), \quad (3)$$

where N denotes the number of trials. Whether the animal's choice behavior in a given session was better accounted for by the reinforcement learning model or by the win–stay–lose–switch strategy was determined by the Bayesian information criterion (BIC) (Burnham and Anderson, 2002):

$$\mathrm{BIC} = -\log L + k\log N, \quad (4)$$

where k is the number of model parameters (1 for the win–stay–lose–switch strategy model and 2 for the reinforcement learning model).

Analysis of neural data

Time course of overall activity. To examine how the overall excitability of neurons in the ACCd changed throughout the course of a trial, we calculated the rate of spikes during the three 0.5 s windows corresponding to the delay period, the fixation period for the chosen target, and the feedback period in each trial. In addition, the rate of spikes during the 0.2 s window before the onset of saccade directed to the chosen target also was calculated. It then was determined whether the activity in each of these epochs significantly increased or decreased as compared with the activity during the 0.5 s fore period (paired Student's t test; $p < 0.05$).

Analysis of activity related to reward and choice. For each neuron we used a Student's t test to test whether its activity during the 0.5 s feedback period differed for the rewarded and unrewarded trials. Animals often broke their fixation during the feedback period in unrewarded trials because this was not penalized. Therefore, to determine whether neural activity related to eye movements was mistaken for reward-related activity, we applied the following regression model to test the effects of reward and eye movements separately:

$$y_t = a_0 + a_1 r_t + a_2 s_t, \quad (5)$$

where r_t is a dummy variable indicating whether the animal is rewarded in trial t or not, and s_t is the saccade latency in trial t defined as the time between the feedback onset and saccade onset. Another multiple linear regression model then was applied to determine whether the activity of a given neuron was influenced by the animal's choices, the computer opponent's choices, and the animal's rewards in the current and previous trials. This analysis was applied separately to the number of spikes counted for a series of nine non-overlapping 0.5 s bins defined relative to the time of target onset or feedback onset, including five time bins beginning 1 s before the target onset (i.e., 0.5 s before the fixation of the central target) (Fig. 1*A*, bottom, gray horizontal bars) and four time bins starting from the fixation of a peripheral target (i.e., 0.5 s before feedback onset) (Fig. 1*A*, bottom, white horizontal bars). For the spike counts in a particular bin of trial t, y_t, this corresponded to the following:

$$y_t = \mathbf{B}[1\ u_t\ u_{t-1}\ u_{t-2}\ u_{t-3}]' + e_t, \quad (6)$$

where u_t is a row vector consisting of three binary variables corresponding to the animal's choice (0 and 1 for leftward and rightward choices, respectively), the computer choice (coded in the same way as the animal's choice), and the reward (0 and 1 for unrewarded and rewarded trials, respectively) in trial t, $\mathbf{B}$ is a vector of 13 regression coefficients, and e_t is the error term. The statistical significance for each regression coefficient was determined with a Student's t test ($p < 0.05$). As shown in Results, the activity of ACCd neurons was influenced more strongly by the animal's reward history than by its choice history or the previous choices of the computer opponent. In addition, the time course of reward-related signals identified by the regression analysis varied substantially across different neurons. To examine the heterogeneity in the time course of reward-related signals, we applied the k-means cluster analysis to the normalized regression coefficients related to reward (Hastie et al., 2001; Lee et al., 2001). The purpose of this analysis was not to demonstrate that the neurons in ACCd form distinct clusters based on the time course of reward-related signals but to determine how the time course of such signals varied across different neurons. We performed this analysis with the number of clusters (or centroids) ranging from 2 to 10. We also tested whether the activity of neurons in the ACCd was influenced by the interaction among the rewards in successive trials by applying a three-way ANOVA, with the rewards in the current trial and the two previous trials as main factors. For the purpose of illustrations we also estimated the spike density functions by using a Gaussian filter ($\sigma = 50$ ms) separately for various subsets of trials as necessary (see Fig. 4). Average values are shown in Results as the mean ± SEM.

Analysis of activity related to value functions and reward prediction errors. We tested whether the activity of ACCd neurons encoded signals related to the value functions and the reward prediction error by applying the following regression model:

$$y_t = d_0 + d_1\{Q_t(R) + Q_t(L)\} + d_2\{Q_t(R) - Q_t(L)\} + d_3\{r_t - Q_t(c_t)\}, \quad (7)$$

where $Q_t(x)$ denotes the value function for the target x, r_t the reward in trial t, c_t the animal's choice in trial t, and $d_0 \sim d_3$ the regression coefficients. We chose to include the sum of the value functions for the two alternative targets and their difference in this regression model rather than the two value functions separately. As described in Results, this was based on the observation that many more neurons in ACCd encoded signals related to the reward in previous trials that would affect the value functions of both targets indiscriminately. We also hypothesized that the neurons might encode the sum of the value functions by sustaining the signals related to the rewards received by the animal across multiple trials. To test this hypothesis, we calculated the correlation coefficient between the regression coefficient for the sum of value function, d_1, and those related to the reward-related terms in the regression model (Eq. 6). Similarly, ACCd neurons might contribute to the computation of reward prediction errors by encoding the difference between the signals related to the reward in a given trial and those related to the rewards in previous trials. This was tested by calculating the correlation coefficient between the regression coefficient related to the reward prediction error, d_3, in the regression model (Eq. 7) and the reward-related terms in the regression model (Eq. 6).

Results

Reinforcement learning and choice behavior

Behavioral data were collected from a total of 53,308 trials (26,579 and 26,729 trials for monkeys D and E, respectively) in 77

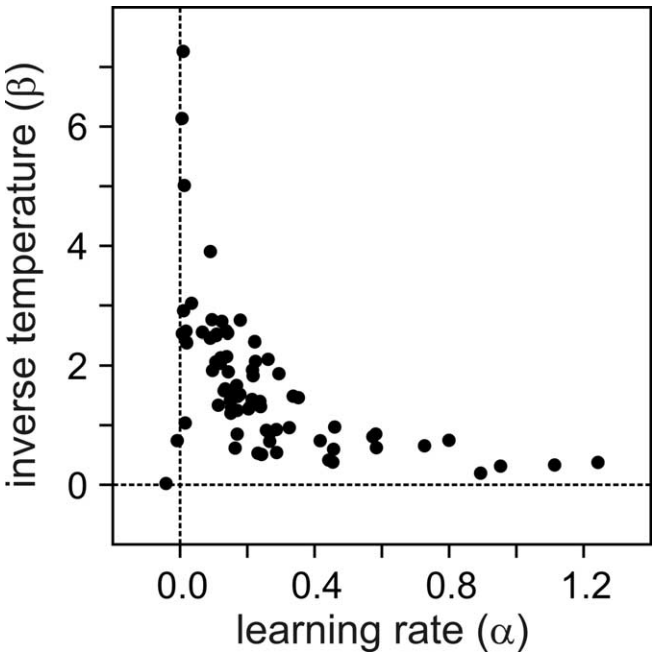

Figure 2. Learning rate and inverse temperature of reinforcement learning model applied to the choice behavior in the matching pennies task. The results from the three sessions in which the inverse temperature was extremely large ($\beta > 1000$) are not shown.

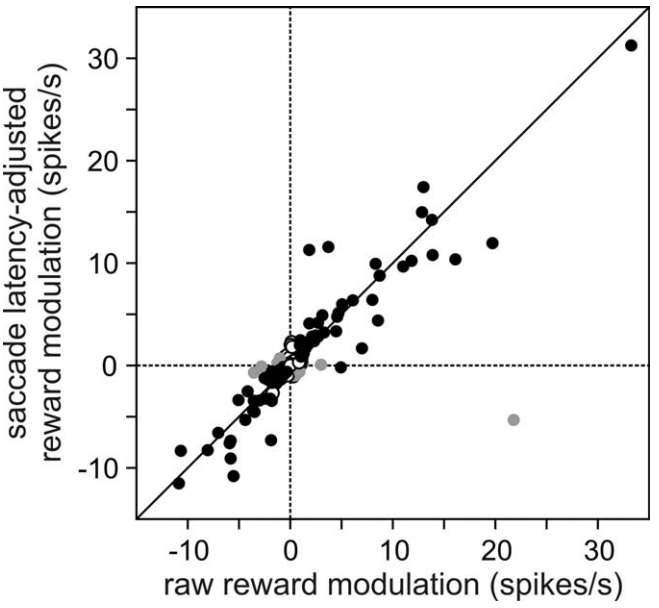

Figure 3. Relationship between the difference in spike rates during the feedback period of rewarded and unrewarded trials (abscissa) and its estimate obtained from a regression model that controlled for the variability in saccade latency (ordinate). Black symbols correspond to the neurons with significant effects of reward in a Student's t test ($p < 0.05$). Gray symbols correspond to the neurons ($n = 12$) for which the activity was significantly related to the saccade latency ($p < 0.05$), but not to the reward in the regression model.

recording sessions (40 and 37 sessions for monkeys D and E, respectively) while the animals performed the matching pennies task. During this task the animal was required to choose the two alternative targets with equal probabilities and independently across successive trials to maximize the overall reward. The animal's performance was relatively close to this optimal strategy. For example, the probability that the animal selected the rightward target was 0.504 ± 0.003, and this was not significantly different from 0.5 (Student's t test; $p = 0.16$). In addition, the hypothesis that the animal chose the two targets with equal probabilities was rejected in only 15.6% of the sessions. Nevertheless, there was a small but systematic bias, as reflected in the overall reward rate (0.494 ± 0.003) that was significantly lower than 0.5 ($p < 0.05$). For example, the animal displayed a small but statistically significant tendency to choose the same target chosen by the computer in the previous trial. This so-called win–stay–lose–switch strategy was found in 51.0 ± 0.3% of the trials; this was significantly higher than 50% ($p < 0.05$), and the animal used this strategy significantly more frequently than 50% in 23.4% of the sessions. We also found that the standard reinforcement learning model accounted for the animal's choice behavior better than the win–stay–lose–switch strategy in 72.7% of the sessions (56 of 77). The average BIC for the reinforcement learning model and the model based on the win–stay–lose–switch strategy was 952.5 and 960.7, respectively, and this difference was statistically significant (paired Student's t test; $p < 10^{-6}$). The average learning rate in the reinforcement learning model was relatively small (0.24) (Fig. 2), indicating that the value functions were updated slowly. This suggests that the animal's choice in a given trial was influenced by the choice outcomes in multiple previous trials.

Table 1. Number (percentage) of ACCd neurons that displayed significant changes in their activity during various epochs compared to the activity in the fore period

	Delay	Pre-saccade	Target fixation	Feedback
Decrease	69 (44.8)	66 (42.9)	52 (33.8)	76 (49.4)
Increase	45 (29.2)	45 (29.2)	56 (36.4)	46 (29.9)
Sum	114 (74.0)	111 (72.1)	108 (70.1)	122 (79.2)

Paired Student's t test; $p < 0.05$; $n = 154$ neurons.

Encoding of reward-related signals in the ACCd

Activity was recorded from 154 neurons in the ACCd (Fig. 1*B*) in two rhesus monkeys (77 neurons from each animal). Compared with the activity during the fore period, the majority of neurons displayed significant changes in their activity during the delay period after target onset, before and after eye movements toward the animal's chosen target, and during the feedback period (Table 1). In addition, most of these neurons (126 neurons; 81.8%) modulated their activity during the feedback period according to whether the animal's choice in the same trial would be rewarded or not (two-tailed Student's t test, $p < 0.05$) (Fig. 3). Some neurons increased their activity during the feedback period of rewarded trials, namely immediately after the computer indicated that the animal would be rewarded, as compared with the activity in unrewarded trials (Fig. 4, bottom). Others decreased their activity when the computer indicated that the animal would be rewarded (Fig. 5, bottom). Overall, the neurons that showed significantly higher activity during the feedback period in rewarded trials than in unrewarded trials (60 neurons) were almost as prevalent as those that showed significantly lower activity in rewarded trials (66 neurons). In addition, the mean t value for the reward-related modulation in the neural activity was 0.522, and this was not significantly different from zero (Student's t test; $p = 0.634$).

During the matching pennies task the animals were required to maintain their fixation throughout the 0.5 s feedback period in rewarded trials. However, the animals often broke their fixation during the feedback period in unrewarded trials, because this was not penalized. The average interval between the feedback onset and the saccade onset was 746.1 ± 0.3 and 352.5 ± 0.7 ms for rewarded and unrewarded trials, respectively. To determine whether the difference in neural activity during the feedback pe-

riod between rewarded and unrewarded trials might be attributable to the difference in the eye movements, we determined the effect of reward on neural activity in a regression model that also included the saccade latency. The regression coefficient related to reward in this regression model was highly correlated with the mean difference in activity between rewarded and unrewarded trials ($r = 0.87$) (Fig. 3). The same regression analysis showed that the neural activity was related significantly to the saccade latency in only 27.3% of the neurons (42 of 154). In only 12 of these neurons (7.8%) was the effect of reward not significant in the regression analysis. Therefore, the effect of eye movements on reward-related activity was relatively small.

Many neurons in the ACCd continued to modulate their activity according to the outcome of the animal's choice in a given trial during the intertrial interval that followed and even during the next trial (Figs. 4, 5). We used a multiple linear regression analysis to determine whether and how the reward in a given trial influenced the activity of the ACCd neurons in the trials that followed. This model also included the animal's choice and the choice of the computer opponent (see Materials and Methods). During the 0.5 s period during the intertrial interval immediately before the fore period, 65.6% (101 of 154 neurons) of the neurons in the ACC significantly modulated their activity according to whether the animal was rewarded or not in the previous trial (Fig. 6, bottom). In addition, whether the animal was rewarded or not two trials before the current trial significantly influenced the activity in 26.0% (40 of 154 neurons) of the neurons. Similarly, during the fore period the activity in 54.6% of the neurons was significantly influenced by whether the animal was rewarded or not in the previous trial, whereas the reward two trials before the trial influenced the activity in 16.2% of the neurons. For some neurons the activity also was modulated by whether the animal was rewarded or not three trials before. The percentage of such neurons in the ACCd was 14.9 and 13.6% immediately before and during the fore period, respectively. During the feedback period 81.8, 41.6, and 19.5% of the neurons modulated their activity significantly according to the reward received by the animal in the current trial and the previous two trials, respectively. Similar to the results based on the fraction of neurons, the magnitude of neural signals related to the reward in the previous trials was reduced gradually as compared with the signals related to the reward in the current trial (Fig. 6, bottom, gray symbols). Therefore, the changes in neural activity related to the reception of reward dissipated over the course of two or three trials in the ACCd.

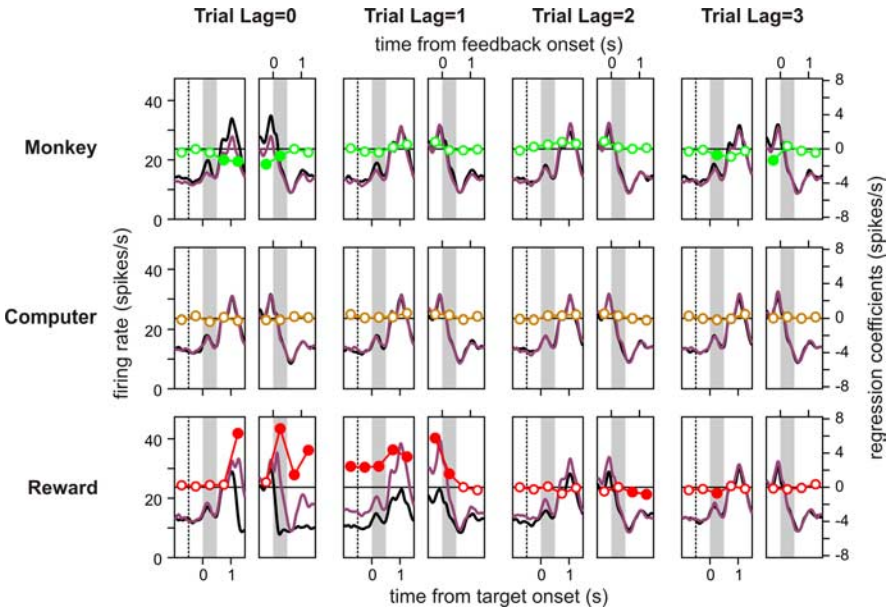

Figure 4. Activity of an example neuron in the ACCd during the matching pennies task. Each pair of small panels displays the spike density functions estimated relative to the time of target onset (left panels) or feedback onset (right panels). They were estimated separately according to the animal's choice (top), the computer choice (middle), or reward (bottom) in the current trial (Trial Lag = 0) or according to the corresponding variables in three previous trials (Trial Lag = 1, 2, or 3). Purple (black) lines correspond to the activity associated with rightward (leftward) choices (top and middle) or rewarded (unrewarded) trials (bottom). Circles show the regression coefficients from a multiple linear regression model, which was performed separately for each time bin. Filled circles indicate the coefficients significantly different from zero (Student's *t* test, $p < 0.05$). The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

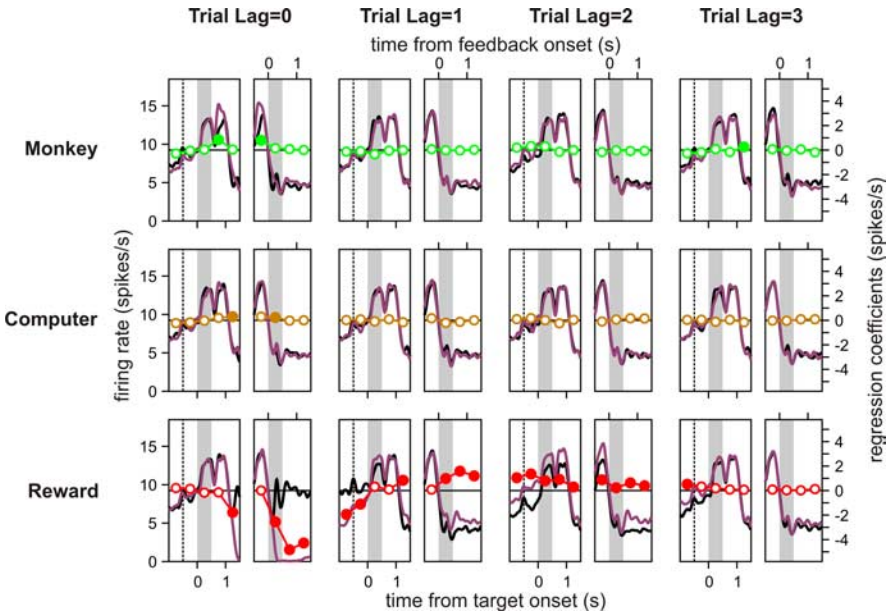

Figure 5. Activity of another example neuron in the ACCd. This is the same format as in Figure 4.

Time course of reward-related signals in the ACCd neurons

The time course of reward-related signals varied substantially across different neurons in the ACCd. For example, the neuron shown in Figure 4 increased its activity during the feedback period in rewarded trials as compared with the activity in unrewarded trials, and this increased activity continued until the feedback period in the next trial (Fig. 4, bottom). In contrast, for some neurons in the ACCd the changes in the activity related to

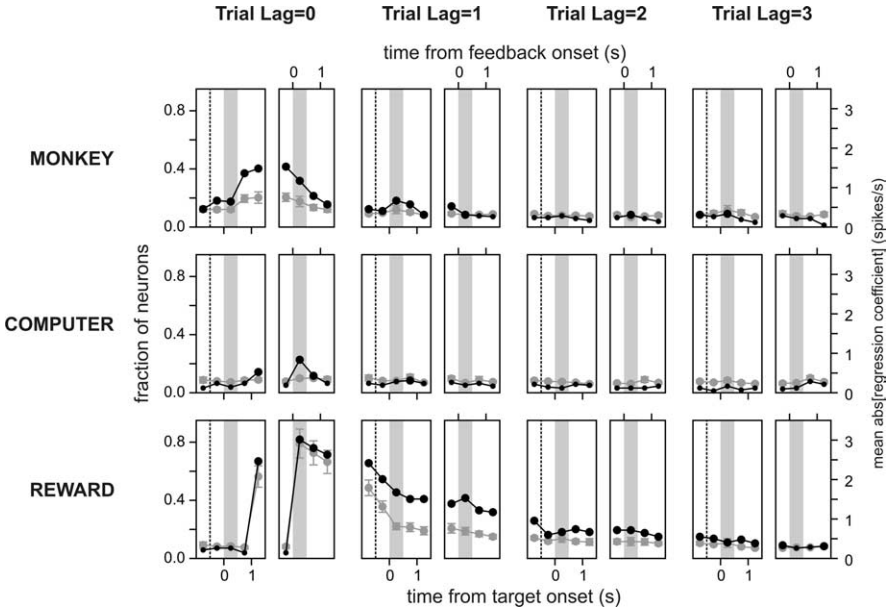

Figure 6. Time course of activity related to the animal's choice (top), the choice of the computer opponent (middle), and reward (bottom) in the population of ACCd neurons. Black symbols (left axis) indicate the percentage of neurons that displayed significant modulations in their activity according to each variable (Student's *t* test; $p < 0.05$). Gray symbols (right axis) indicate the average magnitude of the regression coefficients related to each variable. These values were estimated separately for different time bins, using a series of multiple linear regression models. Large black symbols indicate that the percentage of neurons was significantly higher than the significance level used in the regression analysis (binomial test, $p < 0.05$). The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

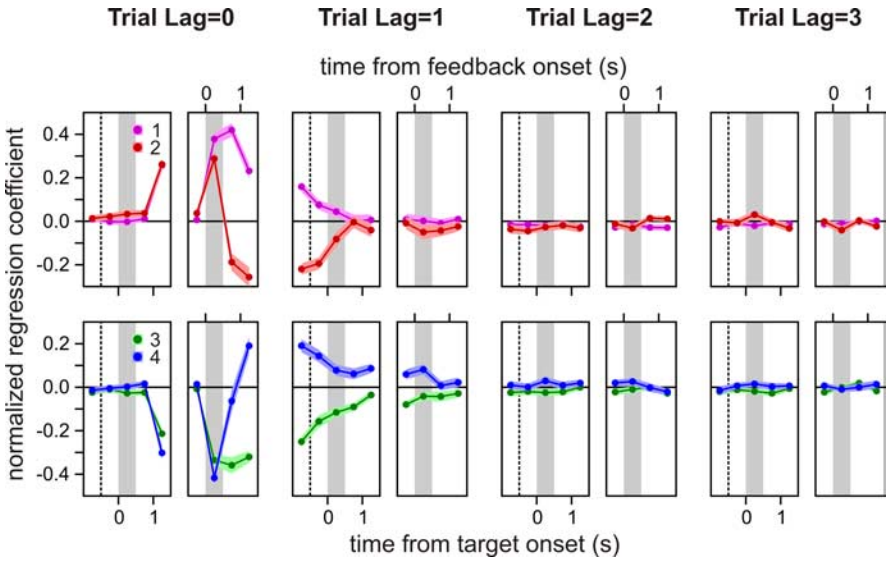

Figure 7. Heterogeneity in the time course of reward-related signals in the ACCd. Normalized regression coefficients were averaged for each of four clusters (centroids) identified with the *k*-means cluster analysis. Shaded region corresponds to the area bounded by the mean ± SEM. The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

the reward in a given trial reversed their polarity during the next several trials. For example, the neuron illustrated in Figure 5 decreased its activity during the feedback period in rewarded trials. The activity of this neuron then remained significantly lower throughout the fore period in the trial that followed if the animal was rewarded in a given trial (Fig. 5, bottom, Trial Lag = 1). However, this neuron tended to increase its activity during the feedback period when the animal was rewarded in the previous trial. There was also a small but significant tendency to increase activity if the animal was rewarded two trials before the current trial (Fig. 5, bottom, Trial Lag = 2). Therefore, compared with when the animal was not rewarded, this neuron responded to the reward-predicting feedback stimulus presented in a given trial (trial t) first by decreasing its activity immediately and then by increasing its activity toward the end of the next trial (trial $t + 1$) and maintaining this increased activity during the subsequent trial (trial $t + 2$).

If the neural activity is increased or decreased consistently by the rewards in two successive trials and if these effects are combined additively, these neurons would encode signals related to the temporal sum of the rewards. In contrast, if the activity is increased by the reward in a given trial but diminished by the reward in the previous trial or vice versa, this activity would be related to the temporal difference of the rewards in the two trials. To determine whether the activity related to the rewards in multiple trials was related more closely to their temporal sum or difference, we examined the regression coefficients related to the reward variables in the regression model. For the spike counts during the feedback period there were 53 neurons in which the regression coefficients for the rewards in the current and previous trials were both significantly different from zero. For 24 of these neurons both regression coefficients had the same sign, whereas their signs were opposite for the remaining 29 neurons. This difference was not significantly larger than expected from the binomial distribution with the equal probabilities ($p = 0.205$). Similarly, there were 27 neurons in which the regression coefficients for the reward in the current trial and the reward two trials before the current trial were both significantly different from zero. For 11 neurons the regression coefficients had the same sign, whereas the signs were different for 16 neurons. Again, this difference was not statistically significant (binomial test; $p = 0.124$). Moreover, the regression coefficients for the reward in the current trial were not significantly correlated with the coefficients for the reward in the previous trial ($r = -0.033$; $p = 0.682$) or two trials before ($r = -145$; $p = 0.073$). Thus, overall, the regression coefficients associated with rewards in successive trials were not correlated strongly, indicating that there was no systematic bias for the neurons in the ACCd to encode the temporal sum or difference of rewards across multiple trials.

We used a series of *k*-means cluster analyses to examine the diversity in the time course of reward-related signals identified with the regression analysis. Although the results varied somewhat according to the number of clusters included in the analysis, two important features were observed consistently. First, neurons were divided into different clusters according to the polarity of reward-related activity during the feedback period. For instance, in the example shown in Figure 7, the neurons in clusters 1 and 2 responded more strongly in rewarded trials than in unrewarded trials, whereas the opposite was true in clusters 3 and 4. Second, for neurons in some clusters, activity increased (cluster 1, $n = 43$ neurons) or decreased (cluster 3, $n = 42$ neurons) (Fig. 7) at the end of rewarded trials, and this differential activity was reduced gradually without reversing its sign. In other clusters, in contrast, activity increased (cluster 2, $n = 33$ neurons) or decreased (cluster 4, $n = 36$ neurons) (Fig. 7) transiently at the end of rewarded trials, but the effect of reward soon was reversed.

Signals related to value functions and reward prediction errors

To investigate whether and how the reward-related signals encoded by the neurons in the ACCd contribute to the computation of value functions and reward prediction errors implicated in the reinforcement learning models, we tested a regression model in which the activity of individual ACCd neurons was given by a linear combination of value functions and reward prediction error (Eq. 7). Consistent with the observation that the activity in ACCd often was modulated consistently by the rewards in previous trials, many neurons significantly modulated their activity according to the sum of the value functions for the two alternative targets. For example, during the fore period and delay period 34.4 and 35.7% of the neurons, respectively, displayed significant modulations in their activity according to the sum of the value functions (Fig. 8, top). In contrast, the fraction of neurons in the ACCd that significantly modulated their activity according to the difference in the value functions was relatively low. During the fore period and delay period, for example, 15.6 and 18.8% of the neurons, respectively, showed such modulations. This indicates that the ACCd might play a relatively minor role in encoding the relative desirability of alternative actions during decision making.

In reinforcement learning the value functions represent how much reward is expected from each of alternative actions, and they are updated based on the animal's experience. During the matching pennies task, therefore, the sum of the value functions indicates the overall rate of reward estimated from the outcomes in previous trials. This implies that the neurons encoding the sum of the value functions might change their activity similarly in response to the rewards in previous trials. For example, the neuron illustrated in Figure 4 increased its activity at the end of the rewarded trial and also when the animal was rewarded in the previous trial (Fig. 4, bottom). Accordingly, the activity of this neuron was significantly correlated with the sum of the value functions throughout the trial as well as the reward prediction error (Fig. 8*B*, top). Overall, the regression coefficients associated with the sum of the value functions during and immediately after the feedback period tend to be significantly correlated with the regression coefficients associated with the reward received by the animal during the previous trials (Fig. 9, top).

Whereas the sum of value functions can be estimated by simply integrating the signals related to the animal's previous rewards, the reward prediction error can be computed by subtracting the signals related to the previous rewards from the signals related to the current reward. Therefore, the activity of neurons encoding the reward prediction error might be influenced antagonistically by the reward in the current trial and those in the previous trials. For example, the neuron illustrated in Figure 5 decreased its activity in response to the reward in the current trial but increased its activity if the animal was rewarded in the previous trials (Fig. 5, bottom). Accordingly, the same neuron modulated its activity during and immediately after the feedback period according to the reward prediction error, whereas the activity of the same neuron was affected only weakly by the sum of the value functions or their difference (Fig. 8*B*, bottom). Similarly, across the population of ACCd neurons the regression coefficients related to the reward prediction error were correlated positively with the coefficients related to the reward in the current trial but tended to be correlated negatively with the coefficients related to the rewards in previous trials. This was true for the activity both during the feedback period (Fig. 9, bottom, gray symbols) and immediately after the feedback period (Fig. 9, bottom, black symbols).

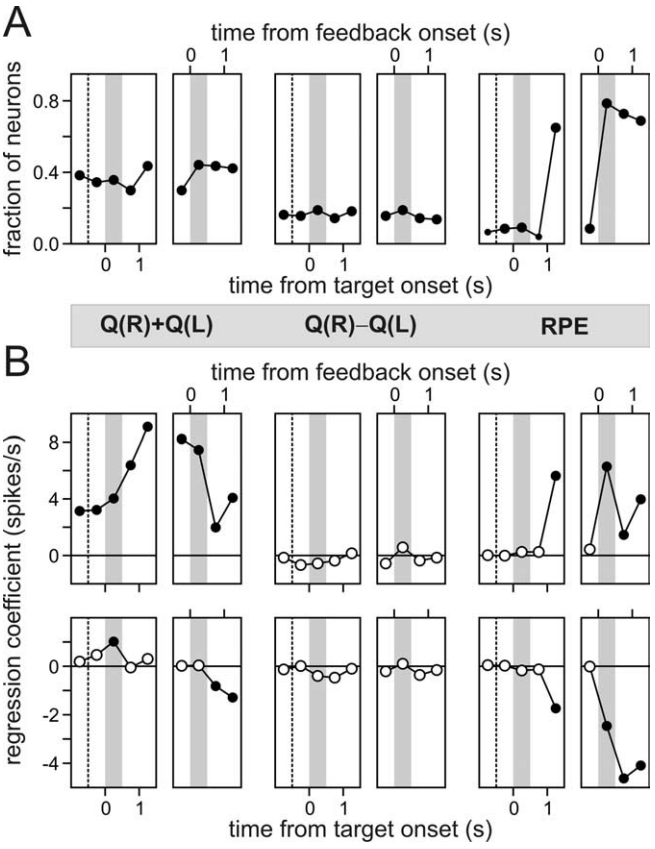

Figure 8. Neural signals related to value functions and reward prediction errors in ACCd. ***A***, Fraction of neurons that significantly modulated their activity according to the sum of the value functions (left), the difference between them (middle), and the reward prediction error during the successive 0.5 s windows used in the regression analysis. Large symbols indicate that the percentage of neurons was significantly higher than the significance level used in the regression analysis (binomial test, $p < 0.05$). ***B***, The regression coefficients associated with the sum of the value functions (left), the difference between them (middle), and the reward prediction error for the same two example neurons shown in Figures 4 (top) and 5 (bottom). Filled symbols indicate that the regression coefficients were significantly different from zero (Student's *t* test, $p < 0.05$). The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

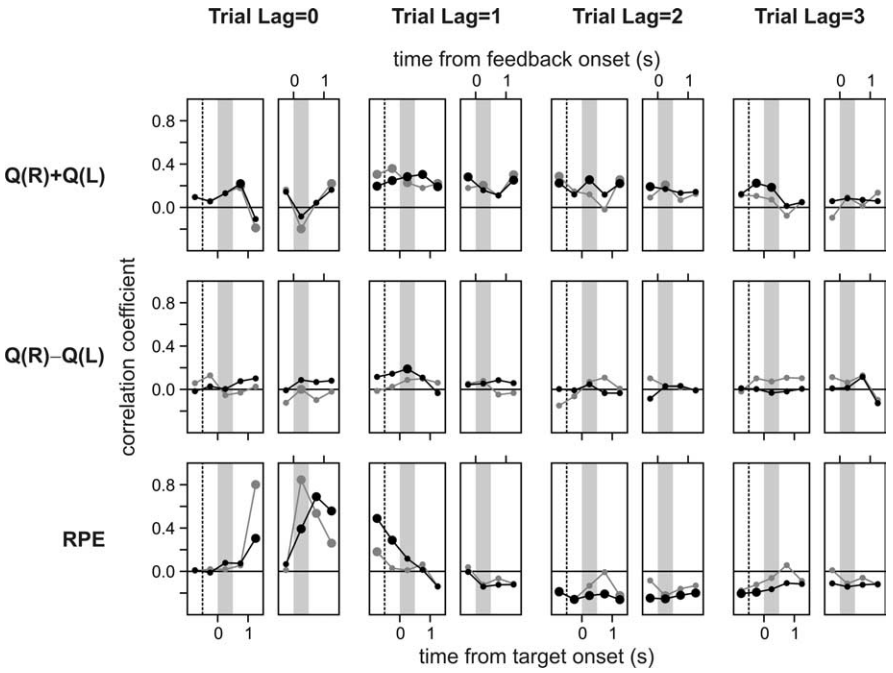

Figure 9. Correlation between the regression coefficients associated with the variables in the reinforcement learning model and those associated with the animal's rewards in the current and previous trials. Gray symbols correspond to the correlation coefficient for the sum of the value functions (top), the difference between the value functions (middle), and the reward prediction error (RPE; bottom) estimated for the activity during the feedback period. Black symbols show the results for the activity during the second 0.5 s window after feedback onset. Large symbols indicate that the correlation was statistically significant (Student's *t* test; $p < 0.05$). The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

Nonlinear effects of reward on neural activity in the ACCd

The results from the regression analysis are based on the assumption that the effects of rewards in successive trials are combined linearly. Therefore, they would provide only approximate descriptions to the real data if there are significant interactions between the rewards in multiple trials. Therefore, the assumption of additivity was tested with a three-way ANOVA in which the main factors included the rewards in two previous trials in addition to the reward in the current trial. As in the regression analysis described above, this ANOVA was applied to a series of 0.5 s bins relative to the target onset or feedback onset. The results from this analysis showed that some neurons displayed significant interactions among the rewards in multiple trials. For example, the neuron illustrated in Figure 4 displayed a significant two-way interaction between the rewards in the current and previous trials ($p < 0.001$). For this neuron the activity was lower during the feedback period in unrewarded trials when compared with the activity in rewarded trials (Figs. 4, 10). In addition, the effect of the reward in the previous trial on neural activity was reduced in unrewarded trials (Fig. 10). Despite this interaction effect, however, the activity during the feedback period in rewarded trials was higher if the animal also was rewarded in the previous trial, suggesting that this neuron encoded the temporal sum of the rewards reliably in rewarded trials. A similar interaction effect was found for the neuron illustrated in Figure 5 ($p < 0.0001$). This neuron became almost completely silent during the feedback period of rewarded trials regardless of the outcomes in the two previous trials (Fig. 11, right column), whereas the effect of reward in the previous trial was apparent in unrewarded trials (Fig. 11, left column). As a result, this neuron encoded the temporal difference of rewards only in unrewarded trials. During the first 0.5 s period after the feedback period the neuron displayed a significant three-way interaction, indicating that the activity after three consecutive unrewarded trials was significantly lower than the activity expected based on the main effects and two-way interactions. Overall, significant main effects of the reward in the current trial were found most frequently during the feedback period (125 neurons; 81.2%) (Fig. 12), and the effects of rewards in the two previous trials decreased gradually (68 and 28 neurons; 44.2 and 18.2%, respectively), consistent with the results from the regression analysis. In addition, during the feedback period the two-way interaction between the reward in the current trial and that in the previous trial was significant in 45 neurons (29.2%).

As illustrated by the example neurons described above, significant interactions between the reward in the current trial and that in the previous trial tended to occur mainly because the strength of signals related to the reward in the previous trial tended to decrease when the overall neural activity was reduced by the outcome of the animal's choice in a given trial. This was quantified by calculating the *t* value for the effect of reward in the previous trial separately according to whether the animal was rewarded or not in the current trial. Overall, the *t* values for the rewarded and unrewarded trials were significantly correlated regardless of whether the neurons decreased ($n = 80$; $r = 0.382$; $p < 0.001$) or increased ($n = 74$; $r = 0.501$; $p < 0.0001$) their activity in rewarded trials (Fig. 13). Thus the effect of reward in the previous trial tended to influence the activity similarly in rewarded and unrewarded trials. In addition, for the neurons that did not show any significant interactions between the rewards in the current and previous trials, the average magnitude of these *t* values did not differ significantly for rewarded and unrewarded trials (paired Student's *t* test; $p > 0.4$) (Fig. 13, unfilled circles). However, for the neurons with significant interactions the magnitude of these *t* values differed significantly for rewarded and unrewarded trials according to whether the neurons increased their activity in rewarded trials or not. For neurons that decreased their activity during the feedback period in rewarded trials, the average magnitude of these *t* values was significantly smaller when the animal was rewarded (paired Student's *t* test, $p < 0.001$) (Fig. 13, left). Similarly, the average magnitude of these *t* values was significantly larger when the animal was rewarded if the activity was higher in rewarded trials ($p < 0.005$) (Fig. 13, right). Thus when a neuron displayed a significant interaction between the reward in the current trial and that in the previous trial, the signals related to the reward in the previous trial were conveyed more reliably by the neurons that increased their activity according to the outcome of the current trial. This was true regardless of whether the reward in the previous trial influenced the activity of a neuron in the same direction as the reward in the current trial or not (Fig. 13).

Encoding of choice-related signals in the ACCd

Compared with the activity changes related to the reward, signals in the activity of ACCd neurons that are related to the animal's choice or the choice of the computer opponent were observed less frequently. For example, the neurons illustrated in Figures 4 and 5 displayed statistically significant changes in their activity related to the animal's choice during the first 0.5 s after the delay period, indicating that both neurons displayed direction-selective eye movement-related activity. The neuron shown in Figure 4 increased its activity more when the animal chose the left-hand target, whereas the neuron in Figure 5 showed the opposite pattern. In both cases, however, this effect was relatively small compared with the effect of reward and was not sustained beyond the feedback period. The neuron in Figure 4 did not show any effect of computer choice, whereas the neuron in Figure 5 showed a small effect during the feedback period. Given that the statistical test was performed in multiple time bins, the possibility that these may correspond to the type I errors cannot be excluded in these individual examples. Nevertheless, the population analysis showed that the fraction of neurons in the ACCd showing significant changes in their activity according to the animal's choice and the choice of the computer opponent was higher than expected by chance. For example, during the delay period 17.5% of the ACCd neurons significantly modulated their activity according to the target chosen by the animal in the same trial (Fig. 6, top), and this percentage increased to 41.6% during the period in which the animal maintained its fixation on the chosen target. During the delay period the percentage of the neurons that displayed significant modulations in their activity according to the animal's choice in the previous trial was 18.2%. Although the fraction of neurons with activity significantly related to the computer choice was above the chance level, this remained relatively low for the time bins included in this analysis. Its maximum value was 22.7%, and this occurred during the feedback period when the computer choice was revealed (Fig. 6, middle).

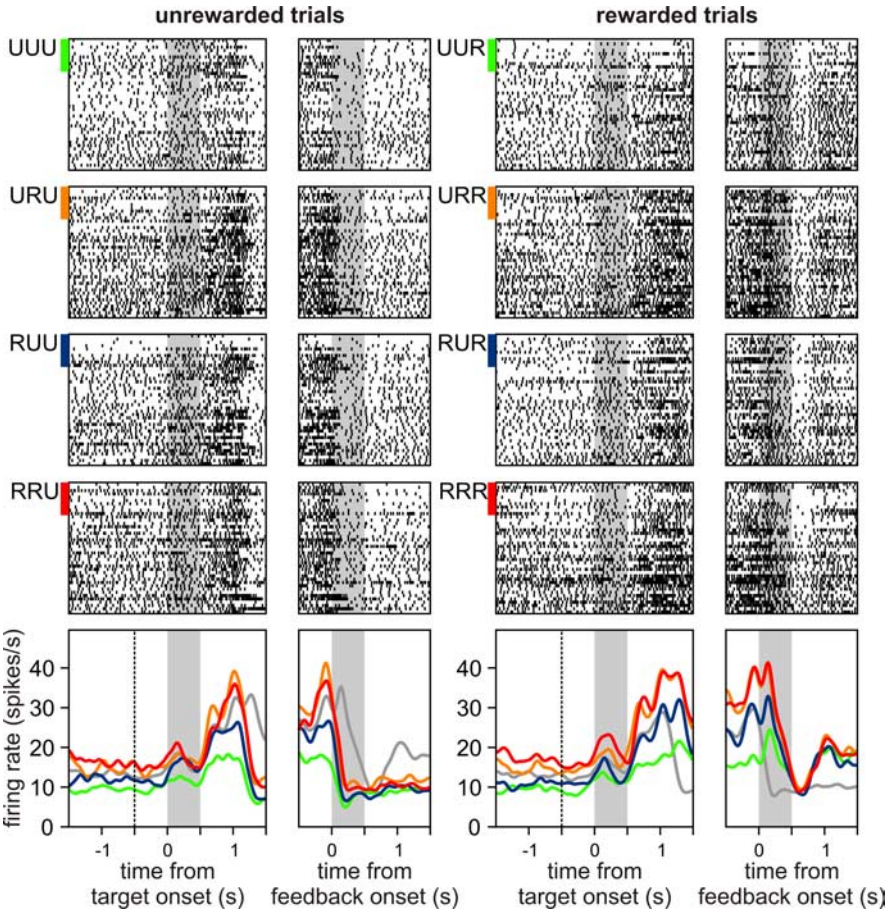

Figure 10. Raster plots and spike density functions for the same neuron shown in Figure 4 sorted by the reward in the current trial (left, unrewarded; right, rewarded) and those in the two previous trials. A three-letter code shown on the left of the raster plots indicates the trials in which the animal was rewarded. For example, RRU indicates that the animal was rewarded in both of the previous two trials, but not in the current trial. Colors of the spike density functions in the bottom panels correspond to those of small bars associated with the raster plots for different reward sequences, except that gray lines in the left (right) column correspond to the average spike density functions for rewarded (unrewarded) trials. The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

Discussion

Reinforcement learning and stochastic decision making

Economists traditionally have applied the principle of utility maximization to account for the choice behavior of human decision makers. However, this begs the question of how the utilities of various objects and behaviors can be estimated properly in the first place. Reinforcement learning theory provides a solution to this problem (Sutton and Barto, 1998). Algorithms in this framework maintain a set of value functions to estimate the desirability of a particular state or a particular action in a given environment. Value functions are adjusted according to the reward prediction error, namely, discrepancy between the reward predicted by the current set of value functions and the actual reward received. Finally, actions are chosen to maximize the value functions, but this often is done probabilistically so that even the actions with smaller value functions sometimes are chosen. This facilitates exploration and allows the decision maker to discover the consequences of previously unexplored actions (Daw et al., 2006).

During the matching pennies task used in the present study the optimal strategy is to select the two targets with equal probabilities and independently across trials. Nevertheless, the animals in the present study displayed small but systematic deviations from this optimal strategy. First, the probability of using the so-called win–stay–lose–switch strategy sometimes exceeded the chance level significantly. Second, the learning rate in the reinforcement learning model used in this study was relatively small, indicating that the animals tended to apply the win–stay–lose–switch strategy across multiple trials. Thus if the animal was rewarded for choosing a particular target in a given trial, the animal was more likely to choose the same target not only in the next trial but also in subsequent trials. This behavioral finding implies that signals related to the animal's previous choices and their outcomes must be integrated temporally in the brain (Barraclough et al., 2004; Sugrue et al., 2004; Kennerley et al., 2006; Seo et al., 2007).

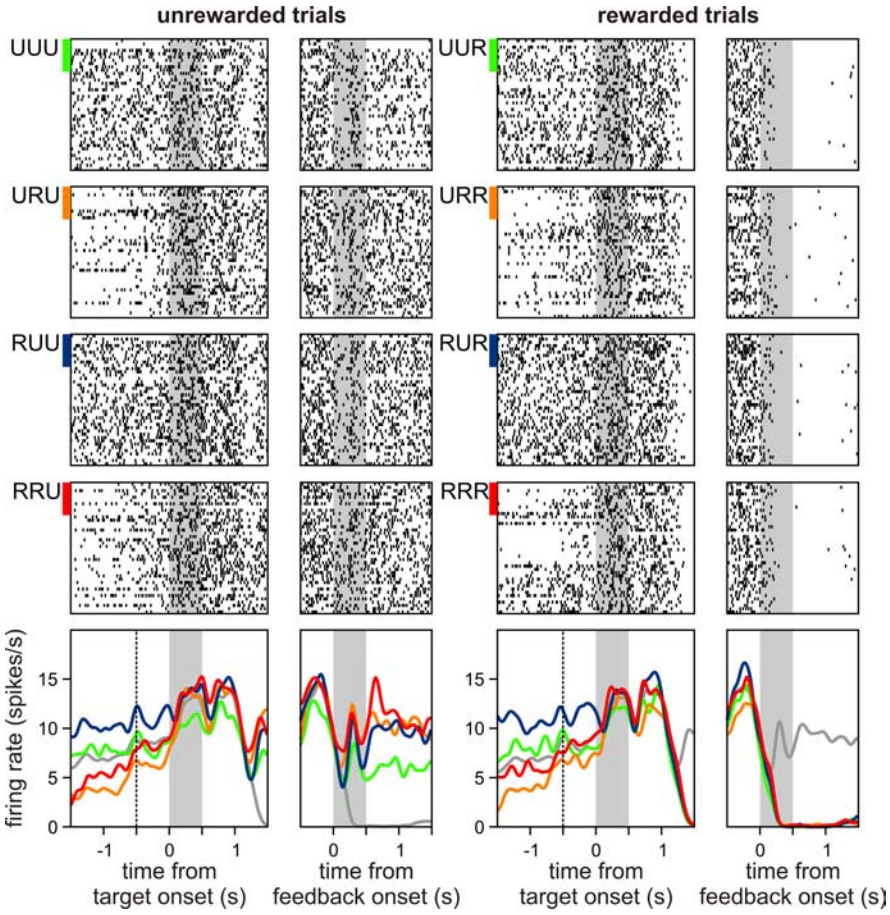

Figure 11. Raster plots and spike density functions for the same neuron shown in Figure 5. This is the same format as in Figure 10.

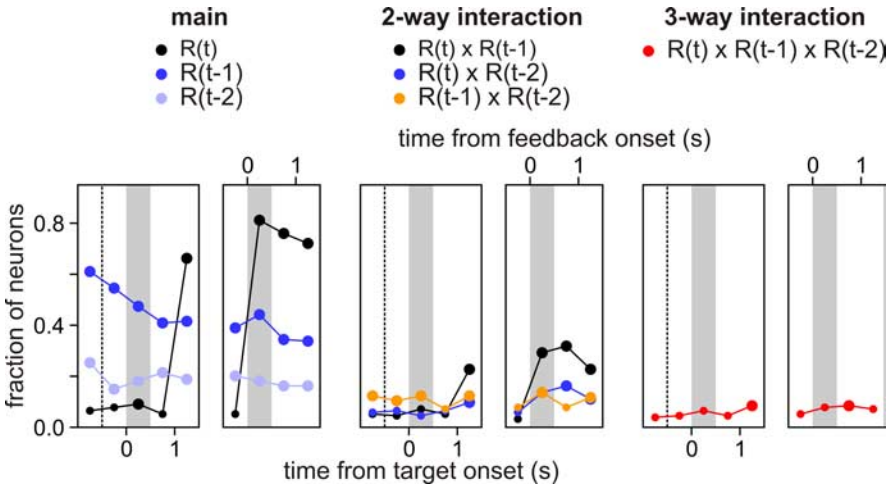

Figure 12. Fraction of neurons showing the significant main effect (left), two-way interactions (middle), and three-way interaction in a three-way ANOVA that included the rewards received by the animal in the current trial, R(t), and the two previous trials, $R(t-1)$ and $R(t-2)$. The dotted vertical lines in the left panels correspond to the onset of the fore period, and the gray background corresponds to the delay (left panels) or feedback (right panels) period.

Reinforcement learning and anterior cingulate cortex

Reinforcement learning models provide a useful framework to investigate the neural mechanisms of decision making. For example, the activity of midbrain dopamine neurons has been linked to reward prediction error (Schultz, 1998). In addition, signals resembling value functions have been identified in a broad network of brain areas, including the dorsolateral prefrontal cortex (DLPFC) (Watanabe, 1996; Leon and Shadlen, 1999), SEF (Ito et al., 2003; Roesch and Olson, 2003), ACCd (Shidara and Richmond, 2002), and the basal ganglia (Hollerman et al., 1998; Kawagoe et al., 1998). Neuroimaging studies in human subjects have identified activity related to expected reward in similar brain areas (O'Doherty et al., 2002; Knutson et al., 2005). Studies in which monkeys made free choices also have shown that the activity in many brain areas modulated activity according to value functions (Platt and Glimcher, 1999; Barraclough et al., 2004; Sugrue et al., 2004; McCoy and Platt, 2005; Samejima et al., 2005).

Previous studies have found that the DLPFC might provide several different types of signals necessary for updating the value functions during the matching pennies task used in the present study (Barraclough et al., 2004; Seo et al., 2007). First, many DLPFC neurons display signals related to the animal's previous choices. This might correspond to eligibility trace, which is necessary to link a particular outcome to its causative action, especially when the outcome is delayed temporally. Second, neurons in the DLPFC often encoded the information about the previous choices of the computer opponent. During the matching pennies task the animal is rewarded only when it selects the same target chosen by the computer opponent. Therefore, signals related to the choice of the opponent might be integrated temporally to estimate the probability that the animal might be rewarded for choosing a particular target. Finally, activity of many neurons in the DLPFC was affected by whether the animal was rewarded in each of the last few trials, and this might be used to estimate the average rate of reward (Seo et al., 2007).

Similar to the neurons in the DLPFC, the majority of the neurons in ACCd encoded signals related to the animal's reward (Niki and Watanabe, 1979; Ito et al., 2003; Amiez et al., 2005). Many neurons in the ACCd also modulated their activity according to the animal's reward history. In most cases reward-related activity in the ACCd could be approximated well by a linear function of rewards received by the animal in successive trials. Some neurons in the ACCd displayed signals related to the animal's previous choices, but this was observed less frequently when compared with the DLPFC. Similarly, the fraction of ACCd neurons encoding the difference in the value functions for the two alternative choices was relatively small, suggesting that the ACCd may play only a minor role in computing and representing the relative

desirability of alternative actions. Previous studies also have found that movement parameters are encoded infrequently by the neurons in ACCd (Ito et al., 2003; Hoshi et al., 2005; Matsumoto et al., 2007). On the other hand, it has been shown that some neurons in the ACCd encoded information about the animal's movement in conjunction with its expected or actual consequences under some circumstances, especially when the animal was required to discover correct actions based on its reward history (Shima and Tanji, 1998; Matsumoto et al., 2003). During the matching pennies task the reward or the lack thereof only weakly influenced the animal's next choice, and this might account for the paucity of neurons in the ACCd that encode movement parameters in the present study.

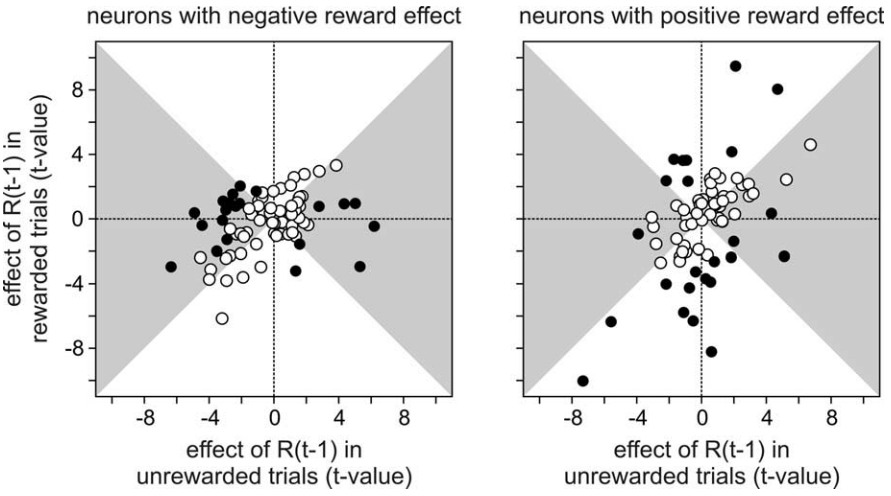

Figure 13. Interaction between the reward in the previous trial and that in the current trial. The effect of reward in the previous trial on neural activity was quantified by using a *t* value separately for rewarded (abscissa) and unrewarded (ordinate) trials. Then the results were plotted separately, depending on whether a given neuron decreased (left, negative reward effect) or increased (right, positive reward effect) its activity during the feedback period of rewarded trials as compared with the activity in unrewarded trials. Filled circles indicate the neurons that displayed significant interactions between the reward in the previous trial and that in the current trial, and the gray background indicates that the magnitude of the *t* value in unrewarded trials is larger than that in rewarded trials.

Reward history and reward prediction error in anterior cingulate cortex

The time course of reward-related signals in the ACCd was quite heterogeneous, and it might reflect different types of temporal filtering or transformation applied to the animal's reward history. More specifically, the activity of some neurons was influenced consistently by the reward in the current and previous trials, suggesting that they might encode the rate of reward necessary to compute the expected reward. In contrast, some neurons modulated their activity antagonistically according to the reward in the current and previous trials, as expected for the neurons encoding reward prediction errors (Amiez et al., 2005; Matsumoto et al., 2007). The reward-related activity observed in the present study often was maintained continuously even during intertrial intervals, whereas signals related to reward prediction errors in the dopamine neurons were transient (Schultz, 1998; Bayer and Glimcher, 2005). This suggests that reward-related signals in the ACCd might provide the information necessary to compute reward prediction errors in dopamine neurons.

In both humans and other animals a particular outcome can produce different emotional reactions and influence subsequent choices differently, depending on whether it is perceived as a gain or loss relative to a certain reference point (Tinklepaugh, 1928; Crespi, 1942; Zeaman, 1949; Kahneman and Tversky, 1979; Flaherty, 1982). Such a reference point may be determined by the overall rate of reward estimated from the recent experience of the decision maker (Helson, 1948), although this might be affected by other contextual factors, such as the outcomes available to other members in a society (Fehr and Schmidt, 1999; Brosnan and de Waal, 2003). The results from the present study suggest that some neurons in the ACCd might encode the overall reward rate, and, therefore, they might be involved in evaluating a hedonic reference point (Frederick and Loewenstein, 1999). Previously, it has been shown that the activity of some neurons in the ACCd was closely related to the expected value of reward, regardless of whether the magnitude of reward was fixed or probabilistic (Amiez et al., 2006). The results from the present study suggest that this might result from the temporal filtering of reward signals in the ACCd. Consistent with these findings, it has been shown that a lesion in the cingulate cortex of rats interferes with the animal's sensitivity to changes in the reward rate (Gurowitz et al., 1970). In addition, lesions (Hadland et al., 2003; Kennerley et al., 2006) or reversible inactivations (Shima and Tanji, 1998; Amiez et al., 2006) in the cingulate cortex of monkeys produce deficits in the animal's ability to learn or update an appropriate association between a particular action and reward. The results from the present study suggest that such a deficit may result from the disruption in the process of evaluating reward signals in the context of the animal's previous experience.

References

Amiez C, Joseph J-P, Procyk E (2005) Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci 21:3447–3452.

Amiez C, Joseph J-P, Procyk E (2006) Reward encoding in the monkey anterior cingulate cortex. Cereb Cortex 16:1040–1055.

Aston-Jones G, Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28:403–450.

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404–410.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141.

Brosnan S, de Waal FB (2003) Monkeys reject unequal pay. Nature 425:297–299.

Burnham KP, Anderson DR (2002) Model selection and multimodel inference. A practical information-theoretic approach, Ed 2. New York: Springer.

Crespi LP (1942) Quantitative variation of incentive and performance in the white rat. Am J Psychol 55:467–517.

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199–204.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Erev I, Roth AE (1998) Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. Am Econ Rev 88:848–881.

Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114:817–868.

Flaherty CF (1982) Incentive contrast: a review of behavioral changes following shifts in reward. Anim Learn Behav 10:409–440.

Frederick S, Loewenstein G (1999) Hedonic adaptation. In: Well-being: the foundations of hedonic psychology (Kaheneman D, Diener E, Schwartz N, eds), pp 302–329. New York: Russell Sage Foundation.

Goldberg ME, Bushnell MC, Bruce CJ (1986) The effect of attentive fixation on eye movements evoked by electrical stimulation of the frontal eye fields. Exp Brain Res 61:579–584.

Gurowitz EM, Rosen AJ, Tessel RE (1970) Incentive shift performance in cingulectomized rats. J Comp Physiol Psychol 70:476–481.

Hadland KA, Rushworth MFS, Gaffan D, Passingham RE (2003) The anterior cingulate and reward-guided selection of action. J Neurophysiol 89:1161–1164.

Haruno M, Kawato M (2006) Different neural correlates of reward expectation and reward expectation error in putamen and caudate nucleus during stimulus-action-reward association learning. J Neurophysiol 95:948–959.

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Data mining, inference and prediction. New York: Springer.

Helson H (1948) Adaptation-level as a basis for a quantitative theory of frames of reference. Psychol Rev 55:297–313.

Hollerman JR, Tremblay L, Schultz W (1998) Influence of reward expectation on behavior-related neuronal activity in primate striatum. J Neurophysiol 80:947–963.

Hoshi E, Sawamura H, Tanji J (2005) Neurons in the rostral cingulate motor area monitor multiple phases of visuomotor behavior with modest parametric selectivity. J Neurophysiol 94:640–656.

Ito S, Stuphorn V, Brown JW, Schall JD (2003) Performance monitoring by the anterior cingulate cortex during saccade countermanding. Science 302:120–122.

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291.

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

Kennerley SW, Walton ME, Behrens TEJ, Buckley MJ, Rushworth MFS (2006) Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9:940–947.

Knutson B, Taylor J, Kaufman M, Peterson R, Glover G (2005) Distributed neural representation of expected value. J Neurosci 25:4806–4812.

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579.

Lee D (2006) Neural basis of quasi-rational decision making. Curr Opin Neurobiol 16:191–198.

Lee D, Port NL, Kruse W, Georgopoulos AP (2001) Neuronal clusters in the primate motor cortex during interception of moving targets. J Cogn Neurosci 13:319–331.

Lee D, Conroy ML, McGreevy BP, Barraclough DJ (2004) Reinforcement learning and decision making in monkeys during a competitive game. Brain Res Cogn Brain Res 22:45–58.

Lee D, McGreevy BP, Barraclough DJ (2005) Learning and decision making in monkeys during a rock-paper-scissors game. Brain Res Cogn Brain Res 25:416–430.

Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24:415–425.

Luppino G, Rozzi S, Calzavara R, Matelli M (2003) Prefrontal and agranular cingulate projections to the dorsal premotor areas F2 and F7 in the macaque monkey. Eur J Neurosci 17:559–578.

Matelli M, Luppino G, Rizzolatti G (1991) Architecture of superior and mesial area 6 and the adjacent cingulate cortex in the macaque monkey. J Comp Neurol 311:455–462.

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

Matsumoto M, Matsumoto K, Abe H, Tanaka K (2007) Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10:647–656.

McCoy AN, Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. Nat Neurosci 8:1220–1227.

Niki H, Watanabe M (1979) Prefrontal and cingulate unit activity during timing behavior in the monkey. Brain Res 171:213–224.

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826.

Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Oxford: Claredon.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400:233–238.

Roesch MR, Olson CR (2003) Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields, and premotor cortex. J Neurophysiol 90:1766–1789.

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

Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1–27.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57:87–115.

Seo H, Barraclough DJ, Lee D (2007) Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb Cortex, in press.

Shidara M, Richmond BJ (2002) Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science 296:1709–1711.

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

Soltani A, Lee D, Wang X-J (2006) Neural mechanism for stochastic behavior during a competitive game. Neural Netw 19:1075–1090.

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782–1787.

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

Tinklepaugh OL (1928) An experimental study of representative factors in monkeys. J Comp Psychol 8:197–236.

Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629–632.

Yechiam E, Busemeyer JR (2005) Comparison of basic assumptions embedded in learning models for experienced-based decision making. Psychon Bull Rev 12:387–402.

Zeaman D (1949) Response latency as a function of the amount of reinforcement. J Exp Psychol 39:466–483.