nature neuroscience

Prefrontal cortex and decision making in a mixed-strategy game

Dominic J Barraclough, Michelle L Conroy & Daeyeol Lee

In a multi-agent environment, where the outcomes of one's actions change dynamically because they are related to the behavior of other beings, it becomes difficult to make an optimal decision about how to act. Although game theory provides normative solutions for decision making in groups, how such decision-making strategies are altered by experience is poorly understood. These adaptive processes might resemble reinforcement learning algorithms, which provide a general framework for finding optimal strategies in a dynamic environment. Here we investigated the role of prefrontal cortex (PFC) in dynamic decision making in monkeys. As in reinforcement learning, the animal's choice during a competitive game was biased by its choice and reward history, as well as by the strategies of its opponent. Furthermore, neurons in the dorsolateral prefrontal cortex (DLPFC) encoded the animal's past decisions and payoffs, as well as the conjunction between the two, providing signals necessary to update the estimates of expected reward. Thus, PFC might have a key role in optimizing decision-making strategies.

Decision making refers to an evaluative process of selecting a particular action from a set of alternatives. When the mapping between a particular action and its outcome or utility is fixed, the decision to select the action with maximum utility can be considered optimal or rational. However, animals face more difficult problems in a multi-agent environment, in which the outcome of one's decision can be influenced by the decisions of other animals. Game theory provides a mathematical framework to analyze decision making in a group of agents[1–4]. A game is defined by a set of actions available to each player, and a payoff matrix that specifies the reward or penalty for each player as a function of decisions made by all players. A solution or equilibrium in game theory refers to a set of strategies selected by a group of rational players[1,5,6]. Nash has proved that any *n*-player competitive game has at least one equilibrium in which no players can benefit by changing their strategies individually[5]. These equilibrium strategies often take the form of a mixed strategy, which is defined as a probability density function over the alternative actions available to each player. This requires players to choose randomly among alternative choices, as in the game of rock-paper-scissors during which choosing one of the alternatives (*e.g.*, paper) exclusively allows the opponent to exploit such a biased choice (with scissors).

Many studies have shown that people frequently deviate from the predictions of game theory[7–21]. Although the magnitudes of such deviations are often small, they have important implications regarding the validity of assumptions in game theory, such as the rationality of human decision-makers[22–27]. In addition, strategies of human decision-makers change with their experience[17–21]. These adaptive processes might be based on reinforcement learning algorithms[28], which can be used to approximate optimal decision-making strategies in a dynamic environment. In the present study, we analyzed the performance of monkeys playing a zero-sum game against a computer opponent to determine how closely their behaviors match the predictions of game theory and whether reinforcement learning algorithms can account for any deviations from such predictions. In addition, neural activity was recorded from the DLPFC to investigate its role during strategic decision making in a multi-agent environment. The results showed that the animal's choice behavior during a competitive game could be accounted for by a reinforcement learning algorithm. Individual prefrontal neurons often modulated their activity according to the choice of the animal in the previous trial, the outcome of that choice, and the conjunction between the choice and its outcome. This suggests that the PFC may be involved in updating the animal's decision-making strategy based on a reinforcement learning algorithm.

RESULTS

Behavioral performance

Two rhesus monkeys played a game analogous to matching pennies against a computer in an oculomotor free-choice task (**Fig. 1a**; Methods). The animal was rewarded when it selected the same target as the computer that was programmed to minimize the animal's reward by exploiting the statistical bias in the animal's choice behavior. Accordingly, the optimal strategy for the animal was to choose the targets randomly with equal probabilities, which corresponds to the Nash equilibrium in the matching pennies game. To determine how the animal's decisions were influenced by the strategy of the opponent, we manipulated the amount of information that was used by the computer opponent (see Methods). In algorithm 0, the computer selected its targets randomly with equal probabilities, regardless of the animal's choice patterns. In algorithm 1, the com-

Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, New York 14627, USA. Correspondence should be addressed to D.L. (dlee@cvs.rochester.edu).

Published online 7 March 2004; doi:10.1038/nn1209

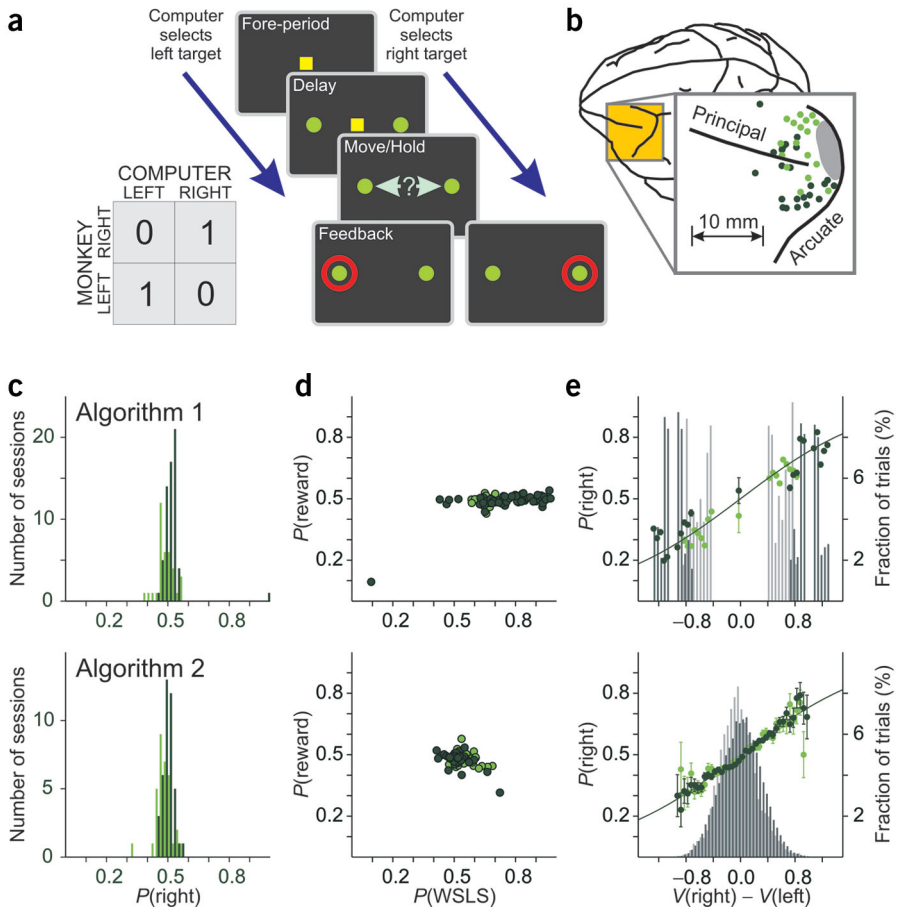

Figure 1 Task and behavioral performance. (**a**) Free-choice task and payoff matrix for the animal during the competitive game (1, reward; 0, no reward). (**b**) Recording sites in PFC. Frontal eye field (gray area in the inset) was defined by electrical stimulation[50]. (**c**) Frequency histograms for the probability of choosing the right-hand target in algorithms 1 and 2. (**d**) Probability of the win-stay-lose-switch (WSLS) strategy (abscissa) versus probability of reward (ordinate). (**e**) Difference in the value functions for the two targets estimated from a reinforcement learning model (abscissa) versus the probability of choosing the target at the right (ordinate). Error bars indicate standard error of the mean (s.e.m.). Histograms show the frequency of trials versus the difference in the value functions. Solid line, prediction of the reinforcement learning model. In all panels, dark and light symbols indicate the results from the two animals, respectively.

puter analyzed only the animal's choice history, but not its reward history. In algorithm 2, both choice and reward histories were analyzed. In both algorithms 1 and 2, the computer chose its target randomly if it did not find any systematic bias in the animal's choice behavior. Therefore, a reward rate near 0.5 indicates that the animal's performance was optimal.

Indeed, the animal's reward rate was close to 0.5 for all algorithms, indicating that the animal's performance was nearly optimal. In algorithm 0, the reward rate was fixed at 0.5 regardless of the animal's behavior, and therefore there was no incentive for the animal to choose the targets with equal probabilities. In fact, both animals chose the right-hand target more frequently ($P = 0.70$ and 0.90 and $n = 5{,}327$ and 1,669 trials, for the two animals, respectively) than the left-hand target. For the remaining two algorithms, the probability of choosing the right-hand target was much closer to 0.5 (**Fig. 1c**), which corresponds to the Nash equilibrium of the matching pennies game. In addition, the probability of choosing a given target was relatively unaffected by the animal's choice in the previous trial. For example, the probability that the animal would select the same target as in the previous trial was also close to 0.5 ($P = 0.51 \pm 0.06$ and 0.50 ± 0.04 and $n = 120{,}254$ and 74,113 trials, for algorithms 1 and 2, respectively). In contrast, the animal's choice was strongly influenced by the computer's choice in the previous trial, especially in algorithm 1. In the game of matching pennies, the strategy to choose the same target selected by the opponent in the previous trial can be referred to as a win-stay-lose-switch (WSLS) strategy, as this is equivalent to choosing the same target as in the previous trial if that choice was rewarded and choosing the opposite target otherwise. The probability of the WSLS strategy in algorithm 1 (0.73 ± 0.14) was significantly higher than that in algorithm 2 (0.53 ± 0.06; $P < 10^{-16}$; **Fig. 1d**). Although the tendency for the WSLS strategy in algorithm 2 was only slightly above chance, this bias was still statistically significant ($P < 10^{-5}$). Similarly, average mutual information between the sequence of animal's choice and reward in three successive trials and the animal's choice in the following trial decreased from 0.245 ($\pm$ 0.205) bits for algorithm 1 to 0.043 ($\pm$ 0.035) bits for algorithm 2.

Table 1 Parameters for the reinforcement learning model.

Algorithm	Monkey	α	Δ_1	Δ_2
1	C	0.176	0.661	–0.554
		(0.130, 0.220)	(0.619, 0.704)	(–0.597, –0.512)
	E	0.170	0.941	–1.064
		(0.143, 0.198)	(0.903, 0.979)	(–1.104, –1.024)
2	C	0.986	0.033	0.016
		(0.983, 0.988)	(0.028, 0.039)	(0.012, 0.021)
	E	0.828	0.195	–0.143
		(0.801, 0.851)	(0.171, 0.218)	(–0.169, –0.118)

α, discount factor; Δ_1 and Δ_2, changes in the value function associated with rewarded and unrewarded targets selected by the animal, respectively. The numbers in parentheses indicate 99.9% confidence intervals.

Reinforcement learning model

Using a reinforcement learning model[19–21,28,29], we tested whether the animal's decision was systematically influenced by the cumulative effects of reward history. In this model, a decision was based on the difference between the value functions (that is, expected reward) for the two targets. Denoting the value functions of the two targets (L and R) at trial t as $V_t(\mathrm{L})$ and $V_t(\mathrm{R})$, the probability of choosing each target is given by the logit transformation of the difference between the value functions[30]. In other words,

$$\mathrm{logit}\, P(\mathrm{R}) \equiv \log P(\mathrm{R})/(1 - P(\mathrm{R})) = V_t(\mathrm{R}) - V_t(\mathrm{L}).$$

The value function, $V_t(x)$, for target x, was updated after each trial according to the following:

$$V_{t+1}(x) = \alpha V_t(x) + \Delta_t(x),$$

where α is a discount factor, and $\Delta_t(x)$ denotes the change in the value function determined by the animal's decision and its outcome. In the current model, $\Delta_t(x) = \Delta_1$ if the animal selects the target x and is rewarded, $\Delta_t(x) = \Delta_2$ if the animal selects the target x and is not rewarded, and $\Delta_t(x) = 0$ if the animal does not select the target x. We introduced a separate parameter for the unrewarded target (Δ_2) because the probability of choosing the same target after losing a reward was significantly different from the probability of switching to the other target for all animals and for both algorithms 1 and 2. Maximum likelihood estimates[31] of the model parameters (**Table 1**) showed that a frequent use of the WSLS strategy during algorithm 1 was reflected in a relatively small discount factor ($\alpha < 0.2$), a large positive Δ_1 (> 0.6) and a large negative Δ_2 (< -0.5) in both animals. For algorithm 1, this led to a largely bimodal distribution for the difference in the value functions (**Fig. 1e**). In contrast, the magnitude of changes in value function during algorithm 2 was smaller, indicating that the outcome of previous choices only weakly influenced the subsequent choice of the animal. In addition, the discount factor for algorithm 2 was relatively large ($\alpha > 0.8$). This suggests that the animal's choice was systematically influenced by the combined effects of previous reward history even in algorithm 2. The combination of model parameters for algorithm 2 produced an approximately normal distribution for the difference in value functions (**Fig. 1e**). This implies that for most trials, the difference in the value functions of the two targets was relatively small, making it difficult to predict the animal's choice reliably. These results suggest that during a competitive game, the monkeys might have approximated the optimal decision-making strategy using a reinforcement learning algorithm.

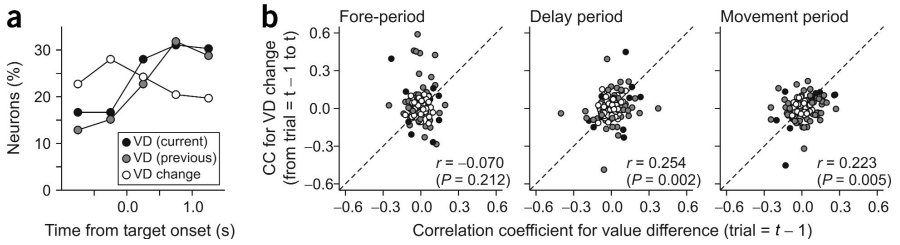

Figure 2 Effects of relative expected reward (*i.e.*, difference in value functions) and its trial-to-trial changes on the activity of prefrontal neurons. (**a**) Percentage of neurons with significant correlation (*t*-test, $P < 0.05$) between their activity and the difference in the value functions for the two targets (VD = $V_\tau(R) - V_\tau(L)$) estimated for the current ($\tau = t$) and previous ($\tau = t - 1$) trials, or between the activity and the changes in the value functions between the two successive trials (VD change = $\Delta_{t-1}(R) - \Delta_{t-1}(L)$). (**b**) Correlation coefficient between the VD change and the activity in a given neuron (ordinate), plotted against correlation coefficient between the VD in the previous trial (*i.e.*, $V_{t-1}(R) - V_{t-1}(L)$) and the activity of the same neuron (abscissa). Black (gray) symbols indicate the neurons in which both (either) correlation coefficients were significantly different from 0 (*t*-test, $P < 0.05$). The numbers in each panel correspond to Spearman's rank correlation coefficient (*r*) and its level of significance (*P*).

Prefrontal activity during a competitive game

The value functions in the above reinforcement learning model were updated according to the animal's decisions and the outcomes of those decisions. To determine whether such signals are encoded in the activity of individual neurons in PFC, we recorded single-neuron activity in the DLPFC while the animal played the same free-choice task. During the neurophysiological recording, the computer selected its target according to algorithm 2. A total of 132 neurons were examined during a minimum of 128 free-choice trials (mean = 583 trials; **Fig. 1b**). As a control, each neuron was also examined during 128 trials of a visual search task in

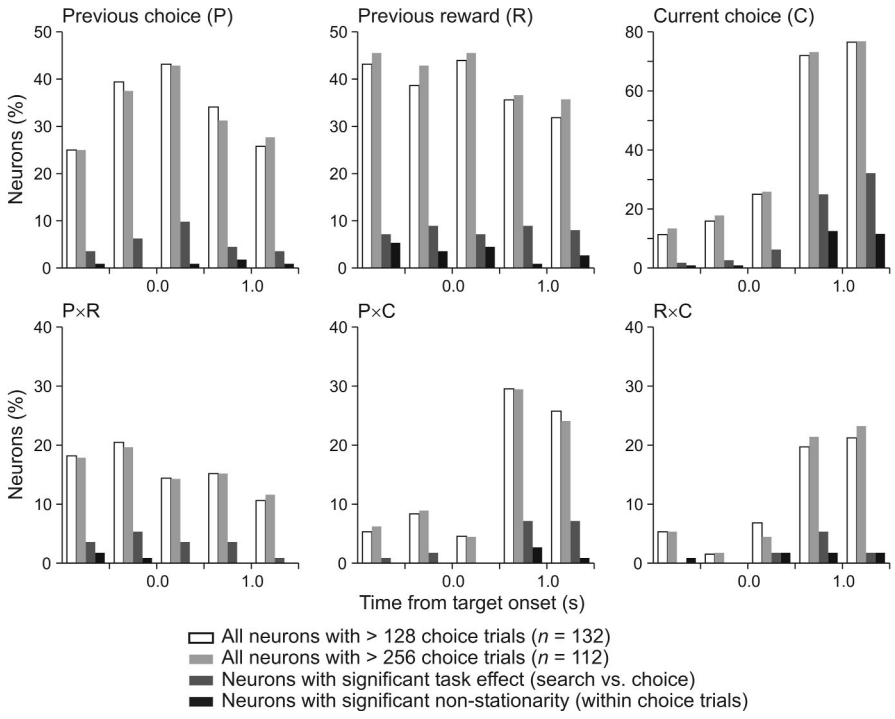

Figure 3 Percentages of neurons encoding signals related to the animal's decision. White bars show the percentage of neurons ($n = 132$) with significant main and interaction effects in a three-way ANOVA (P × R × C). Light gray bars show the same information for the neurons with >256 free-choice trials, which was tested for stationarity in free-choice trials ($n = 112$). Dark gray bars show the percentage of neurons with significant effects in the three-way ANOVA that also varied with the task (search vs. choice) in a four-way ANOVA (Task × P × R × C). This analysis was performed only for the neurons with >256 free-choice trials for comparison with the control analysis to test stationarity. Black histograms show the percentage of neurons with significant effects in the three-way ANOVA that also have significant non-stationarity in a control 4-way ANOVA across the two successive blocks of 128 free-choice trials.

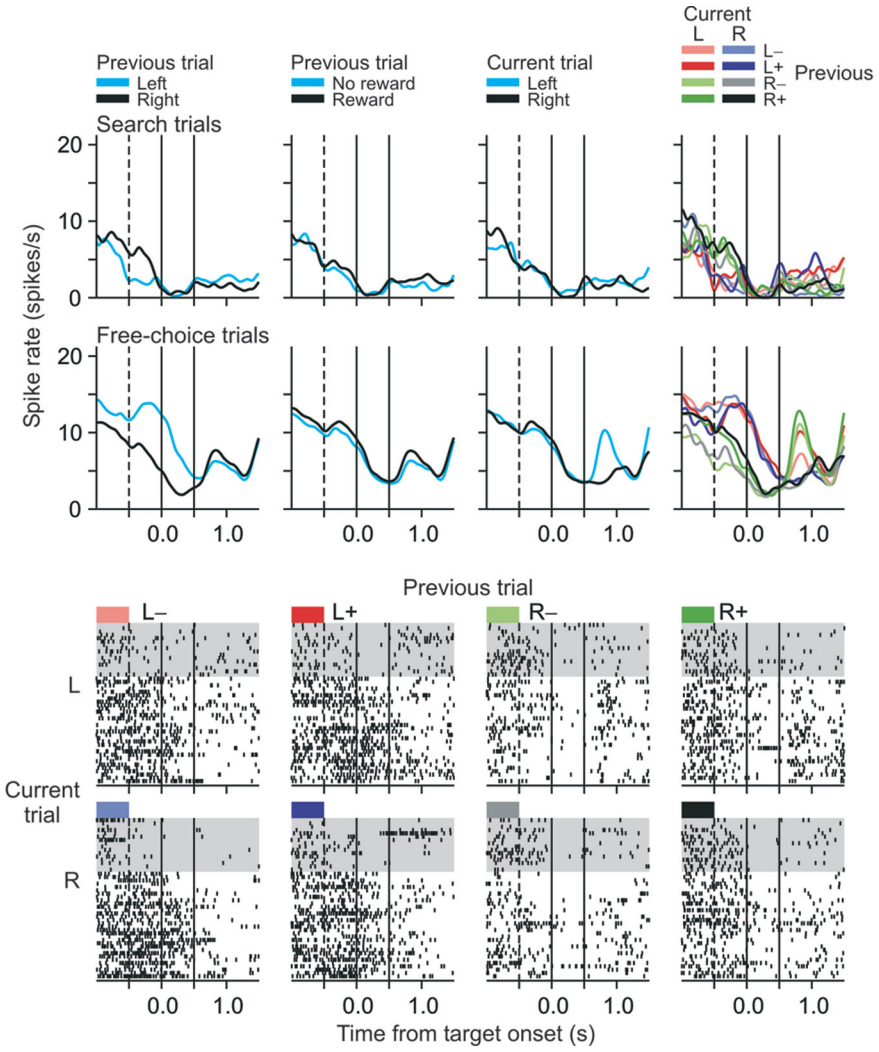

Figure 4 Example neuron showing a significant effect of the animal's choice in the previous trial. Top, spike density functions averaged according to the animal's choice (L and R) and reward (+, reward; –, no reward) in the previous trial and the choice in the current trial. A dotted vertical line indicates the onset of the fore-period, and the two solid lines the beginning and end of the delay period. Bottom, raster plots showing the activity of the same neuron sorted according to the same three factors during the search (gray background) and free-choice (white background) trials.

which the animal's decision was guided by sensory stimuli (Methods).

During the free-choice trials, activity in some prefrontal neurons was influenced by the difference in the value functions for the two targets (that is, $V(R) - V(L)$), although the effects in individual neurons were relatively small (**Fig. 2**). This was not entirely due to the animal's choice and its outcome in the previous trial, as the value functions estimated for the previous trial produced similar results (**Fig. 2a**). If individual PFC neurons are involved in the temporal integration of value functions according to the reinforcement learning model described above, differences in the value functions (*i.e.*, $V(x)$) and their changes (*i.e.*, $\Delta(x)$) would similarly influence the activity of PFC neurons. Interestingly, such patterns were found for the delay and movement periods, but not for the fore-period (Methods; **Fig. 2b**). These results suggest that some prefrontal neurons might be involved in temporally integrating the signals related to previous choice and its outcome to update value functions.

To examine how the activity of individual PFC neurons is influenced by the animal's choices and their outcomes, we analyzed neural activity by three-way ANOVA with the animal's choice and reward in the previous trial and its choice in the current trial as main factors. For 39% of PFC neurons, the activity during the fore-period was influenced by the animal's choice in the previous trial (**Fig. 3**). For example, in the neuron illustrated in **Figure 4**, the animal's choice in the previous trial exerted a significant influence on the activity before and during the fore-period, as well as during the delay period (3-way ANOVA, $P < 0.001$). In addition, activity during the movement period was still influenced by the animal's choice in the previous trial and its outcome, as well as by their interactions with the animal's choice in the current trial. To determine whether any of these effects could be attributed to systematic variability in eye movements, the above analysis was repeated using the residuals from a regression model in which the neural activity related to a set of eye movement parameters was factored out (Methods). The results were nearly identical, with the only difference found in the loss of significance for the effect of the current choice. During the fore-period, 35% of neurons showed a significant effect of the animal's choice in the previous trial on the residuals from the same regression model.

It is possible that the animal's choice in the previous trial influenced the activity of this neuron during the next trial through systematic changes in unidentified sensorimotor events, such as licking or eye movements during the inter-trial interval, that were not experimentally controlled. This was tested by comparing the activity of the same neuron in the search and free-choice trials. For the neuron shown in **Figure 4**, activity during search trials was significantly affected by the position of the target in the previous trial only during the fore-period, and this effect was opposite to and significantly different from that found in the free-choice trials (4-way ANOVA, $P < 10^{-5}$). The raster plots show that this change occurred within a few trials after the animal switched from search to free-choice trials (**Fig. 4**). These results suggest that the effect of the animal's choice in the previous trial on the activity of this neuron did not merely reflect nonspecific sensorimotor events. In 17% of the neurons that showed a significant effect of the animal's previous choice during the fore-period, there was also a significant interaction between the task type (search vs. free-choice) and the animal's choice in the previous trial (**Fig. 3**). This indicates that signals related to the animal's past choice were actively maintained in the PFC according to the type of decision. It is unlikely that this was entirely due to an ongoing drift in the background activity (i.e., non-stationarity), as the control analysis performed on two successive blocks of free-choice trials did not produce a single case with the same effect during the fore-period (**Fig. 3**).

During the fore-period, 39% of neurons showed a significant effect of the reward in the previous trial. For example, the activity of the neu-

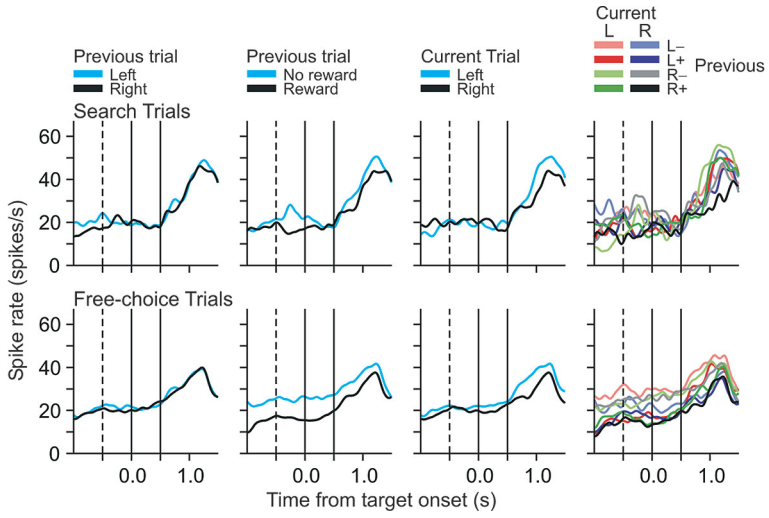

Figure 5 Example neuron showing a significant effect of the reward in the previous trial. Same format as in **Figure 4**.

ron in **Figure 5** was higher throughout the entire trial after the animal was not rewarded in the previous trial, compared to when the animal was rewarded. This effect was nearly unchanged when we removed the eye-movement related activity in a regression analysis, both in this single-neuron example and for the population as a whole. Overall, 37% of neurons showed the effect of the previous reward when the analysis was performed on the residuals from the regression model. The possibility that this effect was entirely due to uncontrolled sensorimotor events is also unlikely, as a substantial proportion of these neurons (21%) also showed a significant interaction between the task type and the previous reward during the fore-period (**Fig. 3**).

To update the value functions in a reinforcement learning model, signals related to the animal's choice and its outcome must be combined, because each variable alone does not specify how the value function of a particular target should be changed. Similarly, activity of the neurons in the PFC was often influenced by the conjunction of these two variables. In the neuron in **Figure 6**, for example, there was a gradual buildup of activity during the fore-period, but this occurred only when the animal had selected the right-hand target in the previous trial, and this choice was not rewarded. During the delay period, the activity of this neuron diverged to reflect the animal's choice in the current trial (**Fig. 6**, arrow). The same neuron showed markedly weaker activity during the search trials, suggesting that information coded in the activity of this neuron regarding the outcome of choosing a particular target was actively maintained in free-choice trials (**Fig. 6**). For the fore-period, 20% of the neurons showed significant interaction between the animal's choice and its outcome in the previous trial ($P < 0.05$; **Fig. 3**). Activity related to eye movements was not an important factor: 90% of these neurons showed the same effect in the residuals from the regression analysis that factored out the effects of eye movements. Furthermore, during the fore-period, 27% of the same neurons showed significant three-way interactions among task type, animal's choice in the previous trial and outcome of the previous trial. In contrast, the control analysis during the first two blocks of the free-choice task revealed such an effect only in 5% of the neurons (**Fig. 3**). These results indicate that signals related to the conjunction of the animal's previous decision and its outcome are processed differently in the PFC according to the type of decisions made by the animal.

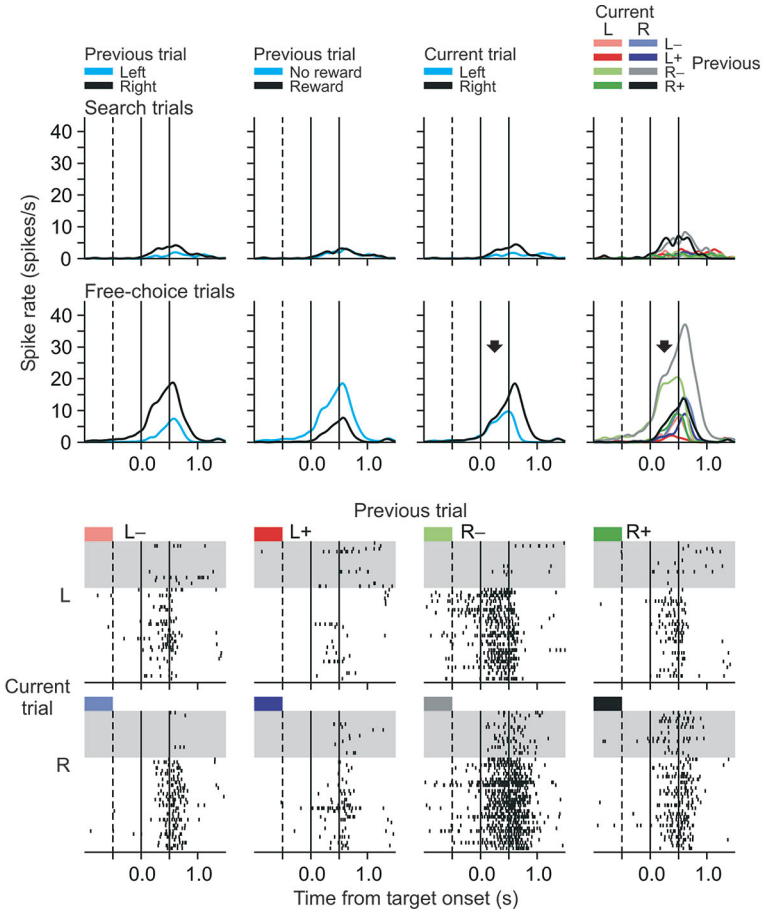

Figure 6 Example neuron with a significant interaction between the animal's choice and its outcome in the previous trial. Same format as in **Figure 4**. Arrows indicate the time when the animal's choice in the current trial was first reflected in the neural activity.

DISCUSSION

Interaction with other intelligent beings is fundamentally different from—and more complex than—dealing with inanimate objects[32,33]. Interactions with other animals are complicated by the fact that their behavioral strategies often change as a result of one's own behavior. Therefore, the analysis

of decision making in a group requires a more sophisticated analytical framework, which is provided by game theory. Matching pennies is a relatively simple zero-sum game that involves two players and two alternative choices. The present study examined the behavior of monkeys playing a competitive game similar to matching pennies against a computer opponent. It is not known whether monkeys treated this game as a competitive situation with another intentional being. Nevertheless, the same formal framework of game theory is applicable to the task used in this study, and as predicted, the animal's behavior was influenced by the opponent's strategy. When the computer blindly played the equilibrium strategy regardless of the animal's behavior, the animals selected one of the targets more frequently. In contrast, when the computer opponent began exploiting biases in the animal's choice sequence, the animal's behavior approached the equilibrium strategy. Furthermore, when the computer did not examine the animal's reward history (algorithm 1), the animals achieved a nearly optimal reward rate by adopting the win-stay-lose-switch (WSLS) strategy. This was possible because this strategy was not detected by the computer. Finally, the frequency of the WSLS strategy was reduced when the computer began exploiting biases in the animal's choice and reward sequences (algorithm 2).

These results also suggest that the animals approximated the optimal strategy using a reinforcement learning algorithm. This model assumes that the animals base their decisions, in part, on the estimates of expected rewards for the two targets and tend to select the target with larger expected reward. During zero-sum games such as matching pennies, strategies of the players behaving according to some reinforcement learning algorithms would gradually converge on a set of equilibrium strategies[5–7]. However, it is important to update the value functions of different targets by a small amount after each play when playing against a fully informed rational player (such as algorithm 2 in the present study). This is because large, predictable changes in the value functions would reveal one's next choice to the opponent. In the present study, the magnitude of changes in the value function varied according to the strategy of the opponent and was adjusted through the animal's experience.

Finally, neurophysiological recordings in the PFC revealed a potential neural basis for updating the value functions adaptively while interacting with a rational opponent. Reward-related activity is widespread in the brain[34–38]. In particular, signals related to expected reward (*i.e.*, value functions) are present in various brain areas[39–43], including the DLPFC[44–48]. Our results showed that neurons in the DLPFC also code signals related to the animal's choice in the previous trial. Such signals might be actively maintained and processed differently in the DLPFC according to the type of information required for the animal's upcoming decisions. Furthermore, signals related to the animal's past choices and their outcomes are combined at the level of individual PFC neurons. These signals might then be temporally integrated according to a reinforcement learning algorithm to update the value functions for alternative actions. Many neurons in the PFC show persistent activity during a working memory task, and the same circuitry might be ideally suited for temporal integration of signals related to the animal's previous choice and its outcome[49]. Although the present study examined the animal's choice behavior in a competitive game, reinforcement learning algorithms can converge on optimal solutions for a wide range of decision-making problems in dynamic environments. Therefore, the results from the present study suggest that the PFC has an important role in optimizing decision-making strategies in a dynamic environment that may include multiple agents.

METHODS

Animal preparations. Two male rhesus monkeys were used. Their eye movements were monitored at a sampling rate of 250 Hz with either a scleral eye coil or a high-speed video-based eye tracker (ET49, Thomas Recording). All the procedures used in this study conformed to National Institutes of Health guidelines and were approved by the University of Rochester Committee on Animal Research.

Behavioral task. Monkeys were trained to play a competitive game analogous to matching pennies against a computer in an oculomotor free-choice task (**Fig. 1a**). During a 0.5-s fore-period, they fixated a small yellow square (0.9 × 0.9°; 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. The central target disappeared after a 0.5-s delay period, and the animal was required to shift its gaze to one of the targets. At the end of a 0.5-s hold period, 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, and the animal maintained its fixation for another 0.2 s. The animal was rewarded at the end of this second hold period, but only if it selected the same target as the computer. The computer had been programmed to exploit certain biases displayed by the animal in making its choices. Each neuron was also tested in a visual search task. This task was identical to the free-choice task, except that one of the targets in the free-choice task was replaced by a distractor (red disk). The animal was required to shift its gaze toward the remaining target (green disk), and this was rewarded randomly with 50% probability. This made it possible to examine the effect of reward on the neural activity. The location of the target was selected from the two alternative locations pseudo-randomly for each search trial.

Algorithms for computer opponent. During the free-choice task, the computer selected its target according to one of three different algorithms. In algorithm 0, the computer selected the two targets randomly with equal probabilities, which corresponds to the Nash equilibrium in the matching pennies game. In algorithm 1, the computer exploited any systematic bias in the animal's choice sequence to minimize the animal's reward rate. The computer saved the entire history of the animal's choices in a given session, and used this information to predict the animal's next choice by testing a set of hypotheses. First, the conditional probabilities of choosing each target given the animal's choices in the preceding n trials ($n = 0$ to 4) were estimated. Next, each of these conditional probabilities was tested against the hypothesis that the animal had chosen both targets with equal probabilities. When none of these hypotheses was rejected, the computer selected each target randomly with 50% probability, as in algorithm 0. Otherwise, the computer biased its selection according to the probability with the largest deviation from 0.5 that was statistically significant (binomial test, $P < 0.05$). For example, if the animal chose the right-hand target with 80% probability, the computer selected the left-hand target with the same probability. Therefore, to maximize reward, animals needed to choose both targets with equal frequency and select a target on each trial independently from previous choices. In algorithm 2, the computer exploited any systematic bias in the animal's choice and reward sequences. In addition to the hypotheses tested in algorithm 1, algorithm 2 also tested the hypothesis that the animal's decisions were independent of prior choices and their payoffs in the preceding n trials ($n = 1$ to 4). Thus, to maximize total reward in algorithm 2, it was necessary for the animal to choose both targets with equal frequency and to make choices independently from previous choices and payoffs.

Neurophysiological recording. Single-unit activity was recorded from the neurons in the DLPFC of two monkeys using a five-channel multi-electrode recording system (Thomas Recording). The placement of the recording chamber was guided by magnetic resonance images, and this was confirmed in one animal by metal pins inserted in known anatomical locations. In addition, the frontal eye field (FEF) was defined in both animals as the area in which saccades were evoked by electrical stimulations with currents <50 μA (ref. 50). All the neurons described in the present study were anterior to the FEF.

Analysis of behavioral data. The frequency of a behavioral event (*e.g.*, reward) was examined with the corresponding probability averaged across recording

sessions and its standard deviation. The values of mutual information were corrected for the finite sample size. Null hypotheses in the analysis of behavioral data were tested using a binomial test or a *t*-test ($P < 0.05$). Parameters in the reinforcement learning model were estimated by a maximum likelihood procedure, using a function minimization algorithm in Matlab (Mathworks Inc.), and confidence intervals were estimated by profile likelihood intervals[31].

Analysis of neural data. Spikes during a series of 500-ms bins were counted separately for each trial. The effects of the animal's choice (P) and reward (R) in the previous trial and the choice in the current trial (C) were analyzed in a 3-way ANOVA ($P \times R \times C$). The effect of the task (search versus free-choice) was analyzed in a four-way ANOVA ($Task \times P \times R \times C$). As a control analysis to determine whether the task effect was due to non-stationarity in neural activity, the same four-way ANOVA was performed for the first two successive blocks of 128 trials in the free-choice task (**Fig. 3**). To determine whether eye movements were confounding factors, the above analysis was repeated using the residuals from the following regression model:

$$S = a_0 + a_1 X_{\mathrm{pre80}} + a_2 Y_{\mathrm{pre80}} + a_3 X_{\mathrm{FP}} + a_4 Y_{\mathrm{FP}} + a_5 X_{\mathrm{SV}} + a_6 Y_{\mathrm{SV}} + a_7 \mathrm{SRT} + a_8 \mathrm{PV} + \varepsilon$$

where S indicates the spike count, X_{pre80} (Y_{pre80}) the horizontal (vertical) eye position 80 ms before the onset of central fixation target, X_{FP} (Y_{FP}) the average horizontal (vertical) eye position during the fore-period, X_{SV} (Y_{SV}) the horizontal (vertical) component of the saccade directed to the target, SRT and PV the saccadic reaction time and the peak velocity of the saccade, and ε the error term.

ACKNOWLEDGMENTS

We thank L. Carr, R. Farrell, B. McGreevy and T. Twietmeyer for their technical assistance, J. Swan-Stone for programming, X.-J. Wang for discussions, and B. Averbeck and J. Malpeli for critically reading the manuscript. This work was supported by the James S. McDonnell Foundation and the National Institutes of Health (NS44270 and EY01319).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 23 December 2003; accepted 12 February 2004
Published online at http://www.nature.com/natureneuroscience/

1. Von Neumann J. & Morgenstern O. *Theory of Games and Economic Behavior* (Princeton Univ. Press, New Jersey, 1944).
2. Fudenberg, D. & Tirole, J. *Game Theory* (MIT Press, Cambridge, Massachusetts, 1991).
3. Dixit, A. & Skeath, S. *Games of Strategy* (Norton, New York, 1999).
4. Glimcher, P.W. *Decisions, Uncertainty, and the Brain* (MIT Press, Cambridge, Massachusetts, 2003).
5. Nash, J.F. Equilibrium points in *n*-person games. *Proc. Natl. Acad. Sci. USA* **36**, 48–49 (1950).
6. Robinson, J. An iterative method of solving a game. *Ann. Math.* **54**, 296–301 (1951).
7. Binmore, K., Swierzbinski, J. & Proulx, C. Does minimax work? An experimental study. *Econ. J.* **111**, 445–464 (2001).
8. Touhey, J.C. Decision processes, expectations, and adoption of strategies in zero-sum games. *Hum. Relat.* **27**, 813–824 (1974).
9. O'Neill, B. Nonmetric test of the minimax theory of two-person zerosum games. *Proc. Natl. Acad. Sci. USA* **84**, 2106–2109 (1987).
10. Brown, J.N. & Rosenthal, R.W. Testing the minimax hypothesis: a re-examination of O'Neill's game experiment. *Econometrica* **58**, 1065–1085 (1990).
11. Rapoport, A. & Boebel, R.B. Mixed strategies in strictly competitive games: a further test of the minimax hypothesis. *Games Econ. Behav.* **4**, 261–283 (1992).
12. Rapoport, A. & Budescu, D.V. Generation of random series in two-person strictly competitive games. *J. Exp. Psychol. Gen.* **121**, 352–363 (1992).
13. Budescu, D.V. & Rapoport, A. Subjective randomization in one- and two-person games. *J. Behav. Dec. Making* **7**, 261–278 (1994).
14. Ochs, J. Games with unique, mixed strategy equilibria: an experimental study. *Games Econ. Behav.* **10**, 202–217 (1995).
15. Sarin, R. & Vahid, F. Predicting how people play games: a simple dynamic model of choice. *Games Econ. Behav.* **34**, 104–122 (2001).
16. Shachat, J.M. Mixed strategy play and the minimax hypothesis. *J. Econ. Theory* **104**, 189–226 (2002).
17. Walker, M. & Wooders, J. Minimax play at Wimbledon. *Am. Econ. Rev.* **91**, 1521–1538 (2001).
18. Fudenberg, D. & Levine, D.K. *Theory of Learning in Games* (MIT Press, Cambridge, Massachusetts, 1998).
19. Mookherjee, D. & Sopher, B. Learning behavior in an experimental matching pennies game. *Games Econ. Behav.* **7**, 62–91 (1994).
20. Erev, I. & Roth, A.E. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**, 848–881 (1998).
21. Camerer, C.F. *Behavioral Game Theory* (Princeton Univ. Press, Princeton, New Jersey, 2003).
22. Simon, H.A. *Models of Man* (Wiley, New York, 1957).
23. Kahneman, D., Slovic, P. & Tversky, A. *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge Univ. Press, Cambridge, UK, 1982).
24. O'Neill, B. Comments on Brown and Rosenthal's reexamination. *Econometrica* **59**, 503–507 (1991).
25. Rapoport, A. & Budescu, D.V. Randomization in individual choice behavior. *Psychol. Rev.* **104**, 603–617 (1997).
26. Chen, H-C., Friedman, J.W. & Thisse, J-F. Boundedly rational Nash equilibrium: a probabilistic choice approach. *Games Econ. Behav.* **18**, 32–54 (1997).
27. Rubinstein, A. *Modeling Bounded Rationality* (MIT Press, Cambridge, Massachusetts, 1998).
28. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: an Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
29. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning. *Machine Learning: Proc. 11th Int. Conf.* pp. 157–163 (Morgan Kaufmann, San Francisco, California, 1994)
30. Christensen, R. *Log-linear Models and Logistic Regression* edn. 2 (Springer-Verlag, New York, 1997).
31. Burnham, K.P. & Anderson, D.R. *Model Selection and Multimodel Inference* 2nd edn. (Springer-Verlag, New York, 2002).
32. Byrne, R.W. & Whiten, A. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans* (Oxford Univ. Press, Oxford, UK, 1988).
33. Whiten, A. & Byrne, R.W. *Machiavellian Intelligence II: Extensions and Evaluations* (Cambridge Univ. Press, Cambridge, UK, 1997).
34. Schultz, W. Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27 (1998).
35. Rolls, E.T. *Brain and Emotion* (Oxford Univ. Press, Oxford, UK, 1999).
36. Amador, N., Schlag-Rey, M. & Schlag, J. Reward-predicting and reward detecting neuronal activity in the primate supplementary eye field. *J. Neurophysiol.* **84**, 2166–2170 (2000).
37. Stuphorn, V., Taylor, T.L. & Schall, J.D. Performance monitoring by the supplementary eye field. *Nature* **408**, 857–860 (2000).
38. Ito, S., Stuphorn, V., Brown, J.W. & Schall, J.D. Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science* **302**, 120–122 (2003).
39. Kawagoe, R., Takikawa, Y. & Hikosaka, O. Expectation of reward modulates cognitive signals in the basal ganglia. *Nat. Neurosci.* **1**, 411–416 (1998).
40. Platt, M.L. & Glimcher, P.W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
41. Shidara, M. & Richmond, B.J. Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science* **296**, 1709–1711 (2002).
42. Ikeda, T. & Hikosaka, O. Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron* **39**, 693–700 (2003).
43. McCoy, A.N., Crowley, J.C., Haghighian, G., Dean, H.L. & Platt, M.L. Saccade reward signals in posterior cingulate cortex. *Neuron* **40**, 1031–1040 (2003).
44. Watanabe, M. Reward expectancy in primate prefrontal cortex. *Nature* **382**, 629–632 (1996).
45. Leon, M.I. & Shadlen, M.N. Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron* **24**, 415–425 (1999).
46. Kobayashi, S., Lauwereyns, J., Koizumi, M., Sakagami, M. & Hikosaka, O. Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *J. Neurophysiol.* **87**, 1488–1498 (2002).
47. Roesch, M.R. & Olson, C.R. Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *J. Neurophysiol.* **90**, 1766–1789 (2003).
48. Tsujimoto, S. & Sawaguchi, T. Neuronal representation of response outcome in the primate prefrontal cortex. *Cereb. Cortex* **14**, 47–55 (2004).
49. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).
50. Bruce, C.J., Goldberg, M.E., Bushnell, M.C. & Stanton, G.B. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J. Neurophysiol.* **54**, 714–734 (1985).