PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 3845–3857
doi:10.1098/rstb.2008.0158
Published online 1 October 2008

Cortical mechanisms for reinforcement learning in competitive games

Hyojung Seo and Daeyeol Lee*

*Department of Neurobiology, Yale University School of Medicine, 333 Cedar Street,
SHM B404, New Haven, CT 06510, USA*

Game theory analyses optimal strategies for multiple decision makers interacting in a social group. However, the behaviours of individual humans and animals often deviate systematically from the optimal strategies described by game theory. The behaviours of rhesus monkeys (*Macaca mulatta*) in simple zero-sum games showed similar patterns, but their departures from the optimal strategies were well accounted for by a simple reinforcement-learning algorithm. During a computer-simulated zero-sum game, neurons in the dorsolateral prefrontal cortex often encoded the previous choices of the animal and its opponent as well as the animal's reward history. By contrast, the neurons in the anterior cingulate cortex predominantly encoded the animal's reward history. Using simple competitive games, therefore, we have demonstrated functional specialization between different areas of the primate frontal cortex involved in outcome monitoring and action selection. Temporally extended signals related to the animal's previous choices might facilitate the association between choices and their delayed outcomes, whereas information about the choices of the opponent might be used to estimate the reward expected from a particular action. Finally, signals related to the reward history might be used to monitor the overall success of the animal's current decision-making strategy.

Keywords: prefrontal cortex; decision making; reward

1. INTRODUCTION

In *Theory of Games and Economic Behaviour* published in 1944, von Neumann & Morgenstern made two fundamental contributions to economics. First, they introduced an axiomatic expected utility theory and provided a set of conditions that are necessary and sufficient to describe the preference of a decision maker among arbitrary choices using a set of numbers referred to as utilities. The theory, for example, assumes that the preference is transitive. In other words, if A is preferred to B and B is preferred to C, this implies that A is preferred to C. It also assumes that the preference between the two options is unaffected when a third option is combined with each of the first two options with the same probability. When these assumptions are satisfied, the entire preference relationship between all available options can be summarized by a utility function so that a particular option is preferred to another option if and only if the utility of the former is greater than the utility of the latter. This implies that the act of choosing a particular option can be characterized as the process of utility maximization, and therefore such choice behaviours are considered rational.

Second, having justified the use of utility function, von Neumann & Morgenstern (1944) then focused on the question of social decision making and created game theory. For animals living in a social group, such as humans and many other non-human primates, the outcomes of their choices are determined not just by the individual's own action, but by the combined actions of all animals interacting in the same group. Assuming that each decision maker or player in the group is rational and hence maximizes the individual's own self-interest as expressed by the utility function, game theory seeks to find an optimal strategy that would be taken by such a rational player.

In game theory, a game can be defined by a pay-off matrix that specifies the utility of an outcome for each player according to the choices of all players in the group. The complexity of a game would increase, of course, with the number of players and the number of choices available to each player. Therefore, the simplest non-trivial game would consist of two players each with two alternative choices. A game is referred to as a zero sum, when the sum of the pay-offs given to all players is zero for all possible outcomes. For example, the game described by the pay-offs shown in figure 1*a*, known as the matching pennies, is a zero-sum game. In this example, the two players are a monkey and its computer opponent (Barraclough *et al.* 2004; Seo & Lee 2007; Seo *et al.* 2007). Each row corresponds to a particular choice available to the monkey, and each column to a particular choice available to the computer opponent. A pair of numbers within each cell of this matrix then specifies the pay-offs given to the two players. For example, if both players choose the rightward target, then the monkey will earn the pay-off of 1 (e.g. one drop of juice) and the computer will lose the same amount (of virtual juice). In the standard matching pennies game, both players earn and lose the same amount of pay-offs for winning and losing, respectively. To avoid having to extract juice from the animal, we changed the pay-off matrix so that when the animal loses, the pay-offs to both players are zero.

[*] Author for correspondence (daeyeol.lee@yale.edu).

One contribution of 10 to a Theme Issue 'Neuroeconomics'.

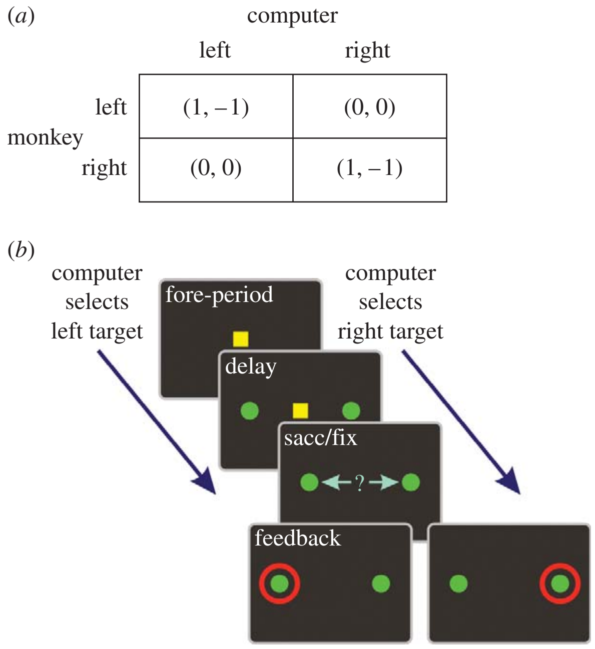

Figure 1. In (*a*) Pay-off matrix for the matching pennies game. The two numbers within each parenthesis corresponds to the pay-offs to the animal and the computer opponent, respectively. (*b*) Spatio-temporal sequence of the matching pennies task.

A strategy in game theory is defined as a probability distribution over a set of alternative actions, and an optimal strategy is the one that gives the maximum expected pay-off possible. A strategy that assigns a non-zero probability to only one action and therefore chooses that action exclusively is referred to as a pure strategy. Otherwise, a strategy is referred to as mixed. In games, the pay-off expected from a particular action for a given player changes according to the choices of other players. When the choices of all the other players are fixed, one or more actions that provide the maximum pay-off to a given player is referred to as a best response. If we assume that all players are rational and try to maximize their pay-offs in response to the actions chosen by all other players, a set of such players would play according to a set of strategies in which the strategy of each player is a best response to the strategies of all other players. This is referred to as Nash (1950) equilibrium. By definition, it is not possible for any player to increase his or her pay-off by deviating individually from a Nash equilibrium. Therefore, assuming that all players are rational, a strategy can be considered optimal for a particular player, if it is a part of a Nash equilibrium. However, such a Nash-equilibrium strategy may not be optimal once some players deviate from the Nash equilibrium.

When the Nash equilibrium for a given game includes a mixed strategy, such games are referred to as mixed-strategy games. For example, the matching pennies game illustrated in figure 1*a* is a mixed-strategy game. To understand this, imagine that the monkey adopts the pure strategy of always choosing the leftward target. Then, the computer opponent simulating a rational agent and therefore trying to maximize its own pay-off would always choose the rightward target, giving rise to the pay-off of 0 to both players. This outcome is not optimal for the animal, since it would be able to increase its pay-off, for example, by choosing the leftward and rightward targets each with a 0.5 probability. With this strategy, the animal would receive the average pay-off of 0.5, not just when the computer chooses either target exclusively, but for any strategy that could be chosen by the computer. Indeed, the strategy to choose the two targets equally often with the probability of 0.5 is the optimal strategy for the monkey, and any other strategy can be potentially exploited by the computer opponent. By the same token, the optimal strategy for the computer is also to choose the two targets with equal probabilities, and these two strategies comprise the Nash equilibrium for the matching pennies game.

Despite such clear predictions from game theory, the choice behaviour of human subjects frequently shows systematic deviations from Nash equilibrium (Camerer 2003). Even for relatively simple two-player mixed-strategy games, such as the matching pennies, human subjects do not converge on Nash equilibrium, and show significant correlation between successive choices, although such a pattern can be potentially exploited by their opponents (Budescu & Rapoport 1994; Mookherjee & Sopher 1994, 1997; Erev & Roth 1998). The results from these studies suggest that human subjects might use certain learning algorithms to improve their decision-making strategies and approximate optimal strategies successively (Lee 2008).

It is possible that the learning algorithms adopted by human subjects during repeated games might also be used by other non-human primates. If so, this would also provide an excellent opportunity to investigate the neural mechanisms for such learning-related processes at work for social decision making. Therefore, we examined whether and how the choice behaviour of rhesus monkeys (*Macaca mulatta*) deviates systematically from a Nash equilibrium during computer-simulated zero-sum games. In this paper, we first summarize the results from these behavioural studies showing that similar to human subjects, monkeys showed systematic biases in their choice sequences that can be accounted for by a relatively simple reinforcement-learning algorithm (Lee *et al.* 2004, 2005). We then describe the findings from neurophysiological experiments conducted in monkeys performing the matching pennies task. We found that neurons in the dorsolateral prefrontal cortex (DLPFC) often encoded signals related to the previous choices of the animal and the computer opponent as well as the animal's reward history (Barraclough *et al.* 2004; Seo *et al.* 2007). By contrast, neurons in the anterior cingulate cortex largely encoded the animal's reward history (Seo & Lee 2007). Finally, we discuss how these various signals might be used to approximate optimal strategies during dynamic decision making in competitive games.

2. REINFORCEMENT LEARNING AND DECISION MAKING

According to the law of effect (Thorndike 1911), the behaviours followed by pleasant outcomes are more likely to recur, whereas the opposite is true for the behaviours followed by aversive outcomes. This suggests that the animal's behaviour can be understood as the product of maximizing pleasant outcomes and minimizing aversive outcomes, as in reinforcement-learning theory (Sutton & Barto 1998). In reinforcement learning, a value function refers to the animal's subjective estimate for the sum of future rewards. Future rewards are often weighted exponentially according to their delays, consistent with the observation that humans and animals often prefer more immediate rewards than delayed ones (McClure *et al.* 2004; Kable & Glimcher 2007; Sohn & Lee 2007; Kim *et al.* 2008). For the matching pennies task used in our study (figure 1), the value function for choosing the leftward and rightward targets in trial t can be denoted as $Q_t(L)$ and $Q_t(R)$, respectively. Based on the value functions, the animal would then choose the rightward target in trial t with the probability given by the following softmax function (Sutton & Barto 1998; Lee *et al.* 2004):

$$P_t(R) = \exp\{\beta Q_t(R)\}/[\exp\{\beta Q_t(L)\} + \exp\{\beta Q_t(R)\}], \tag{2.1}$$

where β, referred to as the inverse temperature in analogy to thermodynamics, determines the randomness of the animal's choices. The probability that the animal would choose the leftward target in the same trial would be $1-P_t(R)$. Thus, the probability that the animal would choose the rightward target increases gradually as the value function for the rightward target increases relative to the value function for the leftward target. A large inverse temperature implies that the animal chooses the target with the higher value function more or less deterministically, whereas a small inverse temperature indicates a relatively stochastic choice behaviour. For example, as β approaches zero, the animal will choose the two targets randomly with equal probabilities, regardless of the value functions. The value functions are updated according to the difference between the reward received by the agent in trial t, R_t and the reward expected by the current value functions. In other words

$$Q_{t+1}(C_t) = Q_t(C_t) + \alpha[R_t - Q_t(C_t)], \tag{2.2}$$

where C_t ($=L$ or R) indicates the animal's choice in trial t and α corresponds to the learning rate. The value function was updated only for the target chosen by the animal in a given trial. This model has two free parameters, α and β, and they were estimated from the behavioural data using a maximum-likelihood procedure (Pawitan 2001; Seo & Lee 2007).

The concepts of value functions in reinforcement-learning theory and utilities in economics play analogous roles, since both of these quantities dictate the decision-maker's choices. Nevertheless, there are some differences. For example, expected utility theory focuses on laying axiomatic foundations for the relationship between utility functions and preferences, and therefore pays little attention to the rules dictating how the utility functions may change through the decision-maker's experience. By contrast, reinforcement-learning theory assumes that the reward signals can be easily obtained from the decision-maker's environment, and primarily deals with the computational algorithms that can efficiently discover a particular course of actions to maximize the future rewards through experience. Therefore, these two approaches are complementary. If the decision maker has full knowledge of his or her environment and sufficient cognitive capacity, formalism provided by the expected utility theory might provide an accurate description for the psychological process of decision making. Such ideal situations, however, may be relatively infrequent and, therefore, humans and animals may have to resort frequently to the solutions described by reinforcement-learning theory.

3. CHOICE BEHAVIOUR OF MONKEYS DURING COMPETITIVE GAMES

We investigated how the choice behaviour of rhesus monkeys changes dynamically during the matching pennies game (Barraclough *et al.* 2004; Lee *et al.* 2004). Three rhesus monkeys (C, E and F) underwent extensive behavioural testing. At the beginning of each trial, the animal first fixated a small yellow square that appeared in the centre of a computer screen (figure 1*b*). After a 0.5 s foreperiod, two identical green peripheral targets were presented along the horizontal meridian, and the animal was required to maintain its fixation on the central target until this was extinguished 0.5 s later. Then the animal was required to shift its gaze towards one of the peripheral targets within 1 s. The computer opponent chose its target at the beginning of each trial, and presented a red ring around its chosen target 0.5 s after the animal shifted its gaze towards one of the targets. If the animal chose the same target as the computer, it was rewarded with a small drop of juice 0.5 s after the onset of the feedback ring. Otherwise, the animal did not receive any reward in that trial (figure 1*a*).

Initially, each animal was tested while the computer opponent blindly adopted the Nash equilibrium and chose the two targets randomly and equally frequently. This condition was referred to as algorithm 0, and was used to examine the intrinsic bias that might be displayed by the animal, because in this condition, the expected pay-offs from the two targets were equal and independent of the animal's choice. Not surprisingly, during this period, the probability that the animal would choose the rightward target varied idiosyncratically across the animals. Each of the three monkeys tested with algorithm 0 chose the rightward target in 70, 90.2 and 33.2 per cent of the trials. Despite this variability in how often the animals chose each of the two targets, all of them were rewarded roughly in 50 per cent of the trials (49.7, 49.9 and 48.9% for monkeys C, E and F, respectively), since the computer selected its targets randomly regardless of the animal's choice behaviours.

In the next stage of the experiment, the computer opponent analysed the animal's choices in all the previous trials of a given session and tested whether the animal displayed any systematic biases in its choice sequence. When the computer detected a significant bias in the animal's choice history, this bias was exploited to predict the animal's choice in the next trial. This is referred to as algorithm 1. For example, the computer tested whether the animal has chosen the two targets equally frequently, whether the animal tended to alternate between the two targets in successive trials, whether the animal tended to switch to the rightward target after choosing the leftward target in two successive trials, and so forth. This was accomplished by testing a series of conditional probabilities that the animal would choose the rightward or leftward target given the choices made by the animal in the last n trials ($n=0$–4). The null hypothesis that each of these conditional probabilities was 0.5 was tested using the binomial test. If none of these null hypotheses was rejected at the significance level of 0.05, it was then assumed that the animal chose the two targets equally often and independently across successive trials, and the computer chose each target with the probability of 0.5. If one or more null hypotheses were rejected, then the computer biased its target according to the conditional probability that was significantly different from 0.5 and deviated maximally from 0.5. If the animal was expected to choose the rightward target, e.g. with the probability of 0.8, then the computer chose the leftward target with the probability of 0.8. Accordingly, in order to maximize its overall reward, the animal was required to choose the two targets equally often and independently across successive trials.

We found that once the computer opponent switched to algorithm 1, the choices of the animals were much more evenly distributed between the two targets, and became more independent across successive trials. The percentage of trials in which the animal chose the rightward target was 48.9, 51.1 and 49 in the three animals, respectively. In addition, the probability that the animal would choose the same target as in the previous trial was close to 0.5 (Lee *et al.* 2004). By contrast, the animal was more likely to choose the same target as in the previous trial after it was rewarded and to switch to the other target otherwise (figure 2). In the matching pennies game, this so-called win–stay lose–switch strategy is equivalent to the strategy of choosing the same target chosen by the computer opponent in the previous trial, since the animal was rewarded only when it chose the same target as the computer. Overall, the three animals tested with algorithm 1 chose their targets according to the win–stay lose–switch strategy in 64.6, 73.1 and 63.3 per cent of the trials, respectively. It should be noted that in algorithm 1, such a frequent use of the win–stay lose–switch strategy was not penalized, since the computer did not analyse the conditional probability of the animal's choice based on the animal's reward history. Therefore, despite the frequent use of the win–stay lose–switch strategy, the animal was rewarded in roughly half of the trials (48.9, 49.1 and 49.5% for monkeys C, E and F, respectively). Each animal was tested with algorithm 1 for several weeks (36, 63 and 26 days for monkeys C, E and F, respectively), and the animals performed on average approximately 1000 trials each day. Interestingly, during the entire period of algorithm 1, the probability that the animal would choose its target according to the win–stay lose–switch strategy gradually increased (figure 2), even though this was not accompanied by an increase in the reward rate. Thus, the animals increased the tendency to adopt the win–stay lose–switch strategy spontaneously (Lee *et al.* 2004). This suggests that the animals might have been more explorative and made their choices more randomly in the initial phase of the task. The fact that the frequency of such explorative behaviours decreased without any changes in the reward rate also suggests that such behaviours might be metabolically costly.

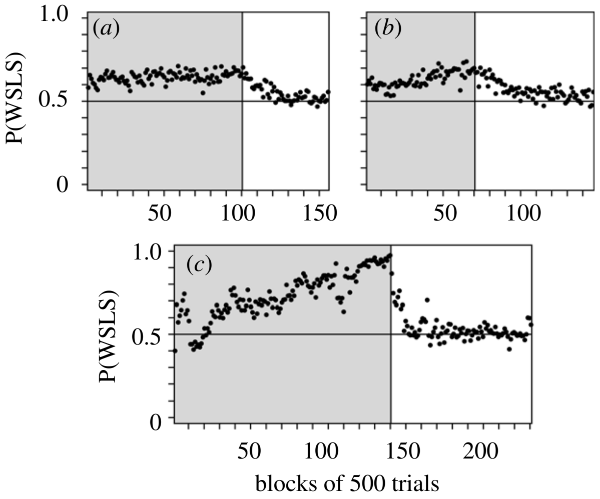

Figure 2. Probability of adopting the win–stay lose–switch strategy, P(WSLS), estimated for successive blocks of 500 trials in three different animals (monkeys (*a*) C, (*b*) F and (*c*) E). Grey and white backgrounds indicate the data obtained against the computer opponent programmed with algorithms 1 and 2, respectively.

A frequent use of the win–stay lose–switch strategy can be detrimental to the decision maker during a competitive game, such as the matching pennies, since it can be exploited by the opponent. Therefore, to test whether monkeys are capable of suppressing the win–stay lose–switch strategy during competitive games, we modified the algorithm used by the computer opponent so that the computer could exploit the win–stay lose–switch and other similar strategies used by the animal. In this so-called algorithm 2, the computer tested a series of conditional probabilities that the animal would choose a particular target given the animal's choices and reward outcomes in the last n trials ($n=1$–4). The computer tested the null hypothesis that each of these conditional probabilities as well as the conditional probabilities tested in algorithm 1 is all 0.5. Then it followed the same rule used in algorithm 1 to bias its choice when this null hypothesis was rejected. We found that once the computer opponent switched to algorithm 2 and began penalizing the frequent use of the win–stay lose–switch strategy, the animal gradually reduced the probability of using the win–stay lose–switch strategy (Lee *et al.* 2004). Overall, three monkeys tested with algorithm 2 chose their target according to the win–stay lose–switch strategy in 54.8, 53.5 and 56.5 per cent of the trials. Compared to the results obtained with algorithm 1, these values were closer to 50 per cent, but they were still significantly higher than 50 per cent. As a result, although the probability that the animal would be rewarded was relatively close to 0.5 (47.6, 47 and 47.8%, for monkeys C, E and F, respectively), it was significantly lower than values obtained for algorithm 1.

If an animal adjusts its strategy according to a reinforcement-learning algorithm, the value function for a given action would increase after the same action is rewarded, and the probability for adopting the win–stay lose–switch strategy would be relatively high. Therefore, a relatively frequent use of the win–stay lose–switch strategy during the matching pennies game suggests that the animals might have adjusted their decision-making strategies according to a reinforcement-learning algorithm. Moreover, the frequency of using the win–stay lose–switch strategy decreased dramatically when the computer opponent switched to algorithm 2. This might be accounted for by some changes in the parameters of a reinforcement-learning model. For example, the probability of using the win–stay lose–switch strategy would increase with the learning rate, because a small learning rate implies only small changes in the animal's strategy after each trial. Alternatively, the probability of using the win–stay lose–switch strategy can also increase with the inverse temperature, since this would reduce the animal's random choices. To distinguish between these two possibilities, we applied the reinforcement-learning model described above separately to the behavioural data obtained from each session. The results showed that the inverse temperature was significantly smaller for algorithm 2 than for algorithm 1 in two animals (monkeys E and F; paired *t*-test, $p<0.01$; figure 3). The difference in the learning rate was more robust, and became significantly smaller during the sessions tested with algorithm 2 in all three animals (paired *t*-test, $p<0.01$). Overall, these results suggest that depending on the strategies used by the computer opponent, the learning rate in the reinforcement learning and in some cases the inverse temperature that controlled the randomness of the animal's choices were adjusted. This might be driven by the process of meta-learning and controlled by long-term changes in the animal's reward probability (Schweighofer & Doya 2003; Soltani *et al.* 2006).

4. ENCODING OF VALUE FUNCTIONS IN THE FRONTAL CORTEX

A large proportion of the brain is devoted to the problem of decision making. In particular, numerous studies have identified signals related to various aspects of reward in many different brain regions. In many cases, such signals appear during the time when the animal is choosing between multiple alternative actions and planning a chosen action, and therefore might correspond to the expected utility or value function for the reward anticipated by the animal (Lee 2006). For example, neurons in the posterior parietal cortex often modulate their activity according to the likelihood that the animal would receive reward following an eye movement directed towards the neuron's receptive field (Platt & Glimcher 1999; Dorris & Glimcher 2004; Sugrue *et al.* 2004; Yang & Shadlen 2007). Similarly, neurons in the basal ganglia as well as the prefrontal cortex and the cingulate cortex often change their activity according to the magnitude, probability and immediacy of expected reward (Watanabe 1996; Hollerman *et al.* 1998; Kawagoe *et al.* 1998; Leon & Shadlen 1999; Kobayashi *et al.* 2002; Shidara & Richmond 2002; Roesch & Olson 2003; McCoy & Platt 2005; Samejima *et al.* 2005; Sohn & Lee 2007). These results suggest that the signals related to the expected utility and value function for the reward anticipated by the animal might be encoded in multiple areas of the brain. How these different areas contribute to specific aspects of decision making is currently an active area of research (Lee *et al.* 2007). For example, an important function of the medial frontal cortex, including the dorsal anterior cingulate cortex (ACCd) and supplementary motor area, might be to integrate the information about the costs and benefits of particular behaviours (Shidara & Richmond 2002; Sohn & Lee 2007; Rushworth *et al.* 2007). In addition, it has been proposed that the ACCd might play a more important role in selecting an action voluntarily and monitoring its outcomes (Walton *et al.* 2004; Kennerley *et al.* 2006; Matsumoto *et al.* 2007; Quilodran *et al.* 2008), whereas the orbitofrontal cortex might be more involved in encoding the subjective value of reward expected from the animal's behaviours (Padoa-Schioppa & Assad 2006; Rushworth & Behrens 2008).

We investigated whether the neurons in the DLPFC and the ACCd modulate their activity according to the value functions during the matching pennies game. Activity was recorded extracellularly from 322 neurons in the DLPFC (Seo *et al.* 2007) and 154 neurons in the ACCd (Seo & Lee 2007). We then tested whether neuronal activity was related to the sum of the value functions associated with the two alternative targets or their difference, using the following regression model:

$$S_t = a_0 + a_1 C_t + a_2\{Q_t(L) + Q_t(R)\} + a_3\{Q_t(L) - Q_t(R)\}, \tag{4.1}$$

where S_t denotes the spike rate in a particular analysis window of trial t; C_t the animal's choice in trial t; $Q_t(L)$ and $Q_t(R)$ the value functions for the leftward and rightward targets, respectively, that were estimated on a trial-by-trial basis using the reinforcement-learning model described above; and $a_0 \sim a_3$ the regression coefficients. If the reinforcement-learning model described the animal's choice behaviour well, the learning rate (α) should be between 0 and 1 and the inverse temperature (β) should be larger than 0. Therefore, neurons were excluded from this analysis, if $\alpha<0$, $\alpha>1$ or $\beta<0$ for the behavioural data that were collected concurrently. As a result, 291 neurons in the DLPFC and 148 neurons in the ACCd were included in this analysis. The sum of value functions would provide the information about the overall reward rate, whereas the difference in the value functions would indicate which choice would be more desirable. Therefore, these two quantities were used in this regression model, rather than the value functions of individual targets (Seo & Lee 2007). For example, if the activity of a given neuron increases similarly with the value functions of both targets, this would largely influence the regression coefficient for the sum of the value functions, but not for the difference in the value functions. In addition, the difference in the value functions would be correlated with the animal's choice, and therefore, we included the animal's choice as a dummy variable in this

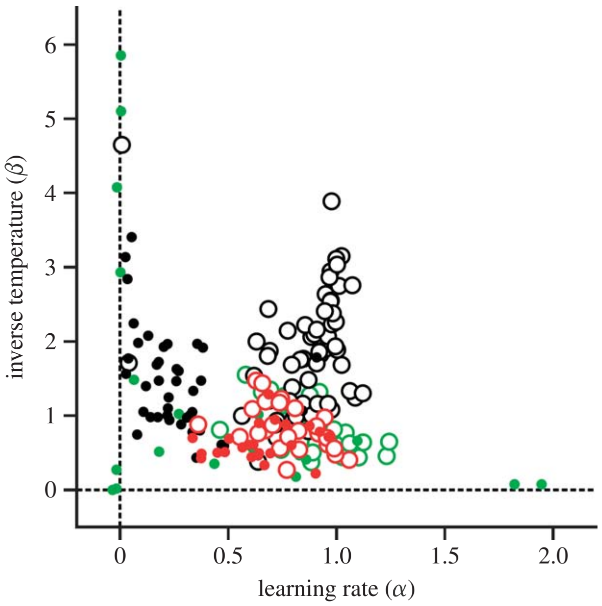

Figure 3. Model parameters for the reinforcement-learning model fit to the choice behaviour during the matching pennies task. A small number of cases ($n=10$ out of 230 sessions) in which the inverse temperature was unusually large (greater than 10) are not shown. Open and filled circles correspond to the results from the sessions tested with algorithms 1 and 2, respectively, and different colours indicate the results obtained from different animals (green, monkey C; black, monkey E; red, monkey F).

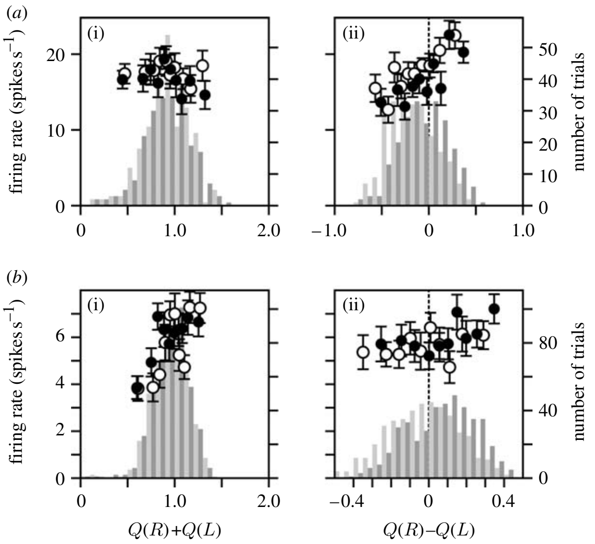

Figure 4. Example neurons recorded in the (*a*) DLPFC and (*b*) ACCd that significantly modulated their activity according to the value functions during the delay period in the matching pennies task. Histograms in (*a*(i),*b*(i)) show the distribution of trials as a function of the sum of the value functions, separately for the trials in which the animal chose the leftward (light grey) or rightward (dark grey) target. The histograms in (*a*(ii),*b*(ii)) show the distribution of the same trials as a function of the difference in the value functions. Neural activity is plotted separately according to whether the animal chose the leftward (open circles) or rightward (filled circles) target. Each symbol corresponds to the average activity in a decile of trials sorted according to the sum of the value functions (*a*(i),*b*(i)) or their difference (*a*(ii),*b*(ii)). Error bars, s.e.m.

regression in order to control for the neural activity directly related to the animal's choice.

We evaluated the statistical significance for each of the regression coefficients included in the above model using two different methods. First, we used a *t*-test to determine the *p*-value for each regression coefficient. Although this is the standard method to evaluate the statistical significance for regression coefficients, it may not be appropriate in the present application, because the value functions estimated for successive trials are not independent but correlated. Since this violates the independence assumption in the regression analysis, the statistical significance determined by a *t*-test is likely to be inflated. Second, to address this concern, we also performed a permutation test. In this method, we randomly shuffled the order of trials and recalculated the value functions according to the shuffled sequences of the animal's choices and rewards. We then recalculated the regression coefficients for the same regression model. This procedure was repeated 1000 times, and the *p*-value for each regression coefficient was given by the frequency of shuffles in which the magnitude of the original regression coefficient was exceeded by that of the regression coefficients obtained after shuffling.

A substantial proportion of the neurons in both the DLPFC and ACCd significantly modulated their activity according to the sum of the value functions, while others modulated their activity according to the difference of the value functions. For example, the DLPFC neuron shown in figure 4*a* modulated their activity according to the difference in the value functions, whereas the ACCd neuron shown in figure 4*b* changed their activity significantly according to the sum of the value functions. When the statistical significance was evaluated with the *t*-test, 33 and 34.5 per cent of the neurons in the DLPFC and ACCd, respectively, showed significant changes in their activity related to the sum of the value functions during the delay period. However, this percentage decreased significantly when the permutation test was used. Results from the permutation test showed that during the delay period, the percentage of neurons significantly modulating their activity according to the sum of the value function was 18.9 and 23.7 per cent for the DLPFC and ACCd, respectively (figure 5, ΣQ, black bars). This suggests that the neural activity related to the sum of value functions might be overestimated when it is tested with the *t*-test, presumably because the value functions for a given target in successive trials are correlated. The proportion of the neurons showing significant modulations related to the sum of the value functions was not significantly different for the DLPFC and ACCd (χ^2-test, $p>0.05$).

In both the DLPFC and ACCd, the proportion of neurons that modulated their activity according to the difference in the value functions for the two targets was lower than that for the sum of the value functions. For example, when examined with the *t*-test, 25.4 and 14.9 per cent of the neurons in the DLPFC and ACCd, respectively, showed significant modulations in their activity during the delay period according to the difference in the value functions. When examined with the permutation test, 13.1 per cent of the neurons

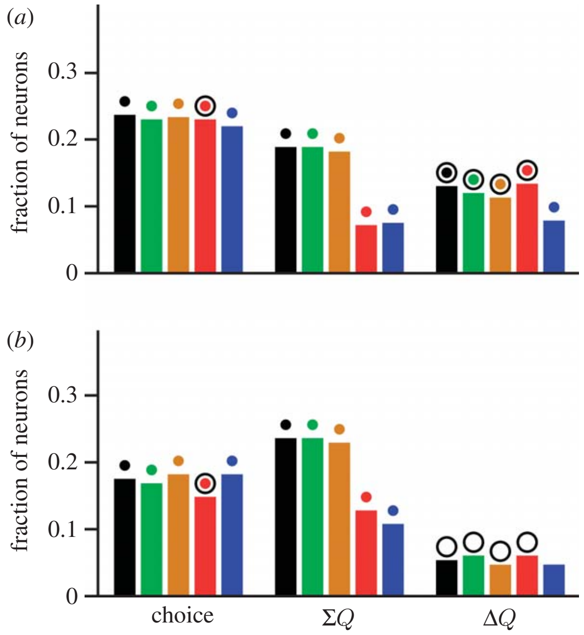

Figure 5. Fraction of neurons in the (*a*) DLPFC ($N=291$) and (*b*) ACCd ($N=148$) that significantly modulated their activity according to the animal's choice, the sum of the value functions for the two targets (ΣQ) and the difference in the value functions (ΔQ) during the delay period in the matching pennies task. The statistical significance of each regression coefficient was determined by a permutation test. Different colours correspond to the results obtained from different regression models. Base model includes only the animal's choice and the linear combinations (sum and difference) of the value functions. Filled circles above the bars indicate that the corresponding fraction is significantly larger than the significance level used ($p=0.05$; binomial test, $p<0.05$), whereas open circles indicate that the difference between the DLPFC and ACCd was significant (χ^2-test, $p<0.05$). Black, base (equation (4.1)); green, base + C (equation (4.2)); brown, base + P (equation (4.3)); red, base + R (equation (4.4)); blue, base $+C+P+$R, where C, P and R indicate the animal's choice, the computer's choice, and reward in the previous trial, respectively, in the regression model.

in the DLPFC showed significant changes in their activity related to the difference in the value functions (figure 5, ΔQ, black bars), and this was still significantly higher than expected by chance given the significance level used ($p=0.05$). By contrast, the proportion of the neurons showing the significant effect of the difference in the value function in the ACCd (5.4%) was not significantly higher than expected by chance. Therefore, there was little evidence for the signals related to the difference in the value functions in the ACCd. In addition, the proportion of neurons showing significant modulations in their activity related to the difference in the value functions was significantly higher in the DLPFC than in the ACCd (χ^2-test, $p<0.05$).

The regression model used in the above analysis provides useful insight into the nature of signals encoded in the activity of individual neurons in the DLPFC and ACCd. However, these results do not test directly the possibility that the neural activity in these two cortical areas encode the value functions only indirectly. In other words, the correlation between neural activity and value functions might be spurious and mediated by some other factors. For example, the value function for a given target is gradually adjusted according to the reward prediction errors (equation (2.2)), so their values in successive trials tend to be correlated. As a result, if the value function for the rightward target is larger than the value function for the leftward target in a given trial, the same is likely to be true in the next trial. The difference in the value functions for the two targets in a given trial may therefore be related to the animal's choice not only in the same trial but also in the previous trial. To test whether the neural activity related to the difference in the value functions might be due to the effect of the animal's previous choice, we included the animal's choice in the previous trial as an additional dummy variable in the regression model described above (equation (4.1)). In other words,

$$S_t = a_0 + a_1 C_t + a_2\{Q_t(L)+Q_t(R)\} + a_3\{Q_t(L)-Q_t(R)\} + a_4 C_{t-1}. \tag{4.2}$$

The results showed that the proportion of neurons that showed significant modulations in their activity related to the sum of the value functions and their difference was little affected by this change (figure 5, green bars). Therefore, the animal's choice in the previous trial did not have any major effect on the activity related to the difference in the value functions.

During the matching pennies game, the probability that the opponent would choose a particular target determines the pay-off expected from the same target. For example, if the computer frequently chooses the rightward target, this would increase the value function for the rightward target relative to the leftward target. Therefore, activity of neurons that change their activity according to the choice of the computer opponent may show significant correlation between its activity and the difference in the value functions for the two targets. This was tested by adding to the regression model a dummy variable corresponding to the choice of the computer opponent in the previous trial.

$$S_t = a_0 + a_1 C_t + a_2\{Q_t(L)+Q_t(R)\} + a_3\{Q_t(L)-Q_t(R)\} + a_4 P_{t-1}, \tag{4.3}$$

where P_t denotes the computer's choice in trial t. When the permutation test was used to evaluate the statistical significance for the regression coefficients in this model, 18.2 and 23 per cent of the neurons in the DLPFC and ACCd, respectively, showed significant modulations in their activity according to the sum of the value functions, whereas the corresponding percentages for the difference in the value functions were 11.3 and 4.7 per cent, respectively (figure 5, brown bars). Therefore, similar to the results from the model that included the animal's previous choice, adding the computer's previous choice did not have a major influence on the proportion of neurons encoding signals related to the value functions. This indicates that the activity related to the value functions in the DLPFC and ACCd did not result entirely from the choice of the computer opponent in the previous trial.

When the animal is rewarded during the matching pennies game, this produces a positive reward prediction error and therefore increases the value function for the target chosen by the animal. The value function of the unchosen target remains unchanged. Therefore, following a rewarded (unrewarded) trial, the sum of the value function would increase (decrease). Therefore, if a particular neuron tended to increase (decrease) its activity in a given trial after the animal was rewarded in the previous trial, then the activity of this neuron might be positively (negatively) correlated with the sum of the value functions. This raises the possibility that at least a part of the signals related to the sum of the value functions might arise from the signals related to the animal's reward in the previous trial. To test this, we included the animal's reward in the previous trial, R_{t-1}, in the regression model. Namely,

$$S_t = a_0 + a_1 C_t + a_2\{Q_t(L) + Q_t(R)\} + a_3\{Q_t(L) - Q_t(R)\} + a_4 R_{t-1}. \quad (4.4)$$

As expected, the proportion of the neurons modulating their activity according to the sum of the value functions decreased significantly when the animal's reward in the last trial was included in the regression model. This was true for both the DLPFC and ACCd (χ^2-test, $p<0.005$; figure 5, red bars). When the permutation test was used to evaluate the statistical significance, the proportion of such neurons in the DLPFC and ACCd was 7.2 and 12.8 per cent, respectively. This was still significantly higher than expected by chance for both areas (binomial test, $p<0.05$). However, it suggests that the neural activity in the DLPFC related to the sum of the value functions was largely due to the effect of the previous reward, whereas in the ACCd it might result from the animal's reward history extending beyond the last trial. Finally, we have also tested a regression model that includes the animal's choice, the choice of the computer opponent and the reward in the previous trial. The results from this model were similar to those obtained from the model that included only the reward in the previous trial in addition to the animal's choice and value functions in the current trial (figure 5, blue bars).

5. ENCODING OF CHOICES AND OUTCOMES IN THE FRONTAL CORTEX

The results described in §4 showed that the activity in the DLPFC and ACCd encodes the sum of the value functions, and that the difference in the value functions might be encoded by some neurons in the DLPFC. In addition, activity in both areas was still correlated with the sum of the value functions, even when the effect of the animal's reward in the previous trial was factored out. Since the sum of the value functions is estimated from the animal's reward history, this suggests that the activity in these cortical areas might be influenced by the reward received by the animal more than a trial before the current trial. The proportion of neurons in the DLPFC modulating their activity according to the difference in the value function was significantly higher than expected by chance and only weakly influenced when all the behavioural variables in the previous trial were included in the regression model. This suggests that neurons in the DLPFC might encode signals related to the animal's choice and reward for multiple trials in the past. To test this, and to further investigate the nature of signals encoded in the DLPFC and ACCd while avoiding the problems related to the serial correlation in the value functions, we applied a regression analysis in which the behavioural variables in the current and previous three trials were included as dummy variables. In other words,

$$S_t = a_0 + \boldsymbol{A_C}[C_t C_{t-1} C_{t-2} C_{t-3}]' + \boldsymbol{A_P}[P_t P_{t-1} P_{t-2} P_{t-3}]' + \boldsymbol{A_R}[R_t R_{t-1} R_{t-2} R_{t-3}]', \quad (5.1)$$

where C_t, P_t and R_t refer to the computer's choice, the choice of the computer opponent and the animal's reward in trial t, and $\boldsymbol{A_C}$, $\boldsymbol{A_P}$ and $\boldsymbol{A_R}$ are row vectors including the corresponding regression coefficients.

Neurons in both the DLPFC and ACCd often modulated their activity according to the animal's choice, the choice of the computer opponent and the animal's reward in current and previous trials (figure 6). The signals related to the animal's choice in the DLPFC gradually increased during the foreperiod and delay period before the animal shifted the gaze towards its chosen target. During the delay period, 19.9 per cent of the neurons in the DLPFC showed significant modulations in their activity according to the animal's upcoming choice (figure 6*a*, trial lag = 0). During the same period, many more neurons (39.8%) modulated their activity according to the animal's choice in the previous trial (figure 6*a*, trial lag = 1), and 11.2 per cent of the neurons changed their activity according to the animal's choice two trials before the current trial (figure 6*a*, trial lag = 2). During the foreperiod and delay periods, the activity of many DLPFC neurons was also affected by the choice of the computer opponent in the previous trial (figure 6*b*) as well as whether the animal was rewarded in the previous trial or not (figure 6*c*). During the delay period, 18 and 32.9 per cent of the neurons in the DLPFC showed significant modulations in their activity according to the computer's choice and reward in the previous trial, respectively. A significant proportion of neurons in the DLPFC changed their activity in relation to the reward received by the animal even two (10.9%) or three (7.1%) trials before the current trial.

Many neurons in the DLPFC modulated their activity according to more than one of these variables. An example neuron in the DLPFC showing the effect of multiple variables is shown in figure 7. This neuron increased its activity during the eye movement period after fixation target offset when the animal chose the rightward target more than when the animal chose the leftward target. In addition, the activity of this neuron during the delay period increased more when the animal chose the rightward target in the previous trial (figure 7*a*, trial lag = 1) and showed a slight but significant decrease when the animal had chosen the rightward target two trials before the current trial (figure 7*a*, trial lag = 2). The same neuron also increased its activity more when the computer opponent chose the rightward target compared to when the computer chose the leftward target, and this difference was maintained throughout the next trial (figure 7*b*, trial lag = 1). Finally, the activity of this neuron was reduced when the animal was rewarded in a given trial (figure 7*c*, trial lag = 0). When analysed with the regression model that included the linear combinations of value functions for the two targets, this neuron also showed a significant modulation in its activity according to the difference in the value functions for the two targets (figure 4*a*).

Compared to the DLPFC, neurons in the ACCd modulated their activity more frequently according to the reward received by the animal in the current or previous trials. An example neuron in the ACCd that showed the effect of rewards in previous trials is shown in figure 8. By contrast, activity in the ACCd was less frequently affected by the animal's choice or the choice of the computer opponent (figure 6). During the delay period, only 18.2 and 7.8 per cent of the ACCd neurons modulated their activity according to the animal's choice and the computer's choice in the previous trial, respectively. The proportion of neurons that changed their activity according to the animal's choice or the computer's choice two trials before the current trial was not significantly higher than the significance level used ($p = 0.05$). By contrast, during the delay period, 45.5, 18.2 and 11 per cent of the ACCd neurons changed their activity significantly according to whether the animal was rewarded or not in each of the last three trials, respectively. In addition, the proportion of the neurons that changed their activity during the feedback period according to whether the animal was rewarded or not in the same trial was significantly higher in the ACCd (81.8%) than in the DLPFC (68.90%; χ^2-test, $p < 0.05$). Therefore, although the signals related to the animal's choice and the computer's choice were more weakly represented in the ACCd than in the DLPFC, ACCd neurons showed more robust modulations in their activity according to the reward in the current and previous trials.

In summary, during the matching pennies game, neurons in the lateral (DLPFC) and medial (ACCd) frontal cortex represent at least three different types of signals that are related to the animal's choice and its outcome (i.e. reward) in addition to the choice of the computer opponent. In the DLPFC, signals related to the animal's choice and reward were both strongly represented, whereas in the ACCd, reward-related signals were dominant. In both areas, some neurons also encoded signals related to the choice of the computer opponent, but this was relatively weak. All of these signals decayed gradually over the course of two to four trials.

6. FUNCTIONAL SIGNIFICANCE OF SIGNALS RELATED TO CHOICE AND OUTCOME

The results described above indicate that signals related to the animal's choice, the computer's choice and reward persisted across several trials in the primate frontal cortex. Therefore, they might contribute to the process of monitoring the outcomes from previous choices and updating the animal's decision-making strategies accordingly. For example, the signals related to the animal's previous choices might be necessary to link the animal's particular action to its outcome, if the outcome of a particular choice is revealed only after a certain delay. Such memory signals are referred to as eligibility trace (Sutton & Barto 1998), and might be essential for determining how the value functions should be adjusted (Kim *et al.* 2007; Lau & Glimcher 2007; Seo *et al.* 2007). In this study, we have considered a relatively simple reinforcement-learning model in which only the value function for the action chosen in the current trial was updated according to its outcome. It remains to be seen whether the animal's choice behaviour during various decision-making tasks can be better accounted for by the model endowed with eligibility traces. Although the anatomical locus or loci in which the value functions are updated are not known, this requires a convergence of signals related to the value functions and reward prediction errors. This may occur in the DLPFC, since a significant number of neurons in the DLPFC encoded the value functions of alternative actions differentially. In addition, dopamine neurons that encode reward prediction errors (Schultz 1998) project to the DLPFC (Brown *et al.* 1979; Lewis *et al.* 2001). The presence of eligibility trace in the DLPFC raises the possibility that this might be used during the process of updating value functions in the DLPFC. Interestingly, the neural signals related to the animal's previous choices have been observed in the striatum even when the animal was required to choose its action according to a sensory stimulus (Kim *et al.* 2007; Lau & Glimcher 2007). In addition, striatal neurons receive dense projections from the dopamine neurons (Reynolds & Wickens 2002) and also encode the value functions for specific actions (Samejima *et al.* 2005). Therefore, it is possible that the striatum might also play a role in updating the value functions.

Compared to the ACCd, neurons in the DLPFC were more likely to encode the signals related to the computer's previous choices. These signals might play an important role in updating the animal's decision-making strategies when the task involves competitive interactions with other decision makers. For example, during the matching pennies game, the probability that the animal would be rewarded for choosing a particular target is equivalent to the probability that the same target would be chosen by the computer opponent. Therefore, the signals related to the previous choices of the opponent might be used to update the value functions of alternative actions. Finally, neurons in the DLPFC and ACCd commonly displayed modulations in their activity according to the animal's reward history. Signals related to the animal's previous rewards can provide some information about the local rate of reward, namely, how often the animal has been rewarded recently. This information can be then used as a reference point against which the desirability of reward in a particular trial is evaluated (Helson 1948; Flaherty 1982; Sutton & Barto 1998; Frederick & Loewenstein 1999). For example, the same reward might be considered more desirable and influences the animal's future behaviour more strongly when it was preceded by a number of unrewarded trials. Indeed,

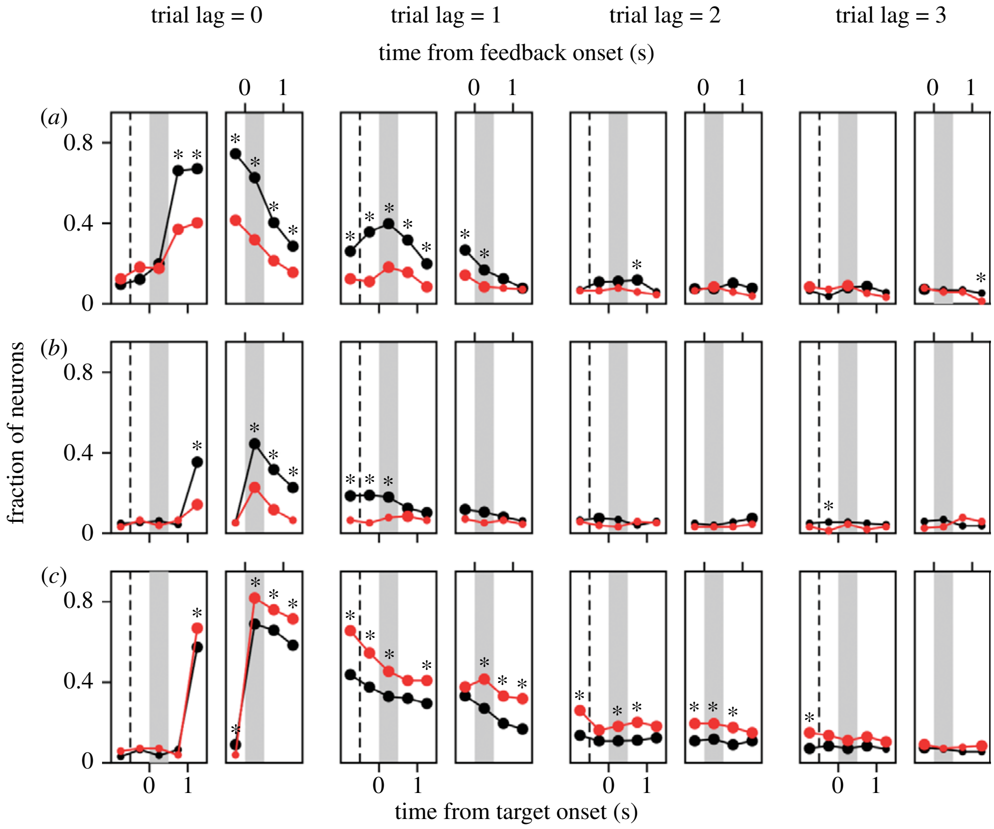

Figure 6. Time course of activity related to the (*a*) animal's choice, (*b*) the choice of the computer opponent and (*c*) reward in the population of neurons in the DLPFC (black) or ACCd (red). Each symbol indicates the fraction of neurons that displayed significant modulations in their activity according to the corresponding variable (*t*-test, $p<0.05$). These were estimated separately for different time windows using a linear regression model. Large circles indicate that the percentage of neurons was significantly higher than the significance level used in the regression analysis (binomial test, $p<0.05$). Asterisks indicate that the difference between the two cortical areas was statistically significant (χ^2-test, $p<0.05$). Dotted vertical lines in the left panels correspond to the onset of foreperiod. Grey background indicates the delay (left panels) or feedback (right panels) period.

reward-related signals in the ACCd were quite heterogeneous (Seo & Lee 2007), and this might reflect the processes of computing average reward rate and evaluating how the outcome of a particular choice deviates from this reference point. For example, ACCd neurons that increase (or decrease) their activity consistently according to the reward in the current and previous trials might encode the average reward rate. By contrast, some neurons in the ACC modulated their activity antagonistically in response to the reward in the current trial and those in the previous trials. Such neurons might signal the extent to which the reward in the current trial deviates from the local reward rate (Seo & Lee 2007; Matsumoto *et al.* 2007).

7. CONCLUSIONS

In the past several years, remarkable progress has been made in our understanding of neural substrates responsible for monitoring the consequences of voluntary actions and incorporating this information to update decision-making strategies. This progress was facilitated by the use of formal frameworks imported from such diverse disciplines as economics, psychology and machine learning. These frameworks provide the tools necessary to estimate the hidden variables, such as utility and value functions, that mediate the process of decision making (Corrado & Doya 2007). They also provide useful insights into the design of behavioural tasks necessary to identify specific neural substrates of decision making. In particular, a large number of experiments guided by game theory have probed the underlying neural processes involved in socially interactive decision making (Sanfey 2007; Fehr & Camerer 2007; Lee 2008). Some of these experiments focused on the neural correlates of altruistic preferences and cooperation (Rilling *et al.* 2002; Moll *et al.* 2006; Harbaugh *et al.* 2007), whereas others have found that some brain areas, such as the anterior paracingulate cortex, might be specialized in analysing the mental states of other decision makers (McCabe *et al.* 2001; Gallagher *et al.* 2002; Rilling *et al.* 2004*a*). When a group of decision makers have the opportunity to interact repeatedly, their strategies can be influenced by their previous experience. During this process, the neural circuitry involved in reinforcement learning plays an important role (Lee 2008). For example, the activity in the striatum reflects the outcomes of social decision making during the Prisoner's Dilemma game (Rilling *et al.* 2004*b*) and trust game (King-Casas *et al.* 2005).

We have investigated the behavioural choices of rhesus monkeys during a computer-simulated competitive game. Consistent with the findings from behavioural studies in humans, the animals used a relatively simple reinforcement-learning algorithm to arrive at a nearly optimal strategy during this task. In addition, we also

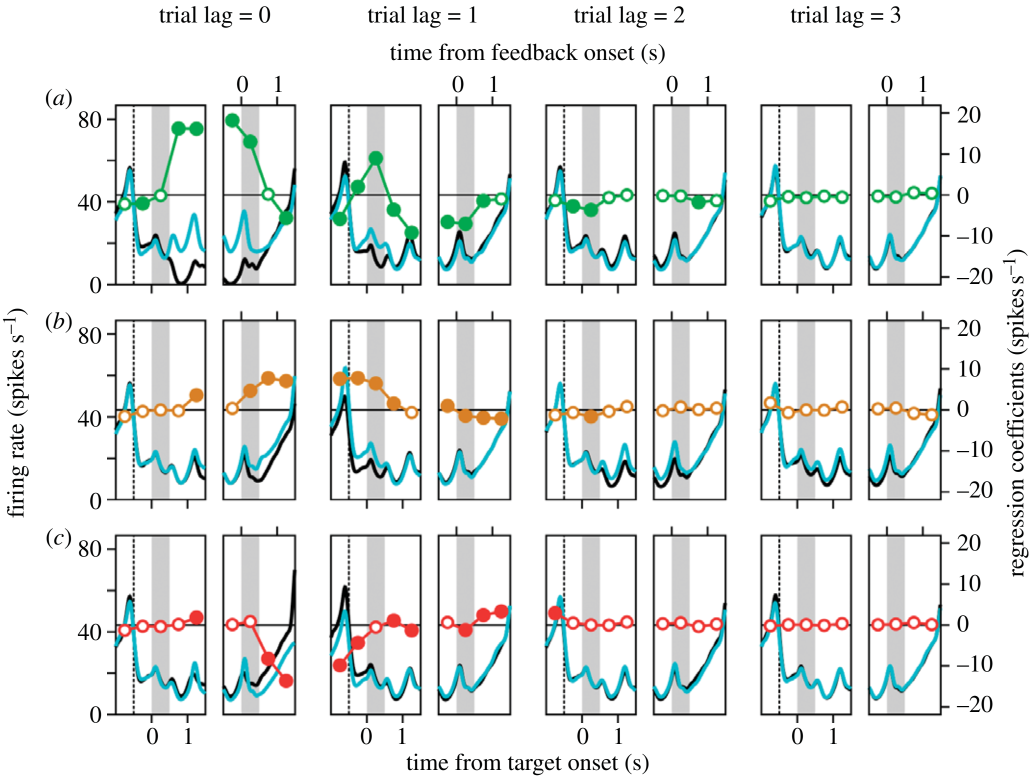

Figure 7. Activity of an example neuron in the DLPFC (also shown in figure 4*a*) during the matching pennies task. Each pair of small panels displays the spike density functions (convolved with a Gaussian kernel, $\sigma = 50$ ms) estimated relative to the time of target onset (left panels) or feedback onset (right panels). They were estimated separately according to the (*a*) animal's choice, (*b*) the computer's choice, or (*c*) reward in the current trial (trial lag $= 0$) or according to the corresponding variables in three previous trials (trial lag $= 1$, 2 or 3). Cyan (black) lines correspond to the activity associated with rightward (leftward) choices (*a*,*b*) or rewarded (unrewarded) trials (*c*). Circles show the regression coefficients from a multiple linear regression model, which was performed separately for a series of 0.5 s windows. Filled circles indicate the coefficients significantly different from zero (*t*-test, $p < 0.05$). Dotted vertical lines in the left panels correspond to the onset of foreperiod. Grey background indicates the delay (left panels) or feedback (right panels) period.

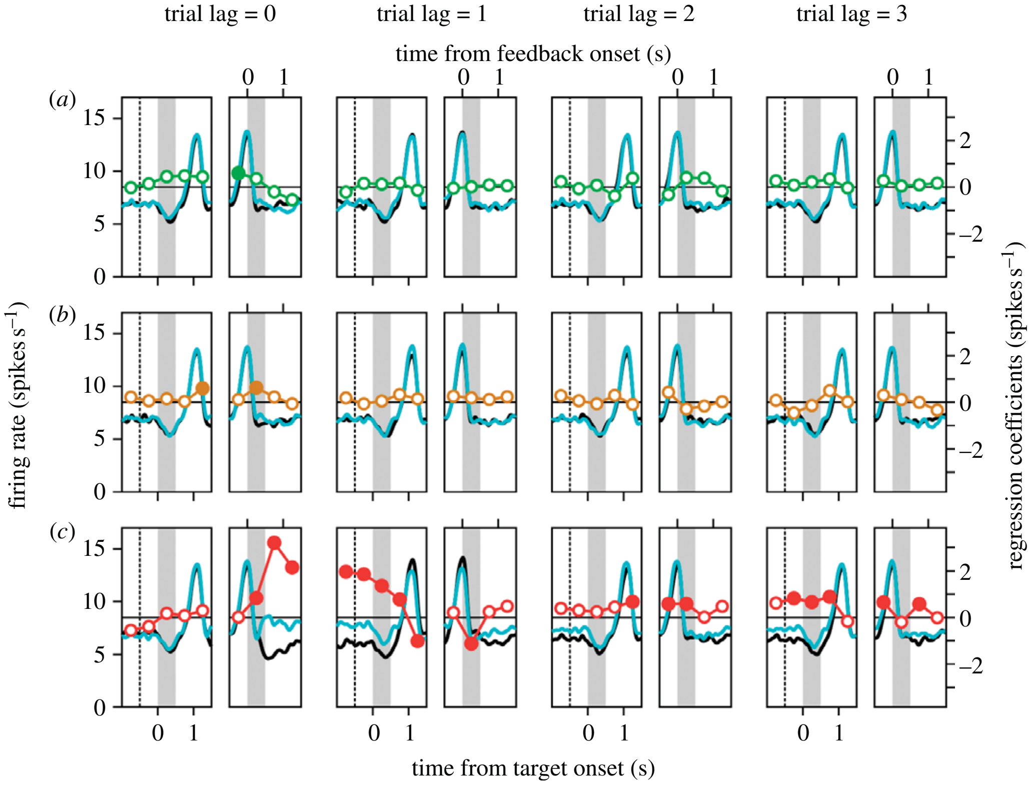

Figure 8. Activity of an example neuron in the ACCd (also shown in figure 4*b*) during the matching pennies task. The description is the same as in figure 7.

found that the activity of neurons in the lateral prefrontal cortex and the ACCd can be selectively linked to specific subprocesses of reinforcement learning. How complex decision-making tasks encountered in our daily lives can be efficiently solved by the brain, however, is still largely unknown, and this will require more intimate interactions across multiple disciplines.

All the procedures used in this study were approved by the University of Rochester Committee on Animal Research, and conformed to the US Public Health and Service Policy on Humane Care and Use of Laboratory Animals and the Guide for the care and use of laboratory animals (Academic Press, 1996).

We are grateful to D. Barraclough, M. Conroy and B. McGreevy for their contributions to the experiments described in this paper. This research was supported by NIH grants NS044270 and MH073246.

REFERENCES

Barraclough, D. J., Conroy, D. & Lee, D. 2004 Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404–410. (doi:10.1038/nn1209)

Brown, R. M., Crane, A. M. & Goldman, P. S. 1979 Regional distribution of monoamines in the cerebral cortex and subcortical structures of the rhesus monkey: concentrations and *in vivo* synthesis rates. *Brain Res.* **168**, 133–150. (doi:10.1016/0006-8993(79)90132-X)

Budescu, D. V. & Rapoport, A. 1994 Subjective randomization in one- and two-person games. *J. Behav. Decis. Mak.* **7**, 261–278. (doi:10.1002/bdm.3960070404)

Camerer, C. F. 2003 *Behavioral game theory*. Princeton, NJ: Princeton University Press.

Corrado, G. & Doya, K. 2007 Understanding neural coding through the model-based analysis of decision making. *J. Neurosci.* **27**, 8178–8180. (doi:10.1523/JNEUROSCI.1590-07.2007)

Dorris, M. C. & Glimcher, P. W. 2004 Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44**, 365–378. (doi:10.1016/j.neuron.2004.09.009)

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

Fehr, E. & Camerer, C. F. 2007 Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn. Sci.* **11**, 419–427. (doi:10.1016/j.tics.2007.09.002)

Flaherty, C. F. 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 foundation of hedonic psychology* (eds D. Kahneman, E. Diener & N. Schwartz), pp. 302–329. New York, NY: Russell Sage Foundation.

Gallagher, H. L., Jack, A. I., Roepstorff, A. & Frith, C. D. 2002 Imaging the intentional stance in a competitive game. *Neuroimage* **16**, 814–821. (doi:10.1006/nimg.2002.1117)

Harbaugh, W. T., Mayr, U. & Burghart, D. R. 2007 Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* **316**, 1622–1625. (doi:10.1126/science.1140738)

Helson, H. 1948 Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychol. Rev.* **55**, 297–313. (doi:10.1037/h0056721)

Hollerman, J. R., Tremblay, L. & Schultz, W. 1998 Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J. Neurophysiol.* **80**, 947–963.

Kable, J. W. & Glimcher, P. W. 2007 The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* **10**, 1625–1633. (doi:10.1038/nn2007)

Kawagoe, R., Takikawa, Y. & Hikosaka, O. 1998 Expectation of reward modulates cognitive signals in the basal ganglia. *Nat. Neurosci.* **1**, 411–416. (doi:10.1038/1625)

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J. & Rushworth, M. F. S. 2006 Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947. (doi:10.1038/nn1724)

Kim, Y. B., Huh, N., Lee, H., Baeg, E. H., Lee, D. & Jung, M. W. 2007 Encoding of action history in the rat ventral striatum. *J. Neurophysiol.* **98**, 3548–3556. (doi:10.1152/jn.00310.2007)

Kim, S., Hwang, J. & Lee, D. 2008 Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron* **59**, 161–172. (doi:10.1016/j.neuron.2008.05.010)

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R. & Montague, P. R. 2005 Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78–83. (doi:10.1126/science.1108062)

Kobayashi, S., Lauwereyns, J., Koizumi, M., Sakagami, M. & Hikosaka, O. 2002 Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *J. Neurophysiol.* **87**, 1488–1498.

Lau, B. & Glimcher, P. W. 2007 Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.* **27**, 14 502–14 514. (doi:10.1523/JNEUROSCI.3060-07.2007)

Lee, D. 2006 Neural basis of quasi-rational decision making. *Curr. Opin. Neurobiol.* **16**, 191–198. (doi:10.1016/j.conb.2006.02.001)

Lee, D. 2008 Game theory and neural basis of social decision making. *Nat. Neurosci.* **11**, 404–409. (doi:10.1038/nn2065)

Lee, D., Conroy, M. L., McGreevy, B. P. & Barraclough, D. J. 2004 Reinforcement learning and decision making in monkeys during a competitive game. *Cogn. Brain Res.* **22**, 45–58. (doi:10.1016/j.cogbrainres.2004.07.007)

Lee, D., McGreevy, B. P. & Barraclough, D. J. 2005 Learning and decision making in monkeys during a rock–paper–scissors game. *Cogn. Brain Res.* **25**, 416–430. (doi:10.1016/j.cogbrainres.2005.07.003)

Lee, D., Rushworth, M. F. S., Walton, M. E., Watanabe, M. & Sakagami, M. 2007 Functional specialization of the primate frontal cortex during decision making. *J. Neurosci.* **27**, 8170–8173. (doi:10.1523/JNEUROSCI.1561-07.2007)

Leon, M. I. & Shadlen, M. N. 1999 Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron* **24**, 415–425. (doi:10.1016/S0896-6273(00)80854-5)

Lewis, D. A., Melchitzky, D. S., Sesack, S. R., Whitehead, R. E., Auh, S. & Sampson, A. 2001 Dopamine transporter immunoreactivity in monkey cerebral cortex: regional, laminar, and ultrastructural localization. *J. Comp. Neurol.* **432**, 119–136. (doi:10.1002/cne.1092)

Matsumoto, M., Matsumoto, K., Abe, H. & Tanaka, K. 2007 Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.* **10**, 647–656. (doi:10.1038/nn1890)

McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. 2001 A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl Acad. Sci. USA* **98**, 11 832–11 835. (doi:10.1073/pnas.211415698)

McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. 2004 Separate neural systems value immediate and delayed monetary rewards. *Science* **306**, 503–507. (doi:10.1126/science.1100907)

McCoy, A. N. & Platt, M. L. 2005 Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat. Neurosci.* **8**, 1220–1227. (doi:10.1038/nn1523)

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R. & Grafman, J. 2006 Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc. Natl Acad. Sci. USA* **103**, 15 623–15 628. (doi:10.1073/pnas.0604475103)

Mookherjee, D. & Sopher, B. 1994 Learning behavior in an experimental matching pennies game. *Games Econ. Behav.* 7, 62–91. (doi:10.1006/game.1994.1037)

Mookherjee, D. & Sopher, B. 1997 Learning and decision costs in experimental constant sum games. *Games Econ. Behav.* **19**, 97–132. (doi:10.1006/game.1997.0540)

Nash, J. F. 1950 Equilibrium points in *n*-person games. *Proc. Natl Acad. Sci. USA* **36**, 48–49. (doi:10.1073/pnas.36.1.48)

Padoa-Schioppa, C. & Assad, J. A. 2006 Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226. (doi:10.1038/nature04676)

Pawitan, Y. 2001 *In all likelihood: statistical modelling and inference using likelihood*. Oxford, UK: Claredon Press.

Platt, M. L. & Glimcher, P. W. 1999 Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238. (doi:10.1038/22268)

Quilodran, R., Rothé, M. & Procyk, E. 2008 Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* **57**, 314–325. (doi:10.1016/j.neuron.2007.11.031)

Reynolds, J. N. J. & Wickens, J. R. 2002 Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* **15**, 507–521. (doi:10.1016/S0893-6080(02)00045-X)

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S. & Kilts, C. D. 2002 A neural basis for social cooperation. *Neuron* **35**, 395–405. (doi:10.1016/S0896-6273(02)00755-9)

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. 2004*a* The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* **22**, 1694–1703. (doi:10.1016/j.neuroimage.2004.04.015)

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. 2004*b* Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* **15**, 2539–2543. (doi:10.1097/00001756-200411150-00022)

Roesch, M. R. & Olson, C. R. 2003 Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *J. Neurophysiol.* **90**, 1766–1789. (doi:10.1152/jn.00019.2003)

Rushworth, M. F. S. & Behrens, T. E. J. 2008 Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397. (doi:10.1038/nn2066)

Rushworth, M. F. S., Behrens, T. E. J., Rudebeck, P. H. & Walton, M. E. 2007 Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behavior. *Trends Cogn. Sci.* **11**, 168–176. (doi:10.1016/j.tics.2007.01.004)

Samejima, K., Ueda, Y., Doya, K. & Kimura, M. 2005 Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340. (doi:10.1126/science.1115270)

Sanfey, A. G. 2007 Social decision-making: insights from game theory and neuroscience. *Science* **318**, 598–602. (doi:10.1126/science.1142996)

Schultz, W. 1998 Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27.

Schweighofer, N. & Doya, K. 2003 Meta-learning in reinforcement learning. *Neural Netw.* **16**, 5–9. (doi:10.1016/S0893-6080(02)00228-9)

Seo, H. & Lee, D. 2007 Temporal filtering of reward signals in the dorsal anterior cingulate cortex. *J. Neurosci.* **27**, 8366–8377. (doi:10.1523/JNEUROSCI.2369-07.2007)

Seo, H., Barraclough, D. J. & Lee, D. 2007 Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cereb. Cortex* **17**, i110–i117. (doi:10.1093/cercor/bhm064)

Shidara, M. & Richmond, B. J. 2002 Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science* **296**, 1709–1711. (doi:10.1126/science.1069504)

Sohn, J.-W. & Lee, D. 2007 Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *J. Neurosci.* **27**, 13 655–13 666. (doi:10.1523/JNEUROSCI.2982-07.2007)

Soltani, A., Lee, D. & Wang, X.-J. 2006 Neural mechanism for stochastic behavior during a competitive game. *Neural Netw.* **19**, 1075–1090. (doi:10.1016/j.neunet.2006.05.044)

Sugrue, L. P., Corrado, G. S. & Newsome, W. T. 2004 Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782–1787. (doi:10.1126/science.1094765)

Sutton, R. S. & Barto, A. G. 1998 *Reinforcement learning: an introduction*. Cambridge, UK: MIT Press.

Thorndike, E. L. 1911 *Animal intelligence: experimental studies*. New York, NY: MacMillan.

von Neumann, J. & Morgenstern, O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Walton, M. E., Devlin, J. T. & Rushworth, M. F. S. 2004 Interactions between decision making and performance monitoring within prefrontal cortex. *Nat. Neurosci.* **7**, 1259–1265. (doi:10.1038/nn1339)

Watanabe, M. 1996 Reward expectancy in primate prefrontal neurons. *Nature* **382**, 629–632. (doi:10.1038/382629a0)

Yang, T. & Shadlen, M. N. 2007 Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080. (doi:10.1038/nature05852)