Available online at www.sciencedirect.com

SCIENCE DIRECT

Cognitive Brain Research 25 (2005) 416 – 430

COGNITIVE BRAIN RESEARCH

www.elsevier.com/locate/cogbrainres

Research Report

Learning and decision making in monkeys during a rock–paper–scissors game

Daeyeol Lee*, Benjamin P. McGreevy, Dominic J. Barraclough

Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, NY 14627, USA

Accepted 12 July 2005
Available online 10 August 2005

Abstract

Game theory provides a solution to the problem of finding a set of optimal decision-making strategies in a group. However, people seldom play such optimal strategies and adjust their strategies based on their experience. Accordingly, many theories postulate a set of variables related to the probabilities of choosing various strategies and describe how such variables are dynamically updated. In reinforcement learning, these value functions are updated based on the outcome of the player's choice, whereas belief learning allows the value functions of all available choices to be updated according to the choices of other players. We investigated the nature of learning process in monkeys playing a competitive game with ternary choices, using a rock–paper–scissors game. During the baseline condition in which the computer selected its targets randomly, each animal displayed biases towards some targets. When the computer exploited the pattern of animal's choice sequence but not its reward history, the animal's choice was still systematically biased by the previous choice of the computer. This bias was reduced when the computer exploited both the choice and reward histories of the animal. Compared to simple models of reinforcement learning or belief learning, these adaptive processes were better described by a model that incorporated the features of both models. These results suggest that stochastic decision-making strategies in primates during social interactions might be adjusted according to both actual and hypothetical payoffs.

Theme: Neural basis of behavior
Topic: Cognition

Keywords: Game theory; Mixed strategy; Motivation; Prefrontal cortex; Reward; Zero-sum game

1. Introduction

Many disciplines of science, such as economics, psychology, and neuroscience, seek quantitative models to describe the process of decision making. In most cases, it is assumed that the desirability of alternative actions is evaluated in terms of a common currency, making it possible for a decision maker to select an action with the optimal outcome. However, often neglected is the fact that this process needs to be empirically adjusted for individual animals, especially when they face a complex and dynamic environment. Furthermore, this adaptive process might be tuned for a given species during evolution so that the complexity of learning rules matches its environment. For example, for solitary carnivorous animals, a relatively simple learning rule based on random sampling and comparison of different outcomes might be sufficient to optimize its hunting strategy. For animals living in social groups, however, learning rules for decision making are likely to take a more complex form. This is because one's action can modify the decision-making strategies of other individuals and thereby influence the outcome of future interactions with them.

The analysis of decision making in a social group is the topic of game theory [40]. A game is defined by a set of choices or strategies available to each player, and a payoff matrix that specifies the outcome (utility) to each player

* Corresponding author. Fax: +1 585 271 3043.
E-mail address: dlee@cvs.rochester.edu (D. Lee).
URL: http://www.bcs.rochester.edu/~dlee/ (D. Lee).

0926-6410/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogbrainres.2005.07.003

according to the choices of all players. A solution of a game refers to a set of strategies that would be selected by "rational" players each trying to maximize his or her utility. Accordingly, it was an important discovery when Nash proved that any N-player game includes at least one such solution. This is known as Nash equilibrium and defined as a set of strategies from which no players can increase their payoffs by changing their strategies individually [25]. Unfortunately, this important concept has theoretical and practical limitations. First, a game can have multiple Nash equilibria, and it is difficult to determine which equilibrium should be preferred. Second, a large number of empirical studies have demonstrated that people deviate, often systematically, from such equilibrium. These limitations led to the proposals that learning might play an important role in optimizing decision-making strategies. In fact, many studies have shown that various learning models describe the observed pattern of decision making better than the equilibrium predictions [2,3,6,7,12,13,15,23,24,30,31]. In the present study, we have analyzed the choice behavior of monkeys during a simple zero-sum game with ternary choices, known as rock–paper–scissors. This was motivated by two considerations. First, rigorous comparative studies of choice behavior in non-human primates can potentially provide important insights into the evolutionary origins of human decision-making process. Second, such primate models of decision making would also provide important opportunities to understand the neural mechanisms of human decision making. For example, classical game theory and other standard economic models have always postulated certain variables, such as utility, that cannot be measured directly, making it difficult to test such theories rigorously. Recent advances in neuroscience, especially an emerging field of neuroeconomics, might make it possible to obtain precise measures of quantities that have been hitherto merely theoretical [16,41].

In our previous studies [1,20], we have examined the choices monkeys made during a binary zero-sum game, known as matching pennies. By training monkeys to play such a competitive game against a computer opponent, we showed that the animal's behavior can be modified by the strategies of its opponent. The Nash equilibrium in matching pennies requires a player to make two choices randomly with equal probabilities. Compared to a baseline condition, in which the animal was rewarded randomly, the choice of the animal became more random when the computer started exploiting statistical biases displayed by the animal in its choices. However, such biases did not disappear completely even when the computer analyzed the animal's choice as well as its reward history. These biases were consistent with the predictions of a reinforcement learning model, suggesting that the animals approximated the equilibrium strategy through experience. Due to the simplicity of the task used in our previous study, however, it was not possible to distinguish among alternative learning models.

Typically, models of adaptive decision making postulate a set of variables, one for each action, that are related to the probabilities of choosing different actions. These variables have been referred to as value functions [39], propensities [12], or attractions [6], and they are updated iteratively through the experience of the player [5,13,15,29]. In reinforcement learning models, value functions are updated strictly based on the outcome of a player's choice [12,39]. For example, in matching pennies, if a player selects the head, only the value function for the head is updated according to the outcome of his or her choice. In belief learning models, on the other hand, it is assumed that players choose their actions based on their beliefs as to how other players would behave [27]. At one extreme, this could be entirely based on the most recent choices of other players, which is referred to as Cournot dynamics [8]. In other words, decision makers may choose an option which is the best response to the most recent choices of other players they are interacting with. The other extreme is fictitious play, where the probability for a given choice of another player can be estimated based on its empirical frequency from the entire history that can be observed [27]. In weighted fictitious play, this approach was modified to give more weights to recent choices by other players [5,7]. Once the beliefs about the choices of other players are formed, they can be used to generate the expected payoffs for different choices of a given player. These expected payoffs can then be converted to the probability of choosing an action, and therefore play the role analogous to that of value functions in reinforcement learning [6,13]. It should be noted that the expected payoffs in a belief learning model are updated not only for a particular action chosen by a given player, but for all actions according to the hypothetical payoffs that the player would have received by choosing each action, given the choices of other players in previous trials. For example, if a player selects the head and wins in a matching pennies game, the value function for the head might increase and that for the tail might decrease. However, for games with binary choices, such as matching pennies, reinforcement learning and belief learning models make similar predictions and therefore are difficult to distinguish. If the player's choice depends on the difference between the value functions of two choices, these two models would become equivalent, since any increase in the value function for one choice would be equivalent to the decrease in the value function for the other choice by the same amount. These two models make distinct predictions, however, when the number of alternative choices is increased from two [24]. Thus, in order to understand the nature of learning in decision making, we examined in the present study the choice behavior of monkeys during a rock–paper–scissors game. The results showed that reinforcement learning models performed better than belief learning models. However, a hybrid model that incorporated the features of both models provided an even better fit to the data. In addition, in one animal, analysis of conditional probabilities revealed some features of belief learning model. These results suggest that a

learning process of monkeys in decision making might not be fully accounted for by simple reinforcement learning models.

2. Methods

2.1. Animal preparation and apparatus

Two male rhesus monkeys (*Macaca mulatta*, body weight = 7–12 kg) were used in this study. The animal was seated in a primate chair and faced a computer monitor located approximately 57 cm from their eyes. All visual stimuli were presented on the computer monitor. The animal's eye position was sampled at 250 Hz with a high-speed video-based eye tracker (ET49, Thomas Recording, Germany). All the procedures used in the present study were approved by the University of Rochester Committee on Animal Research, and conformed to the principles outlined in the Guide for the Care and Use of Laboratory Animals (NIH publications No 80-23, revised 1996).

2.2. Behavioral task

The animals performed an oculomotor version of a rock–paper–scissors game, similar to the matching pennies game used in our previous studies [1,20], except that the present task included 3 different choices (Fig. 1). Three different visual targets were arbitrarily designated as rock, paper, and scissors, respectively. At the beginning of each trial, the computer opponent selected its target according to one of the algorithms described below, and the outcome of the animal's choice was classified as loss, tie, and win, according to the following rule: rock beats scissors, scissors beat paper, and paper beats rock. At the end of each completed trial, the animal was rewarded with one or two drops (a drop = approximately 0.23 ml) of juice for tie and win, respectively. No reward was given for the trial with a loss.

At the beginning of each trial, the animal was required to fixate a yellow square (0.9° × 0.9°; CIE x = 0.432, y = 0.494, Y = 62.9 cd/m^2) presented at the center of the computer screen (Fig. 1). After a 0.5 s fore-period, three identical green disks (radius = 0.6°; CIE x = 0.286, y = 0.606, Y = 43.2 cd/m^2) were presented on the circumference of an imaginary circle (radius = 5°). The animal maintained its fixation on the central square during the following 0.5 s delay period. At the end of this delay period, the central square was extinguished, and the animal was required to produce a saccadic eye movement towards one of the targets within 1 s and maintain its fixation for a 0.5 s hold period. At the end of the hold period, a yellow ring was displayed for 100 ms around the target that was selected by the computer. Simultaneously, a red ring (radius = 1.0°; CIE x = 0.632, y = 0.341, Y = 17.6 cd/m^2) was also displayed around the target that would beat the computer's choice.

2.3. Algorithms of computer opponent

As in our previous study on a matching-pennies game [1], each animal was tested with 3 different algorithms with increasing levels of sophistication.

In algorithm 0, the computer selected three targets randomly with equal probabilities (i.e., p = 1/3). In a rock–paper–scissors game, this mixed strategy corresponds to the Nash equilibrium. Against the computer opponent with this strategy, any strategy adopted by the animal would produce the same expected payoff.

In algorithm 1, the computer stored the entire sequence of choices made by the animal in a given session. In each trial, the computer then used this information to calculate the conditional probabilities that the animal would choose each target given the animal's choices in the preceding N trials (N = 0 to 4). A null hypothesis that this probability is 1/3 was tested for each of these conditional probabilities (binomial test, $p < 0.05$). If none of these hypotheses was rejected, it was assumed that the animal had selected all three targets with equal probabilities independently from its previous choices, and the computer selected its target randomly as in algorithm 0. If one or more hypotheses were rejected, the

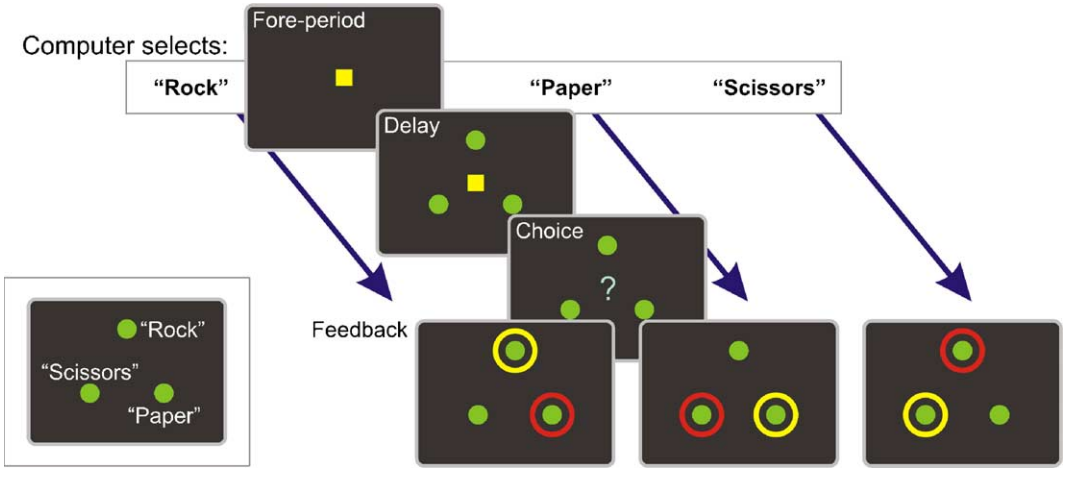

Fig. 1. Spatio-temporal sequence of a free-choice task used in a rock–paper–scissors game.

computer selected its target based on a particular order of conditional probabilities that includes the maximum probability which was significantly different from 1/3. Denoting this set of conditional probabilities for rock, paper, scissors, as p, q, and $1-(p+q)$, the computer selected each of these three targets with the probabilities of $1-(p+q)$, p, and q. For example, if the animal exclusively selects rock (i.e., $p=1$), this would lead the computer opponent to choose paper with certainty. In algorithm 1, therefore, the animal was required to select the three targets randomly with equal probabilities and independently from its previous choices, in order to maximize its total reward.

In algorithm 2, the computer used the entire choice and reward history of the animal in a given session to predict the animal's choice in the next trial. To this end, a series of conditional probabilities that the animal would choose each target, given the animal's choices in the preceding N trials ($N=1$ to 4) along with their payoffs, were calculated. As in algorithm 1, each of these conditional probabilities was tested against the null hypothesis that the corresponding conditional probability is 1/3. If none of these hypotheses was rejected, then the computer selected each target randomly with the probability of 1/3. Otherwise, the computer biased its target selection according to the same rule used in algorithm 1. In algorithm 2, therefore, the animal was required to select its targets not only with equal probabilities and independently from its previous choices, but also independently from the combination of its previous choices and their outcomes.

2.4. Data analysis

2.4.1. Analysis of choice probability and serial dependence

Probability that the animal would choose one of the targets according to a given strategy (e.g., choose rock) was estimated for successive blocks of 100 or 2000 trials in each algorithm. The statistical significance for rejecting the null hypothesis that each of these probabilities was equal to a particular value was evaluated using a binomial test. Whether the difference in a pair of such probabilities was statistically significant was determined with a Z test [36]. The tendency for such probabilities to increase or decrease throughout the course of a particular algorithm was tested with a regression model with the block number and the estimated probability as the independent and dependent variables, respectively. The statistical significance of a regression coefficient was determined with a t test.

2.4.2. Entropy and mutual information

The degree of randomness in the animal's choice sequence was quantified with entropy and mutual information. Both of these measures were evaluated using the choice sequence of the two players in 3 successive trials, since this made it possible to obtain relatively reliable estimates by limiting the number of possible outcomes. Specifically, if there are k possible outcomes and its i-th event has a probability p_i, the entropy H is defined by the following.

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i (\text{bits}).$$

When the entropy was calculated based only on the animal's choice sequence in 3 successive trials, there were a total of 27 possible outcomes ($k=3^3=27$), and the maximum entropy was 4.755 bits. Entropy was also calculated based on the animal's choice sequence in 3 successive trials and the choice of the computer's opponent in the first two of these 3 trials ($k=3^5=243$). The maximum entropy in this case was 7.925 bits. When the entropy is estimated using the probabilities estimated from a finite sample, the estimate for the entropy is biased [22]. To correct for this bias, the entropy was estimated by the following.

$$H = -\sum_{i=1}^{k} \hat{p}_i \log_2 \hat{p}_i + \frac{k-1}{1.3863\ N} (\text{bits}),$$

where p_i denotes the maximum likelihood estimate for p_i, and N the number of samples.

Mutual information was calculated between the animal's choice in 2 successive trials (input) and the animal's choice in the next trial (output), and between the choice sequence of both players in 2 successive trials and the animal's choice in the next trial. This was estimated as the following to correct for the bias due to a finite sample [22].

$$I = -\sum_{i}^{r}\sum_{c}^{j} \hat{p}_{ij} \log_2 \frac{\hat{p}_{ij}}{\hat{p}_i \hat{p}_j} - \frac{(r-1)(c-1)}{1.3863\ N} (\text{bits}),$$

where p_i is the probability of the i-th outcome in the input event ($r=3^2=9$, or $3^4=81$), p_j is the probability of the j-th outcome in the output event ($c=3$), and p_{ij} is the joint probability for the i-th input event and j-th output event.

2.5. Learning models

In order to determine whether and how an animal's choice is influenced by the cumulative effects of its previous choices and their outcomes, a set of learning models were fit to the data. A common feature in all of these models is that a variable, referred to as the value function, is associated with each choice. How value functions for different choices are adjusted after each trial varies across different models. For example, in reinforcement learning, value functions are adjusted strictly according to the outcome of the animal's choice. In contrast, in belief learning models, value functions are adjusted strictly according to the choices of other players (computer opponent in this case), regardless of the choice of the animal. These two different types of models can be considered as two special cases in a spectrum [6]. Therefore, one can also consider a model in which value functions are adjusted according to choices of all players. This is referred to as a general learning model. The

parameters of all models were estimated according to the maximum likelihood procedure [4] using a function minimization algorithm in Matlab (Mathworks Inc. MA).

2.5.1. Reinforcement learning model

In all of the models examined in the present study, the value function at trial t for a given target x ($x = R$, P, or S, for rock, paper, scissors, respectively), $V_t(x)$, was updated after each trial according to the following:

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

where α is a decay rate, and $\Delta_t(x)$ reflects a change in the value function for target x. In the reinforcement learning model, $\Delta_t(x) = \Delta_L$ if the animal selects the target x and loses (i.e., no reward), $\Delta_t(x) = \Delta_T$ if the animal selects the target x and ties with the computer (i.e., small reward), and $\Delta_t(x) = \Delta_W$ if the animal selects the target x and wins (i.e., large reward). $\Delta_t(x)$ is set to 0, if the animal does not select the target x. The probability that the animal would select a given target is then determined according to the softmax transformation. In other words,

$$p_t(x) = \frac{\exp V_t(x)}{\sum_{u \in \{R,P,S\}} \exp V_t(u)}.$$

2.5.2. Belief learning model

This model is similar to the reinforcement learning model, except that the value functions were updated entirely according to the choice of the computer opponent. Therefore, unlike the reinforcement learning model describe above, $\Delta_t(x) = \Delta_L$ for the target that would have been beaten by the computer's choice, $\Delta_t(x) = \Delta_T$ for the target that would have resulted in a tie, and $\Delta_t(x) = \Delta_W$ for the target that would have beaten the computer's choice. It should be noted that these adjustments are applied to all targets regardless of the animal's choice. Since value functions were converted to the probability of choosing different targets via softmax transformation, adding a constant offset to the value functions of all choices does not alter the resulting set of probabilities of choosing different targets. Therefore, Δ_L was set to 0, and the model was fit to the data by choosing the remaining 3 parameters (Δ_T, Δ_W, and α) according to the maximum likelihood procedure.

2.5.3. General learning model

Both the reinforcement learning and belief learning models described above can be generalized by allowing the changes in the value functions to be determined by a combination of the animal's choice and that of the computer opponent. For example, if the animal loses in a given trial, the value functions for all 3 targets might be adjusted simultaneously as in the belief learning model. In other words, $\Delta_t(x) = \Delta_{LL}$ for the target chosen by the animal in a loss trial, $\Delta_t(x) = \Delta_{LT}$ for the target that could have resulted in a tie, and $\Delta_t(x) = \Delta_{LW}$ for the target that could have resulted in a win. In this general learning model, however, the changes applied to the value functions after a loss trial can differ from those applied following a tie or win trial. Therefore, the changes applied to the value functions after a tie trial are denoted as Δ_{TL}, Δ_{TT}, and Δ_{TW}, and they were estimated separately. The corresponding parameters for a win trial are denoted by Δ_{WL}, Δ_{WT}, and Δ_{WW}. As in the belief learning, the constant offset can be subtracted from the value functions of all targets simultaneously without affecting the probability of choice. Therefore, Δ_{LL}, Δ_{TL}, and Δ_{WL} were set to 0, and the remaining parameters were estimated.

2.5.4. Model selection

In general, the performance of a model, as evaluated by the measures based on the sum of squared errors, improves with an increasing number of free parameters used to estimate the model. Therefore, in order to compare the performance of multiple models, it is necessary to correct for the improvement in the model fit expected from the difference in the number of free parameters. Two different methods, both based on the log-likelihood, were utilized in the present study. First, the Akaike's information criterion (AIC), was computed by the following,

$$\mathrm{AIC} = -2 \log L + 2k,$$

where k is the number of free parameters used in a given model [4]. Second, Bayesian information criterion (BIC) was obtained according to the following,

$$\mathrm{BIC} = -2 \log L + k \log N,$$

where N denotes the number of data points. For a relatively large number of data points ($N > 7.4$), BIC penalizes complex models more than AIC [17].

3. Results

3.1. Database

A total of 5765, 82,479, and 81,627 choices of two monkeys were obtained for algorithms 0, 1, and 2, respectively. The number of days and that of trials in which each animal was tested for different algorithms are shown in Table 1.

3.2. Choice and reward probability

Each animal was tested with algorithm 0 for 2 days, and both animals selected rock in less than 8% of the trials and therefore displayed substantial deviations from the Nash equilibrium (Fig. 2). In addition, the probability that the animal would choose rock significantly decreased during algorithm 0, and a regression analysis showed that this trend was significant in both animals (t test, $p < 10^{-5}$). Whereas

Table 1
Number of days and trials tested in each animal and algorithm

Algorithm	Animal	Days	Trials	Trials/day ± SD
0	E	2	3011	1505 ± 371
	F	2	2754	1377 ± 42
1	E	20	42,598	2130 ± 616
	F	31	39,881	1287 ± 336
2	E	19	41,591	2189 ± 579
	F	19	40,036	2107 ± 326

Table 2
Probabilities of choosing rock, paper, and scissors

Algorithm	Animal	p(rock)	p(paper)	p(scissors)
0	E	0.0784	0.5271	0.3946
	F	0.0476	0.2389	0.7135
1	E	0.2820	0.3737	0.3443
	F	0.2448	0.3674	0.3878
2	E	0.2717	0.3838	0.3445
	F	0.2522	0.3790	0.3688

monkey E selected paper most frequently (52.7%), monkey F selected scissors most frequently (71.3%; Table 2). These results are not surprising, since during algorithm 0, the animal would receive on average one drop of juice, regardless of its decision-making strategy. Indeed, each animal received approximately one drop of juice on average when tested with algorithm 0 (Table 3; Fig. 3).

Following the introduction of algorithm 1, the probability that the animal would choose rock increased, although this change was somewhat more delayed in monkey E. The percentage of choosing rock in successive blocks of 100 trials remained below 5% for the first 1500 trials in monkey E, whereas this was the case only for the first 400 trials in monkey F (Fig. 2, left panels). Accordingly, there was a larger decrease in the average amount of reward received by monkey E during the corresponding period (Fig. 3, left panels). In both animals, the average reward increased gradually and reached a level relatively close to that expected for optimal performance by the end of the first day of algorithm 1 (3787 and 1507 trials for monkeys E and F, respectively). However, the probability of choosing rock remained slightly lower than the probability of choosing paper or that of choosing scissors throughout the duration of algorithm 1 (Fig. 2). Even after removing the first 10,000 trials, the percentage of choosing rock was 29.1% and 25.4% for the remaining trials in algorithm 1 for monkeys E and F, respectively, and both of these were significantly lower than 1/3 (binomial test, $p < 10^{-60}$). Accordingly, although the overall average number of rewards during algorithm 1 was larger than 0.95 for both animals, the percentage of loss trials was significantly higher than 1/3 in both animals (36.3% and 35.6%; $p < 10^{-10}$; Table 3, Fig. 4). Overall, the introduction of algorithm 2 produced only relatively small changes in the probability of choosing different targets (Table 2; Fig. 2), the average amount of reward earned by the animal (Table 3; Fig. 3), or the probability of trials with different outcomes (e.g., win or loss; Table 3, Fig. 4).

3.3. Serial dependence and randomness in choice sequence

The null hypothesis that the choices in two successive trials were made independently was rejected for all animals and algorithms by analyzing the 3 × 3 contingency table (χ^2 test; $\chi^2 > 140$, $p < 10^{-16}$, in all cases). To examine specifically how the successive choices deviated from the

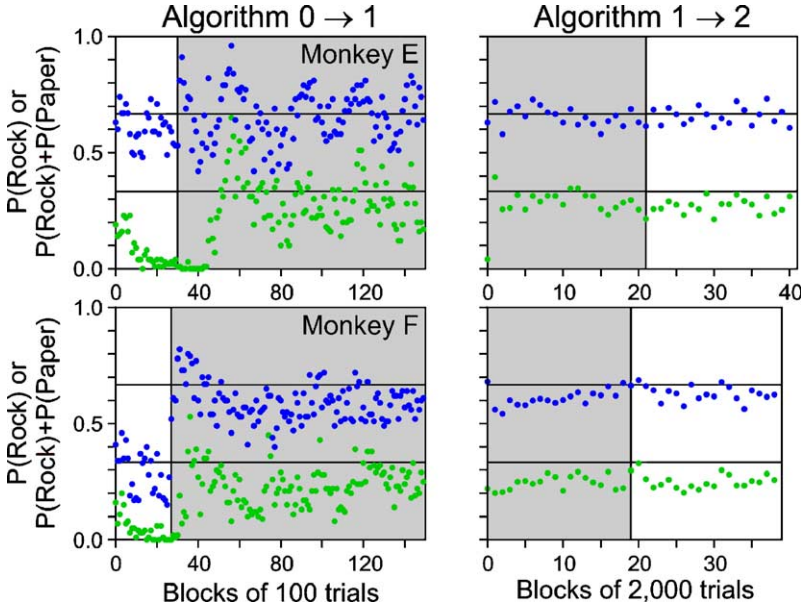

Fig. 2. The frequency of choosing rock (green dots), and the frequency of choosing rock or paper (blue dots), in a blocks of 100 (left) or 2000 (right) trials. Gray background indicates the results from the trials in which the computer opponent selected its targets according to algorithm 1.

Table 3
Probabilities of loss, tie, and win, and average number of rewards

Algorithm	Animal	p(loss)	p(tie)	p(win)	Mean reward
0	E	0.3212	0.3447	0.3341	1.0130
	F	0.3232	0.3362	0.3406	1.0174
1	E	0.3627	0.3147	0.3226	0.9600
	F	0.3558	0.3197	0.3246	0.9688
2	E	0.3695	0.3163	0.3143	0.9448
	F	0.3520	0.3223	0.3257	0.9736

predictions of the independence assumption, conditional probabilities were estimated for choosing each possible combination of two successive choices (e.g., paper followed by scissors). For algorithm 0, the most salient deviation was observed for the probability of choosing rock (Fig. 5). Although this probability was relatively low in both animals, the probability of choosing rock was substantially higher after a trial with the same choice (42.37% and 25.95% for monkey E and F, respectively), and this was significantly higher than the probability of choosing rock after paper or scissors ($p < 10^{-10}$). For algorithms 1 and 2, the deviations from independence were smaller than in algorithm 0 (Fig. 5).

In order to determine how the animal's choice in a given trial was influenced by the choice and its outcome in the previous trial, choices were classified according their relationship with the computer's choice in the previous trial. For example, the animal's choice that would defeat the computer's choice in the previous trial is referred to as the Cournot best response (CBR). Similarly, a Cournot second best response (CSBR) refers to the choice that would tie with the previous choice of the computer. Finally, a Cournot worst response (CWR) refers to the choice that would be defeated by the computer's previous choice. If the animal makes its choice independently from the computer's previous choice, the probability of CBR, CSBR, and CWR would be all equal to 1/3. This null hypothesis was rejected for all animals and algorithms ($p < 0.005$), with the only exception being the probability of CWR in monkey F during algorithm 2 (Table 4). In all cases, the probability of CBR was significantly higher than 1/3 (binomial test, $p < 10^{-10}$). During algorithm 1, interestingly, the probability of CBR increased gradually in monkey E, but decreased in monkey F (Fig. 6). A regression analysis showed that both of these trends were statistically significant (t test, $p < 0.001$).

The Nash equilibrium for a rock–paper–scissors game requires that the animal selects each target with the probability of 1/3 and that successive choices be made independently from the previous choices of the animal and its opponent. To determine how closely the animal's actual choice patterns approached this optimal performance, entropy and mutual information were calculated based on the sequence of choices in 3 successive trials in blocks of 2000 trials (Fig. 7). This analysis was performed only for algorithms 1 and 2, since algorithm 0 was tested for a relatively small number of trials and produced a substantial deviation in the animal's choice behavior from the equilibrium prediction. The entropy for 3 successive choices of the animal was relatively close to the maximum possible value (Fig. 7). The value of entropy during the first block of 2000 in algorithm 1 was relatively low in both animals, indicating that the animal's choice behavior stabilized during this period. These values were, therefore, removed in the calculation of mean entropy. The mean entropy for algorithm 1 was 4.675 and 4.639 bits for monkeys E and F, respectively. The corresponding values for algorithm 2 were 4.650 and 4.664. Therefore, the difference between the

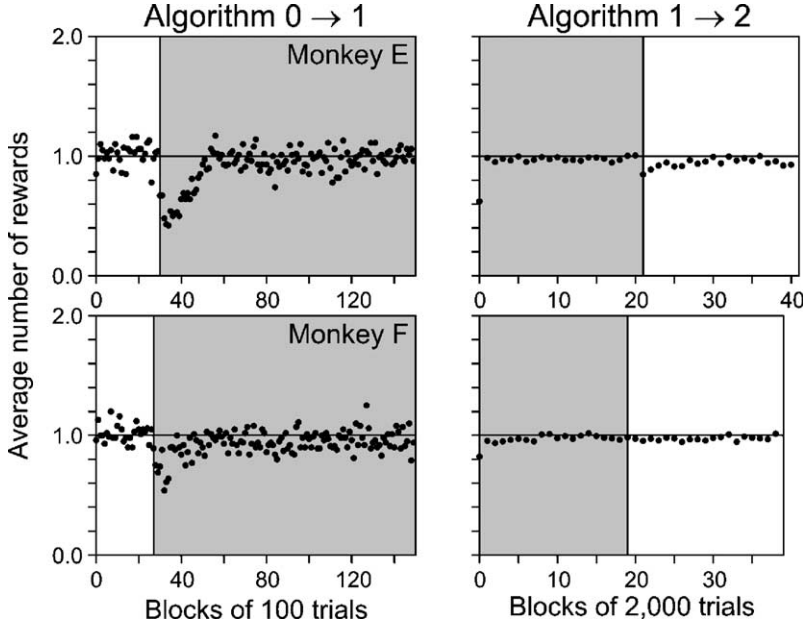

Fig. 3. The average reward received by the animal. Same format as in Fig. 2.

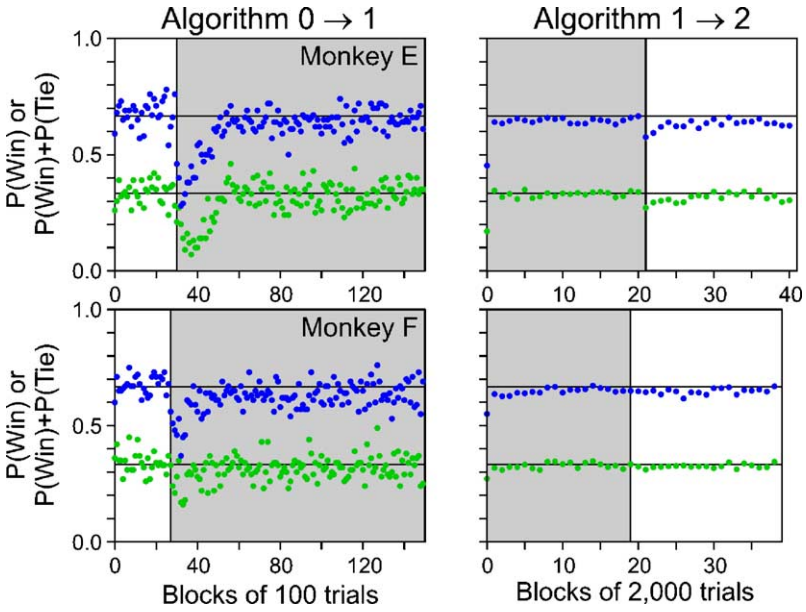

Fig. 4. The frequency of win (green dots) and the frequency of win or tie (blue dots). Same format as in Fig. 2.

mean entropy and the theoretical maximum was less than 0.12 in all cases. The difference in the mean entropy between algorithms 1 and 2 was not significant in either animal. The mutual information between the two previous choices and the current choice was also quite small (Fig. 7), and the average mutual information remained below 0.025 bits in all animals and algorithms when the first block of 2000 trials in algorithm 1 was excluded. The average mutual information in algorithm 2 (0.021 and 0.005 for monkeys E and F) was lower than that in algorithm 1 (0.025 and 0.009), but this difference was significant only in monkey F (t test, $p < 0.01$). A regression analysis showed that, in most cases, the values of entropy or mutual information did not show any significant increase or decrease during the duration of a given algorithm, except that the value of mutual information decreased significantly during algorithm 1 in monkey F ($p < 0.001$), even after the first block of 2000 was removed.

Entropy and mutual information were also calculated between the choices of the animal and the computer opponent during the 2 successive trials and the animal's choice in the next trial. Compared to the entropy computed without taking into account the computer's choice, the mean entropy based on the choice patterns of both players displayed somewhat more substantial deviations from the maximum value. For example, the mean entropies for algorithm 1 were 7.318 and 7.656 bits for monkeys E and F, respectively, whereas the corresponding values for algorithm 2 were 7.694 and 7.730. The difference in the average entropy between the algorithms 1 and 2 was significant in monkey E ($p < 10^{-11}$), but not in monkey F

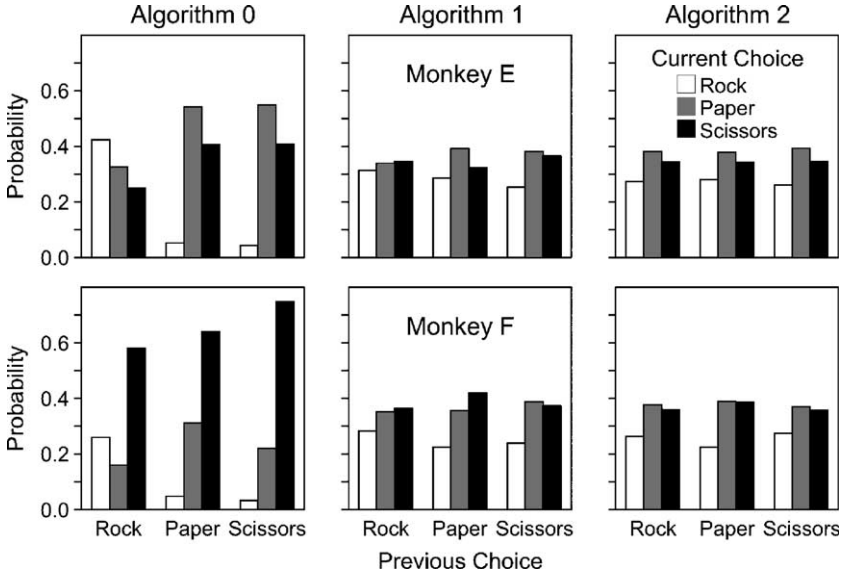

Fig. 5. Conditional probabilities of selecting rock (white), paper (gray), or scissors (black), after selecting rock, paper, or scissors in the previous trial (abscissa).

Table 4
Probability of Cournot worst response (CWR), second best response (CSBR), and best response (CBR)

Algorithm	Animal	p(CWR)	p(CSBR)	p(CBR)
0	E	0.2508	0.2658	0.4834
	F	0.3062	0.2877	0.4061
1	E	0.2528	0.2141	0.5331
	F	0.3154	0.2992	0.3854
2	E	0.3170	0.3232	0.3597
	F	0.3319	0.3078	0.3603

($p = 0.525$). The value of mutual information between the choices of both players during the previous 2 trials and the animal's choice in the current trial was significantly higher in algorithm 1 than in algorithm 2, for both animals (t test, $p < 0.01$). The average mutual information in algorithm 1 was 0.282 and 0.068 bits for monkeys E and F, and the corresponding values were 0.053 and 0.033 bits in algorithm 2. Interestingly, during algorithm 1, both entropy and mutual information changed systematically, but the direction of change was opposite in two animals. In monkey E, entropy gradually decreased. Although this trend was not statistically significant, a regression analysis showed that mutual information increased significantly during the same period ($p < 0.05$), suggesting that the choice behavior of this animal became more dependent on the outcome of previous trials. In contrast, entropy increased and mutual information decreased significantly during algorithm 1 in monkey F ($p < 0.001$).

3.4. Comparison of learning models

As described above, the probability of CBR was higher than the probability of CSBR or that of CWR consistently in all animals and algorithms. This is consistent with at least two different types of learning processes. In reinforcement learning, for example, the probability of selecting a particular target is adjusted according to the utility of reward received from the same target in a given trial. This theory predicts that the probability of CBR would be relatively large following a win trial. Following a loss trial, this theory also predicts that the probability of CWR would be relatively low and that the probability of CBR and CSBR would increase similarly. Another possibility is that the probability of selecting a particular target is adjusted for all targets simultaneously according to the opponent's choice. The prediction of this belief learning theory for a choice after a win trial is the same as that of reinforcement learning, since both theories predict that the probability of CBR would increase. Unlike the reinforcement learning theory, the belief learning theory also predicts that following a loss trial, the probability of CBR would be higher than that of CSBR.

To determine which of these learning theories better describes the pattern of the animal's choice behavior, the probability of CBR, CSBR, or CWR was calculated separately according to the outcome of the previous trial (Fig. 8). For each triplet of frequencies for CBR, CSBR, and CWR following a particular outcome in the previous trial (i.e., loss, tie, or win), a χ^2 goodness-of-fit test was applied. According to this test, the null hypothesis that animals selected targets according to these three different strategies with equal probabilities was rejected, regardless of the previous outcome, for all animals and all algorithms ($p < 0.005$). Nevertheless, there were some interesting differences between the two animals. In algorithm 0, the probability of CBR was relatively high following a win in both animals. The probability of CBR following a loss or tie

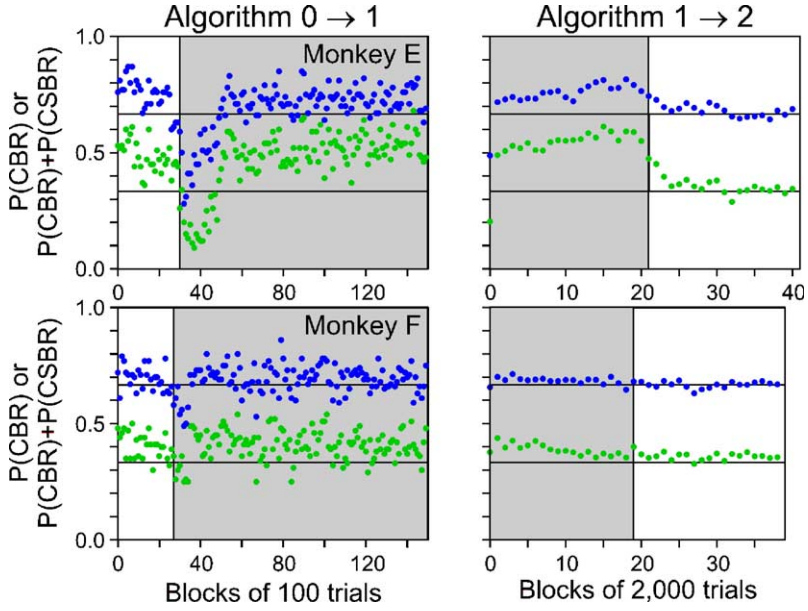

Fig. 6. The frequency of making the Cournot best response (CBR; green dots), and the frequency of making the Cournot best response or second best response (CSBR; blue dots). Same format as in Fig. 2.

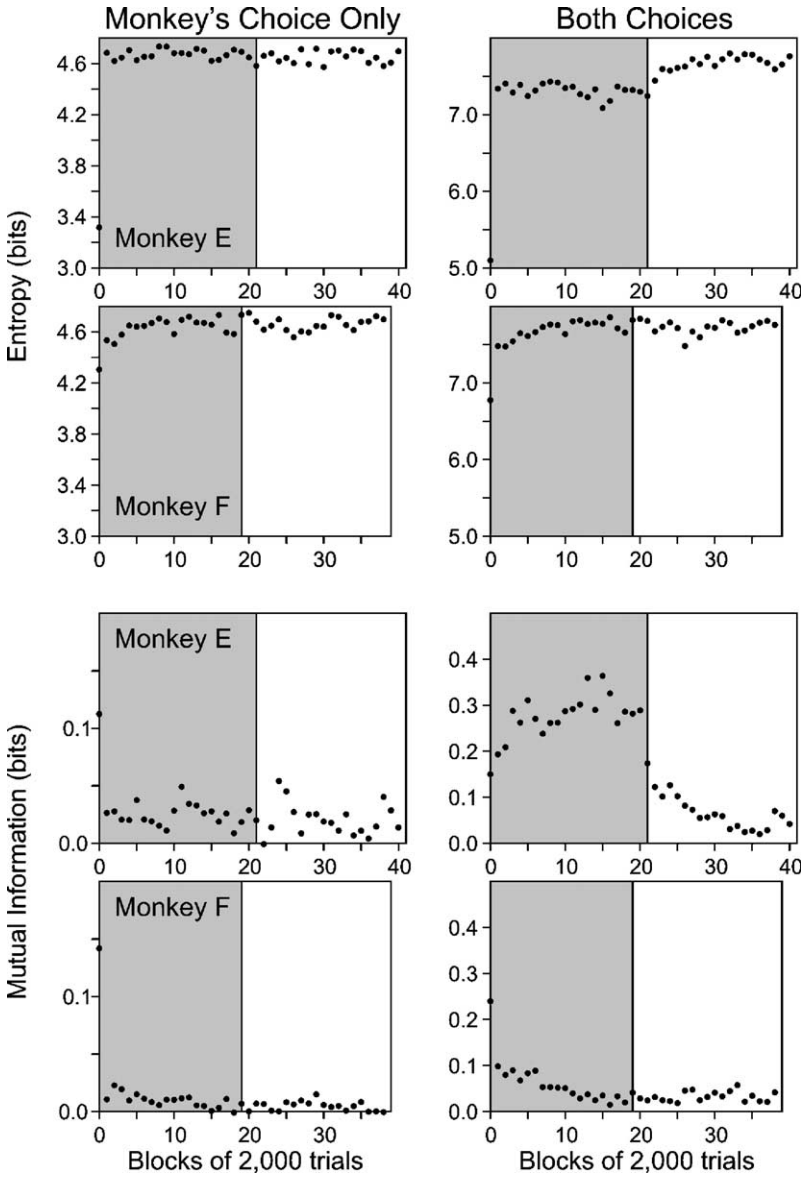

Fig. 7. Top: Entropy of the animal's choices in 3 successive trials (left), and entropy of the animal's choices in 3 successive trials combined with the computer's choices in the first two of such trials (right). Bottom: Mutual information between the animal's choices in 2 successive trials and its choice in the next trial (left), and mutual information between the choices of the animal and its opponent in 2 successive trials and the animal's choice in the next trial (left). Gray background corresponds to algorithm 1.

trial in monkey E was similar to the probability of CBSR or that of CWR. In contrast, monkey F displayed a stronger tendency to select the same target regardless of the outcome in the previous trial, as reflected by a high probability of CWR and that of CSBR following a loss trial and a tie trial, respectively (Fig. 8). During algorithm 1, monkey E displayed a relatively high probability of CBR, regardless of the outcome in the previous trial. The probability of CBR following a win, tie, and loss trial was 0.479, 0.458, and 0.668, respectively. This suggests that monkey E might have updated its strategy according to the rules of belief learning. In contrast, monkey F displayed a relatively high probability of CBR only after a win trial, indicating that it might have performed the task according to the rules of reinforcement learning. In both animals, the probability of CWR was lower than the probabilities of the two remaining strategies after a loss trial, and this is consistent with either of the learning algorithms. Interestingly, following a tie trial, the probability of CSBR was lower than the probability of CBR and that of CWR in both animals. Following the introduction of algorithm 2, the probabilities of CBR, CSBR, and CWR became more similar for all outcomes in both animals. This is not surprising, since a frequent adoption of such systematic strategies could be exploited by the computer opponent in algorithm 2.

To examine quantitatively how the animal's choice in a given trial was influenced by the previous choices of the animal and the computer, 3 different learning models were

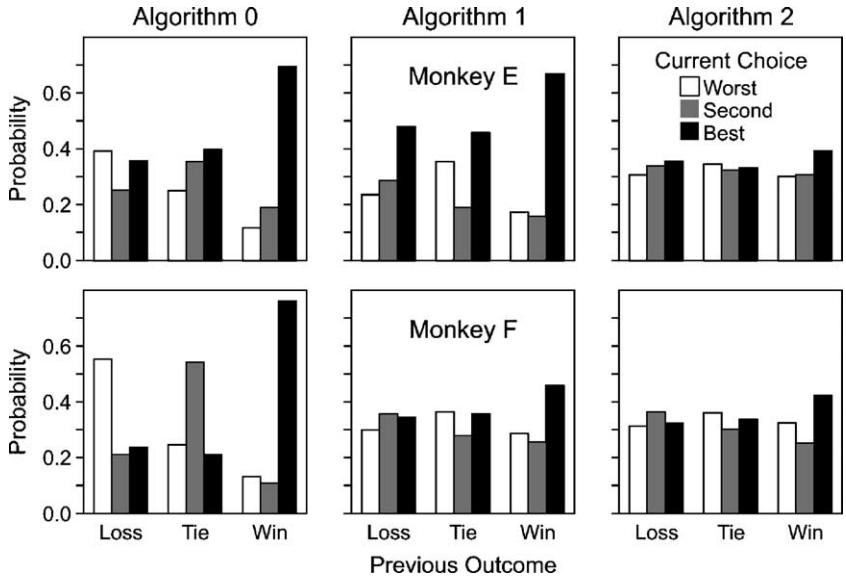

Fig. 8. Conditional probabilities of Cournot worst (white), second best (gray), and best (black) responses, computed separately for different outcomes of a preceding trial (abscissa).

fit to the data as described in the Methods. For algorithm 1, parameters of reinforcement learning indicated that there was a relatively large increase in the value function for the target chosen by the animal in a win trial (Δ_W), relative to the changes in the value functions for targets selected in a tie or loss trials (Table 5). For algorithm 1, decay rate (α) was relatively small in both animals, indicating that the outcomes of the trials immediately prior to the current trial exerted relatively large influences. In contrast, changes in the value functions became smaller in algorithm 2, and the decay rate increased. These results indicate that relatively large influences of the outcomes in the most recent trials found in algorithm 1 were reduced in algorithm 2. This is not surprising, since during algorithm 2, the computer utilized more information to exploit the biases in the animal's choice patterns. The results from the belief learning model (Table 6) were similar to those of the reinforcement learning model. The value function for the target that would have resulted in a win trial increased more than those for the other targets. As in the reinforcement learning model, decay rates were also larger for algorithm 2 than in algorithm 1. However, the log-likelihood for reinforcement learning model was substantially larger than that for belief learning model (Table 8), indicating that reinforcement learning described the animal's choice better. The reinforcement learning model provided a better fit to the data even after correcting for the improvement expected from the use of an additional free parameter, as evaluated by AIC and BIC (Table 9).

Both learning models described above place certain restrictions on how the value functions are updated after each trial. Reinforcement learning model, for example, updates the value function only for the target selected by the animal in a given trial, whereas belief learning model updates the value functions for all targets but these changes are independent of the animal's choice. A more general approach would be to allow the changes in the value function for each target to vary according to the animal's choice in the previous trial as well as its outcome. This model provided a significantly better fit to the data in all animals and algorithms, as indicated by the log-likelihood (Table 8) as well as AIC and BIC (Table 9). In addition, a close examination of the model parameters for this general learning model reveals some features that are consistent with simpler learning models described above (Table 7). For example, during algorithm 1 in monkey E, all the parameters associated with win targets (i.e., Δ_{LW}, Δ_{TW}, and Δ_{WW}) were more positive than those associated with tie targets. This implies that the value functions for a win target increased regardless of the animal's choice, and therefore is consistent with the assumptions of belief learning model. However, the predictions of the belief learning model were not generally borne out in other cases. For example, the signs for the changes in the value functions associated with

Table 5
Parameters of reinforcement learning model

Algorithm	Animal	Δ_L	Δ_T	Δ_W	α
1	E	−0.4764	−0.7343	1.3652	0.2805
	F	−0.0427	−0.1722	0.5190	0.6148
2	E	0.0202	0.0531	0.1356	0.9706
	F	0.0009	0.0150	0.0174	0.9975

Table 6
Parameters of belief learning model

Algorithm	Animal	Δ_T	Δ_W	α
1	E	−0.1553	0.7449	0.2025
	F	−0.0883	0.1726	0.5142
2	E	0.0365	0.1226	0.7112
	F	−0.0874	0.0600	0.7105

Table 7
Parameters of general learning model

Algorithm	1		2	
Animal	E	F	E	F
Δ_{LT}	0.2066	0.0541	−0.0019	−0.0095
Δ_{LW}	0.7013	0.0240	−0.0258	−0.0141
Δ_{TT}	−0.6266	−0.1907	0.0321	0.0159
Δ_{TW}	0.2332	−0.055 6	−0.0290	−0.0048
Δ_{WT}	−0.0522	−0.1194	0.0086	−0.0029
Δ_{WW}	1.3430	0.4470	0.1451	0.0282
α	0.2400	0.6423	0.9708	0.9933

tie targets (Δ_{LT}, Δ_{TT}, and Δ_{WT}) were not consistent. Even monkey E, for which there was some evidence for belief learning, displayed a positive value for the tie target after a loss trial, but negative values in other cases (Table 7). In addition, the change for the win target was consistently positive only after a win trial, and was negative in most cases. These results suggest that the most consistent bias in the animal's choice behavior was the tendency to select the same target again following a win trial.

To determine whether the above learning models account for the data better than simpler models that incorporate only the constant biases displayed by each animal, such as unequal probabilities to choose different targets or the conditional probabilities (i.e., probability of CBR, CSBR, or CWR), the log-likelihood was computed for two such models. The first model is referred to as the Bernoulli model, since each choice is treated in this model as an independent Bernoulli trial with a constant probability for each choice. The second model introduces constant probabilities for CBR, CSBR, and CWR, and therefore referred to as the Cournot model. The Bernoulli model has no memory, since the animal's choice in a given trial is not affected by any other events in the past, whereas the memory in the Cournot model is restricted to the last trial. The values of log-likelihood for these simple models were substantially lower than any of the learning models (Table 8), indicating that the animal's choice was influenced by the cumulative effects of previous trials integrated over multiple trials as assumed in the above learning models. Similarly, the values of AIC and BIC for these simple models showed smaller improvement compared to other learning models (Table 9).

Table 8
Improvement in the log-likelihoods of different learning models, relative to the prediction of Nash equilibrium

Algorithm	1		2	
Animal	E	F	E	F
Equilibrium (0)	−46,799	−43,815	−45,692	−43,984
Bernoulli (2)	285.8	753.0	412.3	624.1
Cournot (2)	3651.2	248.6	67.9	85.1
Reinforcement (4)	4472.8	818.6	1684.1	875.9
Belief (3)	3768.3	342.5	116.4	153.9
General (7)	4908.1	850.8	1725.2	933.3

The number of free parameters in each model is shown in parentheses.

Table 9
Changes in the Akaike's information criteria (AIC) and Bayesian information criterion (BIC), relative to the values obtained for the equilibrium model

Algorithm	1		2	
Animal	E	F	E	F
AIC/BIC				
Equilibrium (0)	93,597	87,628	91,385	87,968
AIC				
Bernoulli (2)	567.7	1502.0	820.7	1244.2
Cournot (2)	7298.5	493.2	131.8	166.1
Reinforcement (4)	8937.6	1629.2	3360.3	1743.9
Belief (3)	7530.7	679.0	226.7	301.7
General (7)	9802.2	1687.5	3436.4	1852.6
BIC				
Bernoulli (2)	550.4	1484.8	803.4	1227.0
Cournot (2)	7281.1	476.0	114.6	148.9
Reinforcement (4)	8903.0	1594.8	3325.7	1709.5
Belief (3)	7504.7	653.2	200.8	275.9
General (7)	9741.6	1627.4	3375.9	1792.4

4. Discussion

4.1. Models of learning in competitive games

Decision making in a social group is characterized by the fact that, in order to make optimal choices, players must take into consideration the predicted behavior of other decision makers in the group. However, this process may not be explicit. For example, in reinforcement learning, value functions, hence the probability of making various choices, are adjusted only by the outcome of a particular choice. Therefore, for this type of learning, the player only needs to know the outcome of its own choices, but not the choices of other players. Nevertheless, if a given game is played repeatedly, value functions are ultimately influenced by the choices of other players that affect the outcome of one's choice. In belief learning, the choices of other players can influence one's choice behavior more directly, since value functions for all choices and therefore the probabilities for choosing them can be simultaneously adjusted after each choice. In the present study, the choice behavior of monkeys playing a competitive game with three alternative choices was examined in order to gain insights into the nature of learning during decision making in non-human primates.

As in our previous studies [1,20], we first examined the choice behavior of each animal in a non-interactive and hence non-competitive situation where the computer's choice was random and independent of animal's choice. The computer selected each target with the probability of 1/3, which corresponds to the Nash equilibrium of the rock–paper–scissors game. Against this static strategy of the

computer, the animal's average payoff is fixed regardless of its decision-making strategy. This is a property of any Nash equilibrium in a zero-sum game. Therefore, it is not surprising that under this condition, each animal displayed an idiosyncratic pattern substantially deviating from the Nash equilibrium. Both animals displayed a bias against the target located directly above the fixation target, which was designated as Rock. Nevertheless, the animal's average payoff was optimal for this game. When the computer began exploiting the statistical biases displayed by the animal in its choice sequences (algorithm 1), the probabilities for choosing different targets became much more similar, although there was still significant bias against one of the targets. Interestingly, the choice behavior of the two animals diverged during the period of algorithm 1. One of the animals (monkey E) gradually increased its tendency to select the target that would beat the computer's choice in the previous trial, whereas this tendency decreased in monkey F. These changes were probably driven by factors intrinsic to the animals, since there were no visible changes in the average payoff during this period. The strategy to choose the best response to the choices of other players in the previous trial is referred to as the Cournot best response (CBR) [6–8,29]. Similarly, the remaining two choices other than the CBR are referred to as the Cournot second best response (CSBR) and the Cournot worst response (CWR). During algorithm 1, the average probability of CSBR for monkey E was 0.53 (Table 4), and the maximum value for a block of 2000 trials was 0.61 (Fig. 6). The probability of CBR was lower in monkey F, and did not exceed 0.44.

The choice behaviors of both animals displayed some features consistent with the predictions of reinforcement learning, especially during algorithm 1. For example, both animals were more likely to choose the same target again if they won in the preceding trial. A relatively low probability for choosing the same target after a loss trial is also consistent with reinforcement learning. Interestingly, the probability of choosing the same target as in the preceding trial was reduced after a tie trial. Within the framework of reinforcement learning, this implies that the value function for a given target is reduced after a tie, suggesting that there was a negative reward prediction error. In addition, one of the animals (monkey E) displayed some features associated with belief learning. For example, reinforcement learning predicts that following a loss trial, the probability for the CBR and CSBR would increase similarly. Although the behavior of monkey F was more or less consistent with this pattern, monkey E was substantially more likely to select the target that corresponds to the CBR in the next trial, as predicted by belief learning models. Since only two monkeys were tested in the present study, it is difficult to determine how often monkeys would adjust its decision-making strategies according to the rules of belief learning, or whether such tendency is a stable personality trait of a given animal that might be preserved across different types of games. This remains to be investigated in future studies.

We also explored the possibility that the animal's choice might be influenced by the cumulative effects of multiple trials in the past, rather than only by the animal's choice and its outcome in the previous trial. The results showed that regardless of the type of learning rules examined, the models based on the temporal integration of value functions accounted for the biases in the animal's choice behavior better than the simple Cournot dynamics. In addition, consistent with the analyses of conditional probabilities, the results also showed that the reinforcement learning model provided a substantially better description for all algorithms and animals than the belief learning model. However, in all cases, the best fit was provided by a model that allows the value functions for different targets to be adjusted according to the previous choice of the animal and its outcome. The essential feature of reinforcement learning is that, in a given trial, only the value function for the action selected by the player is adjusted. In contrast, belief learning allows the value functions for every possible action to be adjusted after each choice, according to the hypothetical payoffs determined by the choices of other players. The general model we examined, therefore, incorporated the features of both reinforcement learning and belief learning models, similar to the experience-weighted attraction (EWA) model proposed by Camerer and Ho [6]. In the EWA model, the changes in the value functions or attractions are determined by the monetary payoffs available to a given player. In the present study, this constraint was removed, and we allowed the parameters of our general learning model to vary more freely, since we did not have any reason to believe that the changes in the value functions are proportional to the amount of juice given to the animal. For example, following a tie trial, the probability that the animal would choose the same target was reduced in algorithm 1, suggesting that the value function was reduced for a tie target. Consistent with this result, in our general learning model, the parameter for the tie target after a tie trial was negative for algorithm 1.

The results from the present study showed that the choice behavior of monkeys during a competitive game can be described better by a reinforcement learning model than by a belief learning model. Similar to the findings in the present study, human players often display systematic deviations from the predictions of Nash equilibrium even during relatively simple games. Results from previous studies suggest that human players might rely more on reinforcement learning, rather than believe learning, during a variety of constant-sum games [12,24], although in some cases, reinforcement learning and belief learning models performed more similarly [13]. As described above, these two different types of learning models share some common features, such as the use of intermediate variables that are related to the actual probabilities of choices. The finding that a more general learning model provides a better description of the choice behavior of monkeys in the present study is consistent with the EWA model of Camerer

and Ho [6]. These results suggest that models incorporating features of both learning models might better account for the animal's behavior. Thus, the choice behavior of both humans and monkeys during a competitive game may not conform to strict assumptions of reinforcement learning or belief learning model. Whether more complex learning models are required to describe choice behaviors of humans and other animals needs to be investigated further [5].

4.2. Implications for the neural correlates of learning during decision making

Reinforcement learning algorithms seek an optimal sequence of actions in a dynamic environment. In this framework, value functions are adjusted according to the reward prediction error [39] or actual payoff [12], and a given action is selected according to a probability that is monotonically related to the value function for the same action, for example, through the softmax transformation. Although belief learning models are based on a different set of assumptions as to how the value functions are adjusted, they share a common feature with reinforcement learning models in that the probability of choosing a particular action is based on a set of hypothetical values, such as expected payoffs [7,13,24] or attractions [6], that are adjusted according to the choices of other players. Therefore, in both reinforcement and belief learning models, signals related to the actual or hypothetical outcome must be generated after each choice, and they must be temporally integrated to compute value functions or attractions that are directly related to the choice probabilities. Therefore, reinforcement learning models and other learning theories of economic decision making provide a useful framework in which to investigate the underlying neural processes of decision making [32].

Indeed, transient neural signals related to the outcome of a behavioral choice as well as signals related to the amount of expected reward have been found in various regions of the primate brain. Transient activity of dopamine neurons in the ventral midbrain can signal reward prediction errors [33]. In addition, neurons in the medial frontal cortex, such as the supplementary eye field [34,37] or the anterior cingulate cortex [18], provide information about the outcome of a behavioral choice. Nevertheless, the function of the dopamine neurons and other neurons carrying transient signals related to the choice outcome is not yet fully understood. For example, it has been recently demonstrated that dopamine neurons transmit information about the uncertainty of upcoming reward in addition to the error in predicting upcoming reward, raising the possibility that uncertainty-related activity of dopamine neurons might facilitate learning in an unpredictable environment [14]. The results from the present study provide more specific predictions for the transient signals used to update the animal's decision-making strategy during a competitive game. In our study, the probability of selecting the same target was reduced after a tie trial in algorithm 1, suggesting that a relatively small reward might produce a negative reward prediction error even when it is quite close to the average value. The results from the present study also suggest that following a particular action, value functions for multiple actions may be adjusted simultaneously, as suggested in belief learning or a more general learning model. This might be achieved by at least two different mechanisms. One possibility is that changes in the value functions for multiple actions might be reflected in the heterogeneity of dopamine neurons signaling reward prediction errors. Alternatively, a particular pattern in the activity of dopamine neurons might be interpreted differently by different neurons in the striatum or the cortex to update the value functions of multiple choices. Further neurophysiological studies are required to distinguish between these alternative scenarios.

Another type of signals that play a central role in reinforcement learning is value function. A value function is an estimate for the temporally discounted sum of all future rewards resulting from a course of actions selected by the animal's current decision-making strategy, and therefore differs from the expected reward resulting immediately from a given action [39]. Nevertheless, they are closely related, since for a trial with a single action, they are equivalent. Therefore, a number of brain areas in which neurons often modulate their activity according to expected reward might be involved in computing value functions and/or using such signals to select the optimal sequence of actions. These areas include the prefrontal cortex [1,11,21,28,35], the posterior parietal cortex [10,26,38], and the basal ganglia [9,19]. However, the neural correlates of value functions and how these signals might be used for the purpose of selecting an optimal sequence of actions are still largely unknown. Nevertheless, the results from the present study suggest that value functions for unselected actions might be updated according to hypothetical payoffs expected from observing the behaviors of other players in a social group. Primate models of competitive interactions as utilized in the present study might, therefore, provide a useful tool to investigate these issues.

Acknowledgments

We thank Lindsay Carr and Ted Twietmeyer for their technical assistance and John Swan-Stone for computer programming. This study was supported by the National Institute of Health.

References

[1] D.J. Barraclough, M.L. Conroy, D. Lee, Prefrontal cortex and decision making in a mixed-strategy game, Nat. Neurosci. 7 (2004) 404–410.
[2] K. Binmore, J. Swierzbinski, C. Proulx, Does minimax work? An experimental study, Econ. J. 111 (2001) 445–464.

[3] D.V. Budescu, A. Rapoport, Subjective randomization in one- and two-person games, J. Behav. Decis. Mak. 7 (1994) 261–278.
[4] K.P. Burnham, D.R. Anderson, Model selection and multimodel inference, A Practical Information—Theoretic Approach, Second ed., Springer-Verlag, New York, 2002.
[5] C.F. Camerer, Behavioral Game Theory: Experiments in Strategic Interaction, Princeton Univ. Press, Princeton, 2003.
[6] C.F. Camerer, T.-H. Ho, Experience-weighted attraction learning in normal form games, Econometrica 67 (1999) 827–874.
[7] Y.-W. Cheung, D. Friedman, Individual learning in normal form games: some laboratory results, Games Econ. Behav. 19 (1997) 46–76.
[8] A. Cournot, Recherches sur les principes mathematiques de la theorie des richesses, 1938, in: N. Bacon (Ed.), Researches into the Mathematical Principles of the Theory of Wealth, English edition, Macmillan, New York, 1897.
[9] H.C. Cromwell, W. Schultz, Effects of expectations for different reward magnitudes on neuronal activity in primate striatum, J. Neurophysiol. 89 (2003) 2823–2838.
[10] M.C. Dorris, P.W. Glimcher, Activity in posterior parietal cortex is correlated with the relative subjective desirability of action, Neuron 44 (2004) 365–378.
[11] R. Elliott, J.L. Newman, O.A. Longe, J.F.W. Deakin, Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study, J. Neurosci. 23 (2003) 303–307.
[12] I. Erev, A.E. Roth, Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria, Am. Econ. Rev. 88 (1998) 848–881.
[13] N. Feltovich, Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games, Econometrica 68 (2000) 605–641.
[14] C.D. Fiorillo, P.N. Tobler, W. Schultz, Discrete coding of reward probability and uncertainty by dopamine neurons, Science 299 (2003) 1898–1902.
[15] D. Fudenberg, D.K. Levine, The Theory of Learning in Games, MIT Press, Cambridge, 1998.
[16] P.W. Glimcher, Decisions, Uncertainty, and the Brain, MIT Press, Cambridge, 2003.
[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[18] S. Ito, V. Stuphorn, J.W. Brown, J.D. Schall, Performance monitoring by the anterior cingulate cortex during saccade countermanding, Science 302 (2003) 120–122.
[19] R. Kawagoe, Y. Takikawa, O. Hikosaka, Expectation of reward modulates cognitive signals in the basal ganglia, Nat. Neurosci. 1 (1998) 411–416.
[20] D. Lee, M.L. Conroy, B.P. McGreevy, D.J. Barraclough, Reinforcement learning and decision making in monkeys during a competitive game, Cognit. Brain Res. 22 (2004) 45–58.
[21] M.I. Leon, M.N. Shadlen, Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque, Neuron 24 (1999) 415–425.
[22] G.A. Miller, Note on the bias of information estimates, in: H. Quastler (Ed.), Information Theory in Psychology, Free Press, Glencoe, 1955, pp. 95–100.
[23] D. Mookherjee, B. Sopher, Learning behavior in an experimental matching pennies game, Games Econ. Behav. 7 (1994) 62–91.
[24] D. Mookherjee, B. Sopher, Learning and decision costs in experimental constant sum games, Games Econ. Behav. 19 (1997) 97–132.
[25] J.F. Nash, Equilibrium points in n-person games, Proc. Natl. Acad. Sci. 36 (1950) 48–49.
[26] M.L. Platt, P.W. Glimcher, Neural correlates of decision variables in parietal cortex, Nature 400 (1999) 233–238.
[27] J. Robinson, An iterative method of solving a game, Ann. Math. 54 (1951) 296–301.
[28] M.R. Roesch, C.R. Olson, Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex, J. Neurophysiol. 90 (2003) 1766–1789.
[29] T.C. Salmon, An evaluation of econometric models of adaptive learning, Econometrica 69 (2001) 1597–1628.
[30] R. Sarin, F. Vahid, Predicting how people play games: a simple dynamic model of choice, Games Econ. Behav. 34 (2001) 104–122.
[31] Y. Sato, E. Akiyama, J.D. Farmer, Chaos in learning a simple two-person game, Proc. Natl. Acad. Sci. 99 (2002) 4748–4751.
[32] W. Schultz, Neural coding of basic reward terms of animal learning theory, game theory, and microeconomics and behavioral ecology, Curr. Opin. Neurobiol. 14 (2004) 139–147.
[33] W. Schultz, A. Dickinson, Neuronal coding of prediction errors, Annu. Rev. Neurosci. 23 (2000) 473–500.
[34] H. Seo, D.J. Barraclough, B.P. McGreevy, D. Lee, Role of supplementary eye field in decision making during a competitive game, Program No. 87.3. 2004 Abstract Viewer/Itinery Planner, Society for Neuroscience, Washington, DC, 2004, (Online).
[35] M. Shidara, B.J. Richmond, Anterior cingulate: single neuronal signals related to degree of reward expectancy, Science 296 (2002) 1709–1711.
[36] G.W. Snedecor, W.G. Cochran, Statistical Methods, Eighth ed., Iowa State Univ. Press, Ames, 1989.
[37] V. Stuphorn, T.L. Taylor, J.D. Schall, Performance monitoring by the supplementary eye field, Nature 408 (2000) 857–860.
[38] L.P. Sugrue, G.S. Corrado, W.T. Newsome, Matching behavior and the representation of value in the parietal cortex, Science 304 (2004) 1782–1787.
[39] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, 1998.
[40] J. von Neumann, O. Morgenstern, The Theory of Games and Economic Behavior, Princeton Univ. Press, Princeton, 1944.
[41] P. Zak, Neuroeconomics, Philos. Trans. R. Soc. London, B 359 (2004) 1737–1748.