Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 19 (2006) 1075–1090
Neural Networks
www.elsevier.com/locate/neunet

2006 Special Issue

Neural mechanism for stochastic behaviour during a competitive game

Alireza Soltani[a,1,2], Daeyeol Lee[b,1,3], Xiao-Jing Wang[a,*,1]

[a] Department of Physics and Volen Center for Complex Systems, Brandeis University, Waltham, MA 02454, USA
[b] Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, NY 14627, USA

Received 30 November 2005; accepted 22 May 2006

Abstract

Previous studies have shown that non-human primates can generate highly stochastic choice behaviour, especially when this is required during a competitive interaction with another agent. To understand the neural mechanism of such dynamic choice behaviour, we propose a biologically plausible model of decision making endowed with synaptic plasticity that follows a reward-dependent stochastic Hebbian learning rule. This model constitutes a biophysical implementation of reinforcement learning, and it reproduces salient features of behavioural data from an experiment with monkeys playing a matching pennies game. Due to interaction with an opponent and learning dynamics, the model generates quasi-random behaviour robustly in spite of intrinsic biases. Furthermore, non-random choice behaviour can also emerge when the model plays against a non-interactive opponent, as observed in the monkey experiment. Finally, when combined with a meta-learning algorithm, our model accounts for the slow drift in the animal's strategy based on a process of reward maximization.

Keywords: Decision making; Reward-dependent stochastic Hebbian learning rule; Reinforcement learning; Meta-learning; Synaptic plasticity; Game theory

1. Introduction

Decision making has been studied using a variety of paradigms in multiple disciplines. For example, economists have often employed tasks based on multiple gambles or lotteries to investigate how decisions are influenced by the decision maker's attitude towards uncertainty (Kahneman & Tversky, 1979). Behavioural ecologists have approached the problem of decision making in the context of foraging (Stephens & Krebs, 1986), whereas psychologists have frequently investigated the choice behaviour using a concurrent schedule of reinforcement (Herrnstein, Rachlin, & Laibson, 1997). All of these paradigms, however, are designed to investigate the process of decision making in a socially isolated individual. In contrast, decision making in a socially interactive context introduces a new principle of optimality (von Neumann & Morgenstern, 1944), since the outcome of one's decision can be influenced by the decisions of others in the same group.

Recently, neuroscientists have begun to investigate the neural basis of decision making using the behavioural paradigms rooted in these various disciplines (see Lee (2006)). In some cases, these studies were carried out in non-human primates, allowing the investigators to examine the activity of individual neurons during various types of decision making. For example, Sugrue, Corrado, and Newsome (2004) found that activity of neurons in intraparietal cortex reflects the relative income from the target in their receptive fields during an oculomotor foraging task based on a concurrent variable-interval schedule. Using an approach based on a standard economic choice theory, McCoy and Platt (2005) found that neurons in posterior cingulate cortex modulate their activity according to the uncertainty of reward expected from a particular target. Barraclough, Conroy, and Lee (2004) and Dorris and Glimcher (2004) have examined the pattern of neural activity in dorsolateral prefrontal cortex and posterior parietal cortex, while the animal interacted competitively with a computer opponent.

* Corresponding author. Tel.: +1 203 785 6297; fax: +1 203 785 5263.
E-mail addresses: alireza.soltani@yale.edu (A. Soltani), daeyeol.lee@yale.edu (D. Lee), xjwang@yale.edu (X.-J. Wang).
[1] Current address: Department of Neurobiology and Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, CT 06520, USA.
[2] Tel.: +1 203 785 6302; fax: +1 203 785 5263.
[3] Tel.: +1 203 785 3527; fax: +1 203 785 5263.

0893-6080/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2006.05.044

In the present study, we focus on the choice behaviour of monkeys playing a simple competitive game, known as the matching pennies (Barraclough et al., 2004). During this task, monkeys were required to choose one of two visual targets in an oculomotor free-choice task, and they obtained a reward only if they chose the same target as the computer opponent in a given trial. The optimal strategy during this game is to choose the two targets randomly and with equal probability, and therefore requires random sequences of choices. Some studies have shown that in general people are relatively poor in generating a random sequence of choices (Bar-Hillel & Wagenaar, 1991; Camerer, 2003), but with feedback they can learn to generate sequences that can pass standard randomness tests (Neuringer, 1986). More interestingly, it has been found that people can generate more random sequences of choices, if they are engaged in a competitive game (Rapoport & Budescu, 1992). Nevertheless, the neural mechanisms responsible for the generation of such a high level of stochastic behaviour is unknown.

Consistent with these behavioural findings obtained in human subjects, the results from the previous study (Barraclough et al., 2004) showed that monkeys can learn to generate nearly random sequences of choices when they receive reward feedback regarding their performance. In addition, the degree of randomness in the animal's choice behaviour varied according to the amount of information utilized by the computer opponent to predict the animal's choice (Lee, Conroy, McGreevy, & Barraclough, 2004). In other words, the animal's behaviour became more random, when the computer utilized additional information about the animal's previous choices and their outcomes. Furthermore, a simple reinforcement learning model was proposed to account for the fact that the animal's choice was systematically influenced by the computer's choices in previous trials (Lee et al., 2004).

Here, we show that a biophysically-plausible network model of decision making endowed with plastic synapses can not only generate random sequences of choices, but also capture other important features of animal's choice behaviour during the matching pennies task. First, monkeys displayed a bias in their choice behaviour when playing against a non-responsive computer opponent selecting its target randomly, regardless of the animal's behaviour. To understand the nature of such a bias, we analyze the steady-state behaviour of the reinforcement learning model described in Lee et al. (2004) and that of our model, and derive the conditions under which a biased and non-random choice behaviour can emerge. Second, when the computer opponent was partially exploitive and used only the information about the animal's previous choices but not their outcomes, the animal's choice strategy displayed a slow drift over the period of many days. We implement a meta-learning algorithm (Schweighofer & Doya, 2003) in our model, and show that it can account for the gradual change in the animal's strategy. To our knowledge, this study is the first to propose a possible explanation for the slow, gradual behavioural change on the timescale of many days observed experimentally.

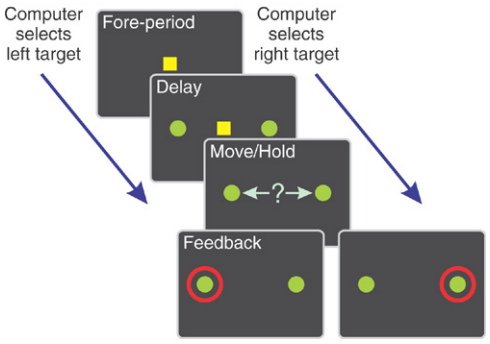

Fig. 1. Spatial layout and temporal sequence of the free-choice task.

2. Choice behaviour of monkeys in the matching pennies game

2.1. Experimental methods

A detailed description of the experimental methods used to collect the behavioural data during a matching pennies task has been published previously (Lee et al., 2004). In the following, the behavioural task used in this study is only briefly described. Three rhesus monkeys (C, E, and F) were trained to perform in an oculomotor free-choice task according to the rule of the matching pennies game (Fig. 1).

The animal began each trial by fixating a small yellow square at the centre of the computer screen. After a 0.5 s fore-period, two identical green disks were presented along the horizontal meridian. After a delay period of 0.5 s, the central square was extinguished and the monkey was required to make 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 red circle was displayed for 0.2 s around the target that the computer had selected. The monkey was rewarded with a drop of fruit juice if it selected the same target as the computer.

The strategy or algorithm used by the computer opponent during this matching pennies game increased its complexity through three successive stages. In the first stage, referred to as algorithm 0, the computer selected one of the two targets randomly, each with 50% probability, and therefore, the animal's expected payoff was not influenced by its own choice behaviour. The computer's strategy in algorithm 0 corresponds to the Nash equilibrium of the matching pennies game. In the next stage (algorithm 1), the computer used the entire sequence of the animal's previous choices in a given day to predict the monkey's next choice by testing a set of hypotheses. The conditional probability that the monkey would choose each target given the monkey's choices in the preceding N trials ($N = 0$ to 4) was calculated and tested against a null hypothesis that this probability is 0.5 (binomial test, $p < 0.05$). If none of these hypotheses was rejected, the computer selected its target randomly as in algorithm 0. If one or more hypotheses were rejected, then the computer biased its target selection using the conditional probability with the largest deviation from 0.5 that was statistically significant. If the conditional probability of choosing a given target that was selected by this procedure was

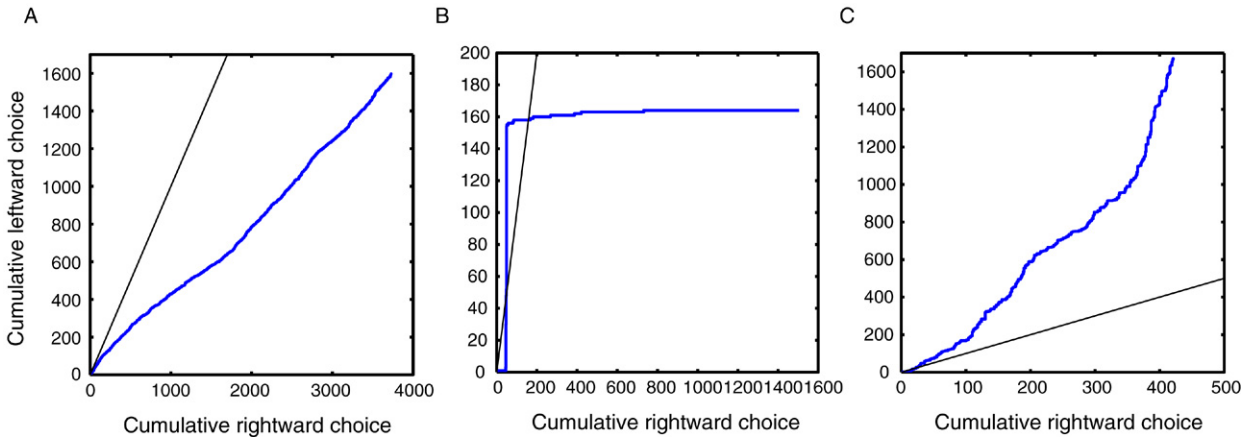

Fig. 2. Instability in monkeys' choice behaviour in algorithm 0. The cumulative choices of the leftward target is plotted against the cumulative choices of the rightward target in three different monkeys (in the last 2 days of algorithm 0). (A) Choice behaviour in monkey C was the most stable but it was biased toward the rightward target. (B) Monkey E showed very unstable choice behaviour, such that at the end it chose only the rightward target. (C) In monkey F, the choice behaviour was biased toward the leftward target. There are switches between the two targets but lengths of consecutive choices on the leftward target are larger. The black line corresponds to the choices made with equal probabilities.

p, the computer selected the same target with the probability of $1 - p$. Therefore, in order for the monkey to obtain the maximum reward in algorithm 1, the two targets should be selected with equal probabilities and independently from the animal's previous choices. In the final stage (algorithm 2), the computer used the entire history of the animal's choice and reward in a given session to predict the monkey's choice in the next trial. Thus, in algorithm 2, a set of conditional probabilities that the monkey would choose a given target given the monkey's choices and their payoffs in the preceding N trials ($N = 1$ to 4) were calculated and tested, as well as the conditional probabilities tested in algorithm 1. If none of these hypotheses deviated significantly from 0.5, then the computer selected each target randomly with equal probability (50%). Otherwise, the computer biased its target selection according to the same rule used in algorithm 1. In algorithm 2, therefore, the monkey 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.

2.2. Summary of animal's choice behaviour

The choice behaviour of monkeys during the matching pennies game has been described previously (Barraclough et al., 2004; Lee et al., 2004). Here, the results that are particularly relevant to the behaviour of our biophysical network model are briefly summarized. First, each of the animals tested in this experiment displayed a bias to choose one of the targets more frequently during algorithm 0 in which the computer opponent passively selected both targets with equal probabilities without exploiting the statistical biases displayed by the animal (Fig. 2).

When the computer started to exploit the animal's preference for one of the targets in algorithm 1, such biases rapidly diminished. However, these biases were not entirely removed. The probability of choosing the rightward target in algorithm 1 was 0.489, 0.511, and 0.490, for monkeys C, E, and F, respectively, and they were all significantly different from 0.5($p < 0.01$). The probability of choosing each target in algorithm 2 was also close to 0.5, although it was significantly different from 0.5. In addition to the bias to choose one of the targets more frequently, animals also displayed a tendency to choose the same targets in two consecutive trials or alternate between the two targets (Lee et al., 2004).

During the matching pennies game, the animals also displayed the bias to choose the same target chosen by the computer in the previous trial. This is equivalent to choosing the target that was rewarded or would have been rewarded in the previous trial, and is referred to as a win-stay-lose-switch (WSLS) strategy. The probability that the monkey would choose its target according to this strategy was significantly larger than 0.5 ($p < 10^{-10}$) for all monkeys and for all algorithms, except for algorithm 0 in monkey E (Table 1). The probability of WSLS strategy was especially high in algorithm 1, because in this algorithm the computer did not examine the reward history and monkeys were therefore not penalized for frequently using WSLS strategy. Interestingly, the probability of WSLS strategy increased gradually during algorithm 1, and this trend was statistically significant in all monkeys (Fig. 3). Following the introduction of algorithm 2, the probability of WSLS strategy declined towards 0.5 in all monkeys although this bias remained statistically significant even in algorithm 2 for all monkeys (Table 1).

The change in monkeys' choice behaviour over the course of the experiment has been previously quantified using entropy (Lee et al., 2004). In particular, entropy which was computed based on the bivariate sequence consisting of the animal's choices and computer's choices, decreased during algorithm 1 mirroring the gradual increase in the use of WSLS strategy, and increased after the introduction of algorithm 2. In summary, these results show that animals learned to adopt new strategies over the course of the experiment in order to maximize their reward.

Table 1
Probabilities of choice, reward, and a few simple strategies for monkeys' choice behaviour in the experiment of matching pennies

Algorithm	Monkey	P (Right)	P (Reward)	P_{ind}(Same)	P (Same)	P (WSLS)
0	C	0.7002*	0.4969	0.5802	0.5726	0.6674*
	E	0.9017*	0.4985	0.8228	0.9808*	0.5081
	F	0.3320*	0.4892	0.5565	0.6727*	0.5718*
1	C	0.4886*	0.4894*	0.5003	0.5202*	0.6462*
	E	0.5110*	0.4911*	0.5002	0.4963	0.7314*
	F	0.4899*	0.4951*	0.5002	0.5043	0.6333*
2	C	0.4857*	0.4766*	0.5004	0.5137*	0.5478*
	E	0.4911*	0.4695*	0.5002	0.4878*	0.5345*
	F	0.4717*	0.4778*	0.5016	0.4693*	0.5650*

P (Right), probability of choosing the right-hand target. P (Reward), probability of reward. P_{ind}(Same), probability of choosing the same target as in the previous trial estimated from P (Right); (P_{ind}(Same) $= P_R^2 + (1 - P_R)^2$). P (Same), actual probability of choosing the same target as in the previous trial. P (WSLS), probability of using the win-stay-lose-switch strategy. The asterisk indicates that the deviation from the null hypothesis is significant at the level of $p = 0.01$. The null hypothesis was $p = 0.5$ in all cases, except that for P (Same), it was P_{ind}(Same).

3. Stability of equilibrium strategy

In algorithm 0, the computer chose the two targets randomly with an equal probability, independent of the monkey's choice, which corresponds to the Nash equilibrium in the matching pennies game. Thus, in this condition, the animal's choice was always rewarded with 50% probability for both targets. Nevertheless, each animal displayed a significant bias for choosing one of the two targets (Figs. 2 and 3), indicating that they deviated from the Nash equilibrium. This bias was extreme for monkey E, but it was statistically significant in all three animals. The deviation of the animal's choice behaviour from the Nash equilibrium in algorithm 0 is consistent with the observation that human subjects do not adopt an equilibrium strategy if the computer opponent plays according to the Nash equilibrium (Liberman, 1962; Messick, 1967). Similarly, for a matching pennies game played repeatedly between two human subjects, if one subject approaches the equilibrium, the other subject often deviates from it (Mookherjee & Sopher, 1994). These observations have been interpreted as lack of incentive for the subject to adopt the equilibrium strategy. Another possible explanation is that the underlying learning mechanism makes the equilibrium strategy unstable. We examined this possibility using the framework of the reinforcement learning model, which has been used to model the choice behaviour in the matching pennies game (Barraclough et al., 2004; Lee et al., 2004).

In the framework of reinforcement learning (Sutton & Barto, 1998), the choice behaviour in a matching pennies game can be determined probabilistically by two value functions corresponding to the two alternative targets. The value functions provide estimates for the reward expectation for each choice in a given trial, and are updated after each trial according to its outcome as follows

$$V_i(t+1) = \alpha V_i(t) + \Delta_i(t) \tag{1}$$

where α is a decay factor ($0 \le \alpha \le 1$), and $\Delta_i(t)$ refers to the change in the value function for the choice i at time t. Here, it is assumed that the initial values for V_i are zero, $\Delta_i(t) = \Delta_1$ if the target i is selected and rewarded, $\Delta_i(t) = \Delta_2$ if the target i is selected and not rewarded, and $\Delta_i(t) = 0$ if the target i is not selected. The probability of choosing the rightward target at a given time is equal to

$$P_R = \frac{1}{1 + \exp(-(V_R - V_L))}. \tag{2}$$

The difference between this model and more standard reinforcement learning models like Q-learning is that in this model both value functions are updated in each trial. Because the choice probability depends only on the difference between the two value functions, this two-dimensional model can be reduced to a model with one dynamical variable. If we define $U \equiv V_R - V_L$, then U is updated by:

$$U(t+1) = \alpha U(t) \pm \Delta_1 \quad (\text{or } \pm\Delta_2 \text{ in unrewarded trials}) \tag{3}$$

where the plus (minus) sign corresponds to the case where the rightward (leftward) target is selected. If the computer opponent selects between the two targets randomly with the same probabilities, then on average the update rule for U can be described as

$$U(t+1) = \alpha U(t) + (P_R - 0.5)\Delta \tag{4}$$

where $\Delta \equiv \Delta_1 + \Delta_2$. This dynamical system reaches a steady state when $U(t+1) = U(t)$, and the steady-state value of U is given by

$$U_{ss} = \frac{(P_R - 0.5)\Delta}{(1 - \alpha)}. \tag{5}$$

This equation shows how the steady state of U depends on the choice probability which itself is a function of U (see Eq. (2)). So, in order to determine the steady-state choice behaviour of the model, the last equation should be solved for P_R.

First, if $\Delta = 0$, then the only solution for the Eq. (3) is $U_{ss} = 0$ which is equivalent to $P_R = 0.5$. This solution is a stable steady-state for the choice behaviour dynamics, so under this condition the model selects the two choices with equal probability. If Δ is positive, Eq. (3) always has one trivial solution at $P_R = 0.5$ and this solution is a stable steady-state if $\frac{(1-\alpha)}{\Delta} \ge 0.25$ (Fig. 4A). On the other hand, if $\frac{(1-\alpha)}{\Delta} < 0.25$, then $P_R = 0.5$ solution becomes an unstable steady-state and two new stable steady-states emerge (Fig. 4B). So, under

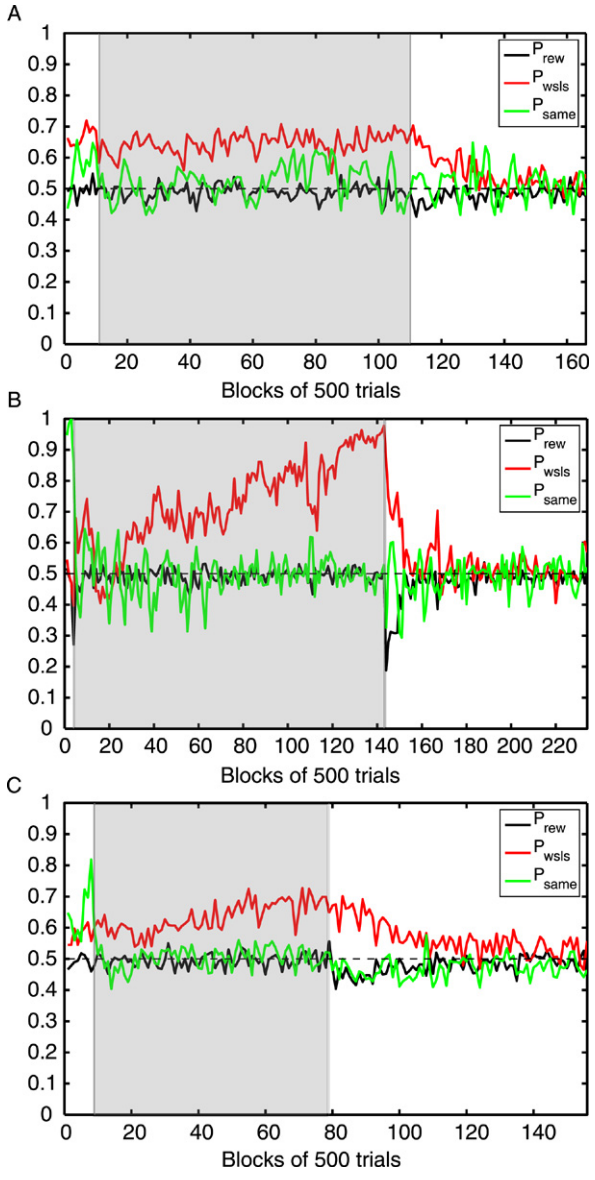

Fig. 3. Slow change in monkeys' choice behaviour over the course of the experiment. In each panel the average probability of choosing the same target as in the previous trial, P_{same}, probability of using WSLS strategy, P_{wsls}, and probability of harvesting reward, P_{rew}, are plotted for each monkey; (A): monkey C, (B) monkey E, (C) monkey F. The gradual change in P_{wsls} is present in all monkeys' choice behaviour but it is most prominent in the behaviour of monkey E. Each probability is computed over a block of 500 trials. To distinguish the behaviour in three different algorithms, blocks in algorithm 1 are shaded.

this condition, the choice behaviour is biased toward one of the targets with a choice probability determined by the value of $\frac{(1-\alpha)}{\Delta}$. If Δ is negative, Eq. (3) has only one steady-state solution at $P_R = 0.5$ and this solution is stable if $\frac{(1+\alpha)}{|\Delta|} \geq 0.25$ and unstable otherwise (Fig. 4C and D). Notice that in this case, an unstable steady-state at $P_R = 0.5$ can be achieved only if the value for $|\Delta_2|$ is large enough ($\Delta_2 < 0$). This means that in a trial with no reward, the value function for the chosen target will be reduced sufficiently so it is more likely that the model chooses the other target in the next trial.

Table 2
Maximum log likelihood estimates of reinforcement learning model parameters for monkeys' choice behaviour in algorithm 0

Monkey	Δ_1	Δ_2	α	$\frac{1-\alpha}{\Delta_1+\Delta_2}$
C	0.0308*	0.0030	0.9921*	0.2337
E	2.0227*	1.1515*	0.6843*	0.0995
F	0.8684*	0.1409*	0.7704*	0.2275

The parameter values were obtained using data from the last 2 days of algorithm 0. The asterisks indicate the estimated model parameters that were significantly different from zero ($p < 0.01$).

To determine whether this analysis can account for the biased choice behaviour seen in algorithm 0, the model parameters for the above reinforcement learning model were examined (see Lee et al. (2004) and Table 2). We found that for the model parameters estimated for the choice behaviour in algorithm 0, Δ is positive and $\frac{(1-\alpha)}{\Delta} < 0.25$, and this was true for all animals. These results suggest that a biased choice behaviour observed in algorithm 0 can emerge under certain conditions as a result of learning dynamics.

4. A biophysical model for probabilistic decision making

4.1. Description of network model

Details about the architecture of our network model can be found in Wang (2002) (see also Brunel and Wang (2001)). Briefly, the decision-making network consists of 2000 integrate-and-fire neurons (1600 excitatory, and 400 inhibitory) which are grouped into three populations of excitatory neurons and a single population of inhibitory neurons (Fig. 5). Two of the excitatory populations (240 neurons each) are selective to the leftward and rightward targets and the third excitatory population (1120 neurons) is nonselective. Each neuron receives input and sends output through realistic AMPA, NMDA, and GABA receptors (Wang, 2002). In addition to the recurrent synaptic currents, all neurons receive a background input from 800 afferent neurons that are external to the decision network and have background firing rate of 3 Hz. In this network, the two populations of excitatory neurons compete against each other through the population of inhibitory neurons. This inhibition produces the so-called winner-take-all property, so that a few hundred milliseconds after the onset of sensory stimulus, the activity in one population increases and suppresses the activity in the other population. Consequently, the network's choice in each trial can be read out according to which neural population has a higher firing rate.

In the simulation of this network's behaviour during the matching pennies game, neurons in the two selective populations receive an additional input following the presentation of the two choice targets. This input is mediated by 0.625% of afferent neurons that increase their firing rates from 3 to 12 Hz. Synapses between these afferent neurons and the excitatory neurons are assumed to be plastic and binary (O'Connor, Wittenberg, & Wang, 2005; Petersen, Malenka, Nicoll, & Hopfield, 1998), with two discrete states, potentiated state with peak conductance of $g_+ = 3.0$ nS and a depressed state with peak conductance of $g_- = 2.1$ nS.

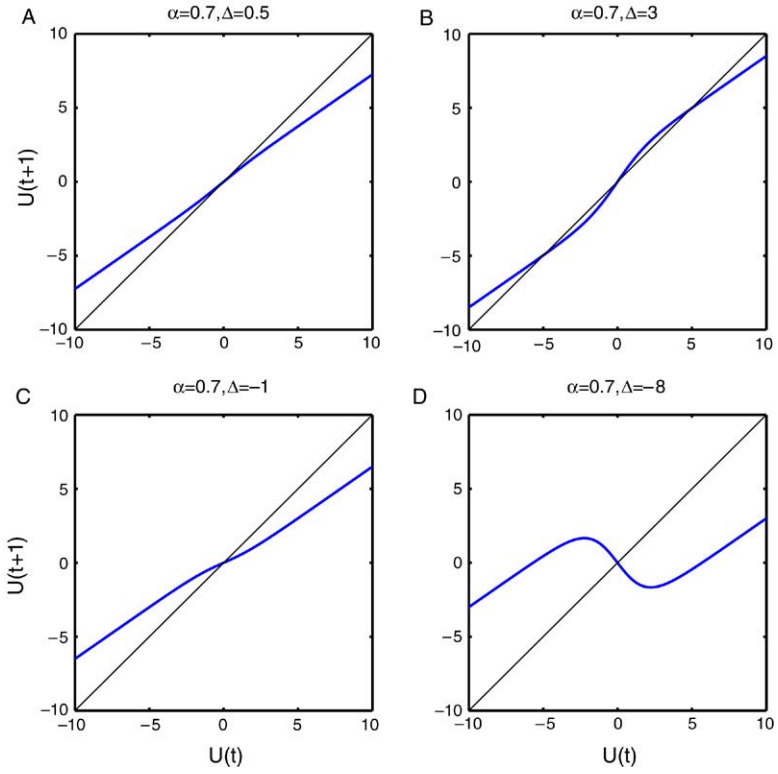

Fig. 4. Stability analysis of the reinforcement learning model in algorithm 0. A steady state is given by the intersection of the update rule (blue curve) and the identity line (block line). For a fixed value of α, as the absolute value of Δ becomes larger, choice behaviour at $P_R = 0.5$, or equivalently at $U(t) = 0$, becomes unstable. As shown in the top panels, if Δ is positive, as Δ increases the stable steady-state at $U(t) = 0$ (A) becomes unstable and two new stable steady-states emerge (B). Bottom panels (C and D) show the case in which Δ is negative. In this case a more negative value of Δ results in instability at $U(t) = 0$ (D). This instability results in alternation between the two targets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

At a given moment, for a set of synapses associated with a particular population i of excitatory neurons, a fraction of such synapses c_i are in the potentiated state, whereas the remaining fraction $1 - c_i$ are in the depressed state. This parameter c_i is called the 'synaptic strength' and it determines the overall input to neurons in selective population i. We assume that firing rates of the input neurons to both selective populations of excitatory neurons are similar, so the difference in the overall inputs to the two populations depends only on the states of their plastic synapses. In the present study, we do not model a read-out network for decision making explicitly. Instead, we assume that at the time when the difference between the average firing rates of the two selective populations exceeds a fixed threshold of 10 Hz (for an interval of at least 50 ms), the population with a higher firing rate determines the choice of the network.

4.2. Neural activity and choice behaviour of model network

Due to the winner-take-all dynamics, the model network is capable of making a binary decision, even when the strengths of plastic synapses are relatively similar for the two populations of excitatory neurons. This is demonstrated by the examples shown in Fig. 6, for the case when the synaptic strength for the population of neurons selective to the rightward target is slightly larger. Because neural spike discharges are intrinsically stochastic, the network's choice can change from trial to trial but it is more frequently biased towards the target with a stronger input. In this example, the right population wins the competition and determines the network's choice in approximately 55% of the trials.

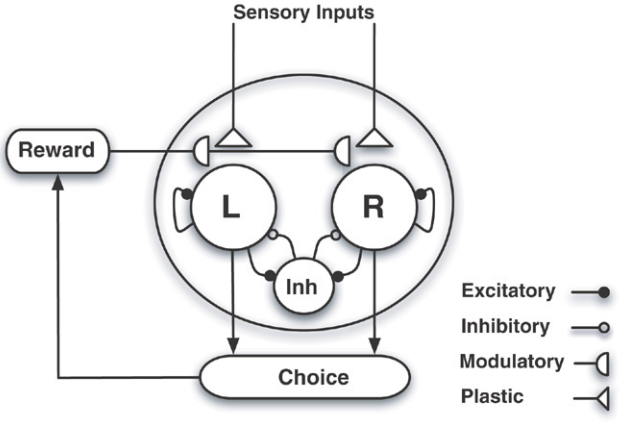

Fig. 5. Schematic model architecture. The core of the model consists of two populations of excitatory neurons which are selective to the two target stimuli and compete against each other through feedback inhibition. Upon the presentation of stimuli, neurons in the two selective populations receive similar inputs through plastic synapses. At the end of each trial these plastic synapses undergo a stochastic Hebbian learning rule which is gated by the all-or-none reward signal.

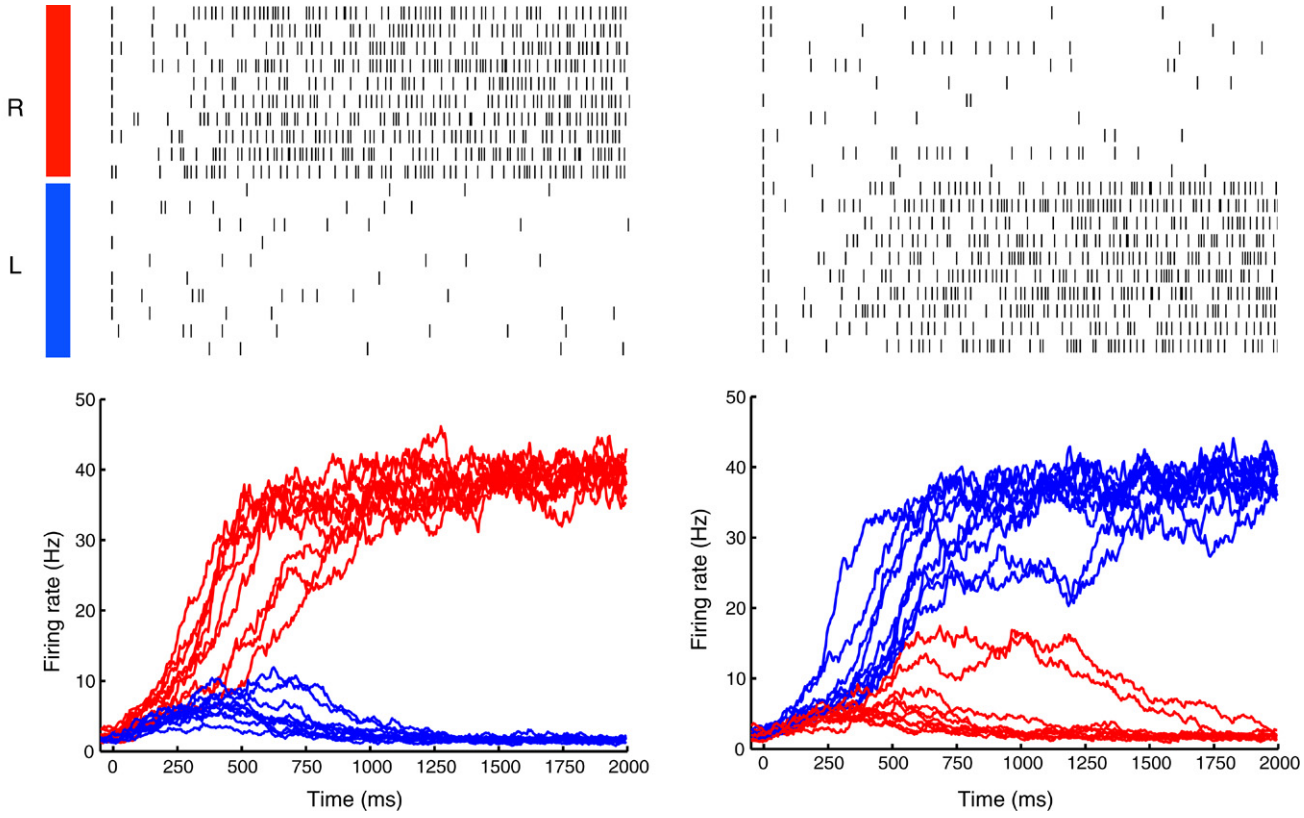

Fig. 6. Examples of neural activity in the decision-making network in 20 simulated trials. The left panels show the population activity of the neurons, and the spike trains of example neurons in the two selective populations in trials in which the right population (red traces) wins the competition. Similarly the right panels show the activity in trials in which the left population (blue traces) wins the competition. In these simulations the synaptic strength onto the right populations is set to $c_R = 52\%$, and synaptic strength onto the left populations is set to $c_L = 48\%$.

We further quantified the probabilistic choice behaviour of the network by computing the choice probability as a function of different synaptic strengths, c_R and c_L. We found that the choice probability is approximately only a function of the difference between the synaptic strengths. This is shown in Fig. 7, which shows, for three different overall synaptic strengths, the probability of choosing the rightward target as a function of the difference between the two synaptic strengths, $c_R - c_L$. The choice probability as a function of the difference in synaptic strengths can be fitted by a sigmoid (softmax) function.

$$P_R = \frac{1}{1 + \exp\left(-\frac{c_R - c_L}{\sigma}\right)} \tag{6}$$

where P_R is the probability of choosing the rightward target. This is interesting because many models in reinforcement learning (Sutton & Barto, 1998) and game theory (Camerer, 2003) assume such a decision criterion to map valuation to action, and our model represents a neuronal instantiation of it.

The value of σ in Eq. (6) indicates the randomness of the network's choice behaviour, that is a larger value of σ denotes a network with more random choice behaviour. In the model, the value of σ is determined by the structure of the network and the range of the actual difference in the overall currents passing through the plastic synapses for the two populations of excitatory neurons. For example, if the difference in the overall synaptic currents fluctuates within a small range, then σ would be large and the model behaves stochastically. The overall synaptic conductance (and similarly current) to each neuron is

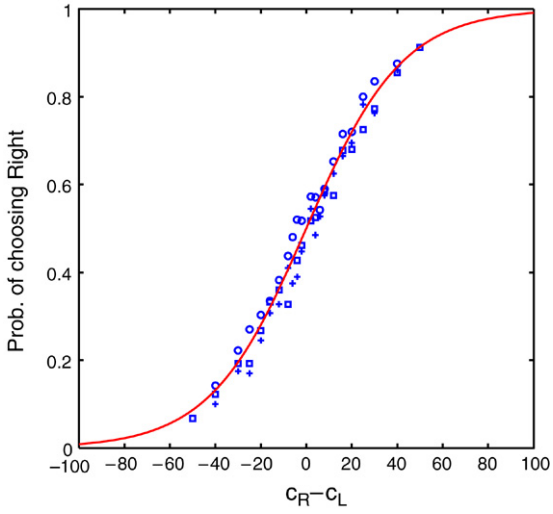

Fig. 7. Choice behaviour of the decision-making network as a function of the difference in synaptic strengths. The choice probability is extracted from the full network simulations (400 trials for each set of synaptic strengths). Different symbols represent different sets of synaptic strengths with different overall synaptic strength $c_R + c_L = 60\%$ (plus), 100% (square), 140% (circle). The red curve shows a sigmoid function fit to all data points (Eq. (6), $\sigma = 21\%$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a function of multiple factors, including the presynaptic firing rate, the peak conductance of the potentiated and depressed states, and the total number of plastic synapses and can be described by:

$$G = N_p f_{st}(cg_+ + (1-c)g_-)\tau_{\text{syn}} \quad (7)$$

where N_p is the number of plastic synapses onto each neuron, f_{st} is the firing rate of the presynaptic neurons, g_+ and g_- are the peak conductance of the synapses in the potentiated and depressed states respectively, and τ_{syn} is the time constant of the AMPA receptors. Thus the difference in the overall synaptic conductance can be written as:

$$G_R - G_L = (c_R - c_L)N_p f_{st}(g_+ - g_-)\tau_{\text{syn}}. \quad (8)$$

As a result, any of the factors in the right side of Eq. (8) can affect the difference in the overall synaptic currents to neurons in selective populations and so it can change the value of σ. For the parameters we used here the value of σ is equal to 21%, but its value can be increased or decreased if any of the factors mentioned above change. For example, the value of σ can be reduced if the firing rates of input neurons increase.

To avoid time-consuming network simulations for individual trials of the experiment, we use the extracted sigmoid function (Eq. (6)) to compute the choice probability for a given set of c_R and c_L values. We then used this choice probability to flip a biased coin and determine the choice of the network in that trial. At the end of each trial, the synaptic strengths are modified according to the outcome of that trial (rewarded or unrewarded) and the learning rule which is presented in the next section.

4.3. Stochastic learning rule gated by reward

Many different types of learning in a natural environment is driven by reward or punishment (Thorndike, 1911). Such information must be translated into an internal signal in the brain to affect the learning process. Accordingly, the regulation of synaptic plasticity by a hetero-synaptic modulatory signal related to reward or punishment is required for any learning rule that can be used to optimize the overall gain of reward. Indeed, dopamine may act as a common currency for such a reward signal in the brain (Schultz, 2000, 2006) and modulate synaptic plasticity (Jay, 2003).

The modulatory effect of dopamine has been studied in different brain areas including the striatum, the hippocampus and the prefrontal cortex (Jay, 2003; Otani, Daniel, Roisin, & Crepel, 2003; Reynolds & Wickens, 2002). In particular, afferents from the cerebral cortex and dopamine inputs arising in the substantia nigra pars compacta converge in the striatum, and corticostriatal synapses undergo plasticity according to the activity of dopamine neurons (Reynolds, Hyland, & Wickens, 2001; Reynolds & Wickens, 2002). Furthermore, in the rat prefrontal cortex, the induction and direction of long-term depression (LTD) and long-term potentiation (LTP) is modulated by dopamine (Huang, Simpson, Kellendonk, & Kandel, 2004; Otani et al., 2003).

In our model, synaptic plasticity is Hebbian and gated by a reward signal. The Hebbian component implies that plasticity is dependent on correlation between pre-synaptic and post-synaptic activity. In addition, the absence of the reward signal can reverse the direction of plasticity. This reward signal is assumed to be binary, indicating whether a reward has been harvested or not. In our simulation, the pre-synaptic side of plastic synapses is always active, because stimuli are presented throughout the trial, so the direction and magnitude of synaptic plasticity is determined entirely by the postsynaptic activity and reward signal.

We also assume that synaptic plasticity is stochastic. In other words, when the condition for plasticity is met, plastic synapses undergo changes with some probability (Amit & Fusi, 1994; Fusi, 2002; Fusi, Drew, & Abbott, 2005). Specifically, for the condition that requires synaptic potentiation, depressed synapses are potentiated with probability q_+. Similarly, for the condition that requires synaptic depression, the potentiated synapses are depressed with probability q_-. As we mentioned before, the impact of the sensory input on each population of excitatory neurons is determined by the synaptic strengths, c_L and c_R, each defined as the fraction of synapses in the potentiated state in a given selective population. Updating rule for these synaptic strengths can therefore be written as:

$$c_i(t+1) = c_i(t) + q_+(r;\nu_i)(1-c_i(t)) - q_-(r;\nu_i)c_i(t) \quad (9)$$

where i = R or L; $q_+(r;\nu_i)$ and $q_-(r;\nu_i)$ are the potentiation and depression rates, respectively. The second term describes the change due to the transition of depressed synapses (fraction $1-c_i$ of synapses are potentiated with probability $q_+(r;\nu_i)$) and the third term describes the change due to the transition of potentiated synapses (fraction c_i of synapses are depressed with probability $q_-(r;\nu_i)$). The learning parameters, $q_+(r;\nu_i)$ and $q_-(r;\nu_i)$, depend on the firing rates of a postsynaptic neuron at the end of a trial ν_i, and the outcome r (rewarded or unrewarded). As we showed in the last section, the firing rate, ν_i, of neurons selective to the chosen (unchosen) target is high (low) at the end of a trial, so we assume only two possible states for the firing rate of postsynaptic neurons, high and low. As a result, each of the four possible outcomes of the decision making (whether the firing rate of the neurons is high or low, and whether a reward is delivered or not) requires two separate learning rates. Thus, in theory, many different learning rules are possible. In the following, we consider two important examples of this learning rule and their connections to the reinforcement learning model (Lee et al., 2004).

4.4. Choice-specific learning rule

The above learning rule can be simplified if one assumes that synapses projecting to inactive neurons are not modified. This means that in each trial, only the synapses projecting to the population of excitatory neurons corresponding to the chosen target are modified. Accordingly, we refer to this rule as a 'choice-specific' learning rule. We also assume that in each trial only one form of plasticity (i.e. potentiation or depression) occurs and the direction of plasticity is determined by the presence or absence of reward. In other words, if the choice in a given trial is rewarded, then synapses are potentiated (i.e. LTP),

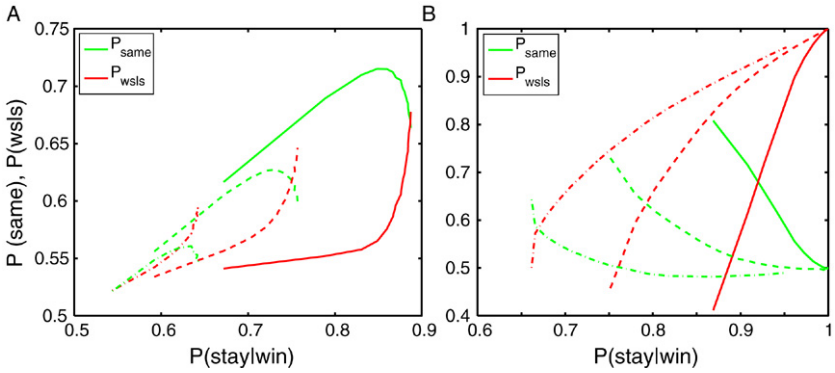

Fig. 8. Performance of the model with different learning rules in algorithm 1. (A) In the model with choice-specific learning rule, the probability of choosing the same target in two consecutive trials, P_{same}, mostly increases as the probability of WSLS strategy, P_{wsls}, increases. In addition there is a limit for P_{wsls} in this model. (B) In the model with belief-dependent learning rule, P_{same} decreases as the P_{wsls} increase, and P_{wsls} can reach to value close to 1. If σ value is large, P_{wsls} can vary over a large range while P_{same} is close to 0.5, consistent with the monkeys' choice behaviour. For these simulations q_+ (or q_r) is fixed at 0.1 while q_- (or q_n) is varied in the range of [0.025, 0.825]. The value of σ is set to 5% (solid), 10% (dash), and 20% (dot–dash).

and if the choice is not rewarded, they are depressed (LTD). This reduces the number of learning parameters to two and the updating rule can be written as:

Right is selected and rewarded:

$$\begin{cases} c_R(t+1) = c_R(t) + (1 - c_R(t))q_+ \\ c_L(t+1) = c_L(t). \end{cases} \tag{10}$$

Right is selected but not rewarded:

$$\begin{cases} c_R(t+1) = c_R(t) - c_R(t)q_- \\ c_L(t+1) = c_L(t). \end{cases}$$

The learning rule for trials in which the leftward target is selected can be obtained by switching the L and R indices. Note that in the framework of reinforcement learning models, this learning rule is equivalent to the state-less Q-learning (Sutton & Barto, 1998).

Although this choice-specific learning rule is based on a plausible mechanism for synaptic plasticity, there is an important discrepancy between the behaviour of the model driven by this learning rule and the animal's behaviour during algorithm 1 of the matching pennies game. According to this learning rule, the model is more likely to choose the same target after a rewarded trial (i.e. win-stay). However, the probability that the model would switch to the other target after an unrewarded trial (i.e. lose-switch) is smaller than the probably of win-stay. This is because at the time of choice, the strength of plastic synapses for the chosen target, c_i would be on average larger than the synaptic strength for the unchosen target. If the choice is rewarded, c_i for the chosen target would increase, and consequently the probability of choosing the same target in the next target would increase further. In contrast, if the choice is not rewarded, then c_i for that target would decrease, but it may still remain larger than the strength of plastic synapses for the unchosen target. Accordingly, the overall probability that the model selects the same target in the two successive trials would be larger than 0.5. In fact, for the choice-specific learning rule, the overall probability of choosing the same target in two consecutive trials (P_{same}) increases with the probability of WSLS strategy (Fig. 8A).

In contrast, the monkeys tested in the matching pennies task displayed approximately the same amount of win-stay and lose-switch behaviours. Moreover, the probabilities of win-stay and lose-switch were both high during algorithm 1, and yet P_{same} was quite close to 0.5. We therefore conclude that the choice-specific learning rule cannot account for the choice behaviour of animals during algorithm 1.

4.5. Belief-dependent learning rule

The choice-specific learning rule described above utilizes only two model parameters, q_+ and q_-, and allow only the synapses for the population of neurons related to the chosen target to be modified. Alternatively, the general learning rule described by Eq. (9) can be simplified by assuming that synapses for the population of neurons related to the unchosen target, and therefore, synapses leading to the inactive neurons, are also modified with the same probability as the synapses of the population for the chosen target, but in the opposite direction. In game theory, the type of learning rules which modify indiscriminately the value functions for chosen and unchosen actions are known as belief learning (Camerer, 2003; Lee, McGreevy, & Barraclough, 2005). Thus, we refer to our second example of update rule as 'belief-dependent learning rule', since this modifies the plastic synapses projecting to both choices at the end of each trial. By assuming that the learning rates in a given trial are similar for both sets of synapses, but differ for rewarded and unrewarded trials, we obtain the following learning rule:

Right is selected and rewarded:

$$\begin{cases} c_R(t+1) = c_R(t) + (1 - c_R(t))q_r \\ c_L(t+1) = c_L(t) - c_L(t)q_r. \end{cases} \tag{11}$$

Right is selected but not rewarded:

$$\begin{cases} c_R(t+1) = c_R(t) - c_R(t)q_n \\ c_L(t+1) = c_L(t) + (1 - c_L(t))q_n \end{cases}$$

where q_r and q_n are the learning rates in the rewarded and unrewarded trials, respectively.

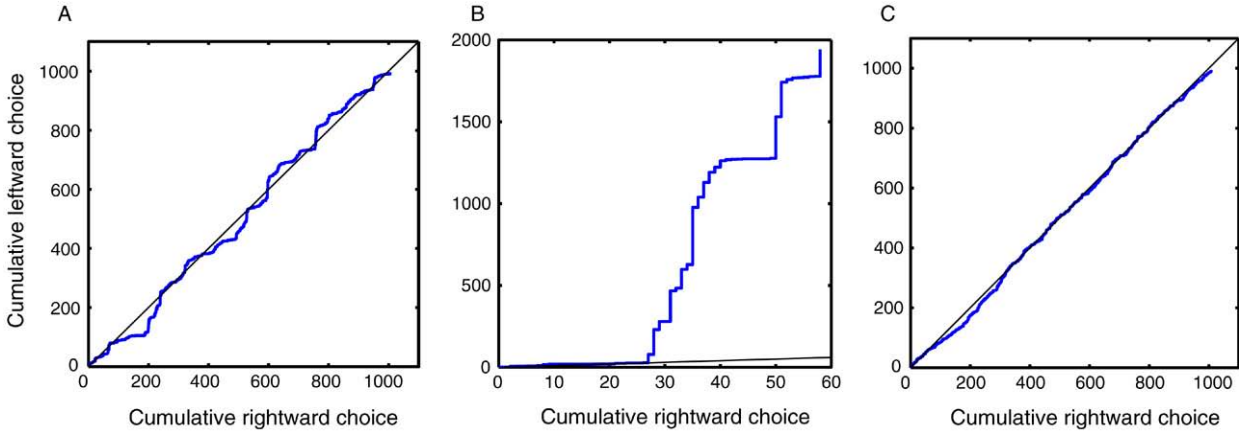

Fig. 9. Examples of different model's choice behaviour in algorithm 0. The model shows different choice behaviours depending on the learning parameters (for fixed $\sigma = 10\%$). (A) For model parameters of $q_r = 0.035$, $q_n = 0.03$, the condition for a stable steady-state at $P_R = 0.5$ is met so the two targets are chosen with equal probability. (B) For $q_r = 0.09$, $q_n = 0.03$ the condition for the stable steady-state at $P_R = 0.5$ is not fulfilled and two new stable steady-states emerge. As a result, the model randomly shows a strong bias for one of the choices (in this example for the leftward choice). (C) If $q_r = 0.1$, $q_n = 0.7$ the only steady state at $P_R = 0.5$ is unstable and the model mostly alternates between the two choices. The black line shows the unity line.

We now show that this learning rule is closely related to the reinforcement learning model used in Lee et al. (2004). First, we define a new variable c which is equal to the difference between c_R and c_L. The update rule for c in rewarded trials can then be written as:

$$c(t+1) = c(t)(1 - q_r) \pm q_r \tag{12}$$

where the plus (minus) sign applies to the trials in which the rightward (leftward) target is chosen. Similarly, for unrewarded trials, the update rule is

$$c(t+1) = c(t)(1 - q_n) \mp q_n \tag{13}$$

where the minus (plus) sign applies to the trials in which the rightward (leftward) target is chosen. Comparing the last two equations with the reinforcement learning model in Eq. (3) shows that this learning rule is equivalent to the learning rule in a reinforcement learning model with two decay factors $(1 - q_r)$ and $(1 - q_n)$ for the rewarded and unrewarded trials, respectively. Furthermore the equivalent changes in the value function, Δ_1 and Δ_2, are equal to q_r and $-q_n$ respectively, but these values are dependent on the decay factors.

Due to these similarities, choice behaviour of the model with the belief-dependent learning rule is quite similar to that of the reinforcement learning model described in Eq. (3). More importantly, the abstract parameters in the reinforcement learning model are grounded to the learning rates at the synaptic level. This implies that value functions of alternative actions can be stored in plastic synapses, and that the decay factor in the reinforcement learning model is equivalent to the probability that plastic synapses are not modified. The dependence of the decay factor on the amount of change in each trial results directly from the fact that the number of available synapses (and the number of possible synaptic states) is limited. This biophysical constraint causes the representation of value function to be limited. At the same time, however, the amount of noise in the decision-making network, measured by σ, can amplify the limited difference between the synaptic strengths and influence the randomness of choice behaviour.

5. Model's behaviour in the game of matching pennies

As described in the previous section, our model with the belief-dependent learning rule at the synaptic level behaves similarly compared to the reinforcement learning model, and the synaptic strength of this model can represent the value function for each choice. In this section, the choice behaviour of our model is characterized further, focusing on the behaviour during algorithm 0 and the robustness of the model.

5.1. Stability of model's behaviour in algorithm 0

In Section 3, we examined the conditions in which the reinforcement learning model shows an unstable choice behaviour during algorithm 0. In this section, we identify conditions in which our network model shows similar behaviour, and illustrate the unstable choice behaviour of the model. Based on the calculations in Section 3 and the comparison between the network model and the reinforcement learning model presented in Section 4.5, one can show that the steady-state value for c ($\equiv c_R - c_L$) is given by

$$c_{ss} = \frac{2(q_r - q_n)(P_R - 0.5)}{q_r + q_n}. \tag{14}$$

Note that in our network model, the choice probability is affected by the value of σ ($P_R = \frac{1}{1+\exp(-\frac{c}{\sigma})}$), so the dynamics of the choice behaviour in algorithm 0 falls into three regimes. If $(q_r - q_n)$ is positive, Eq. (14) always has one solution at $P_R = 0.5$ and this solution is a stable steady-state if $\frac{\sigma(q_r+q_n)}{2(q_r-q_n)} \geq 0.25$. An example of this choice behaviour is shown in Fig. 9A. Although model chooses the two targets with equal probability, it may locally show some preference for one of the targets. If instead $\frac{\sigma(q_r+q_n)}{2(q_r-q_n)} < 0.25$ then $P_R = 0.5$ solution becomes an unstable steady-state and instead two other stable steady-states emerge. As shown in Fig. 9B, the model then shows a clear bias toward one of the targets (which can be any of the two targets). This means that if the difference between the learning rates in

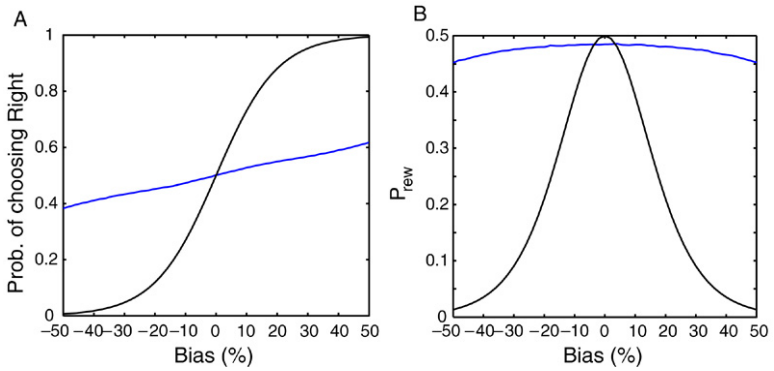

Fig. 10. Choice probability and performance of the model with an intrinsic bias. (A) The black curve shows the probability of choosing the rightward target for a given intrinsic bias, when the plastic synapses are not modified (or there is no feedback). The blue curve shows the probability of choosing the rightward target for the same model which plays against the computer in algorithm 1 and plastic synapses are modified. The bias in the model is drastically reduced, due to feedback and learning dynamics. (B) Performance of the model with an intrinsic bias. The probability of obtaining reward is plotted for different intrinsic biases while the model plays against the computer opponent in algorithm 1 (blue curve). The black curve shows the harvesting rate if the synapses are not modified and only the intrinsic bias determines the choice probability. For each value of the bias (from −50 to 50 with intervals of 1) the average in each condition is computed over 400 days (each day consists of 1000 ± 200 trials) of the experiment and the model parameters are set to $q_r = 0.1$, $q_n = 0.2$, and $\sigma = 10\%$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the rewarded and unrewarded trials is positive and large, the choice behaviour can be biased towards one of the targets. This regime may correspond to monkeys' choice behaviour in the Barraclough et al. (2004) experiment during algorithm 0.

If the learning rates in unrewarded trials is greater than those in rewarded trials ($q_r - q_n \leq 0$), there is only one steady-state at $P_R = 0.5$ for Eq. (14). This steady state is stable if $\frac{\sigma(2-q_r-q_n)}{2(q_n-q_r)} \geq 0.25$ and unstable otherwise. This instability at $P_R = 0.5$ results in the alternation between the two targets (Fig. 9C), and is therefore qualitatively different from the instability resulting when $q_r - q_n \geq 0$. Because the choice behaviour is binary and stochastic, a stable steady-state biased toward one of the targets may not be easily distinguishable from a stable choice behaviour at $P_R = 0.5$. If the choice is slightly biased, only the average choice probability would depart from $P_R = 0.5$ over a long sequence of trials. Overall, these results demonstrate that our model can produce an intrinsically probabilistic choice behaviour. In addition, a biased and non-random choice behaviour can also emerge from the same model. This biased choice behaviour does not necessarily reflect the insensitivity of the model to the reinforcement or feedback, but instead may result from the learning mechanism of the model.

5.2. Intrinsic bias and model robustness

Stability analysis described in the previous section was based on the assumption that the choice behaviour is not intrinsically biased toward one of the choices. However, the decision-making network may have an intrinsic bias, if there is asymmetry in the constant inputs to different populations. If so, one of the targets would be preferred over the other target, and the fixed point or steady state of choice behaviour at $P_R = 0.5$ will be shifted to another point. For the animal's choice behaviour in algorithm 0, it is not possible to distinguish such an intrinsic bias from a bias resulting from the dynamics of learning. If the bias is due to the dynamics of learning mechanism, one can expect that it would resolve once the computer begins to punish such a biased choice behaviour. The question remains as to whether an intrinsic bias in the network can be remedied by the presence of plastic synapses which undergo our learning rule. In the following, we consider this issue for the animal's choice behaviour in algorithm 1 that penalizes a biased choice behaviour and rewards a random sequence of choices. On the one hand, a stochastic behaviour in the reinforcement learning model requires a relatively slow rate of learning, because a large learning rate produces a predictable behaviour, such as win-stay or lose-switch. On the other hand, the slow rate of learning may interfere with the network's ability to reach the unbiased choice behaviour. Here, we try to address this question by studying our model choice behaviour in algorithm 1 (similar results can be obtained in algorithm 2). The intrinsic bias was implemented by adding a constant term to the c value, which is equivalent to a constant current injected to one of the selective populations in the decision-making network.

If the model does not receive any feedback (or similarly if plastic synapses are not modified), the choice selection is biased toward the choice which receives an additional input. This is shown in Fig. 10. If the model receives feedback and synapses are modified according to our belief-dependent update rule, the intrinsic bias is compensated by synaptic changes and the bias in the choice behaviour is dramatically reduced (blue curve in Fig. 10A). This compensation leads to a high overall reward rate (Fig. 10B). This compensation takes place because the plastic synapses related to the population with additional input tend to be depressed as the same population frequently wins the competition and determines the target choice without reward. As a result, the average value of synaptic strength projecting to this population would be lower than the synaptic strength related to the other population. This is illustrated in Fig. 11.

5.3. Comparison with the animal's behaviour

Previously, Lee et al. (2004) fit the choice behaviour of each animal in different algorithms with a reinforcement learning

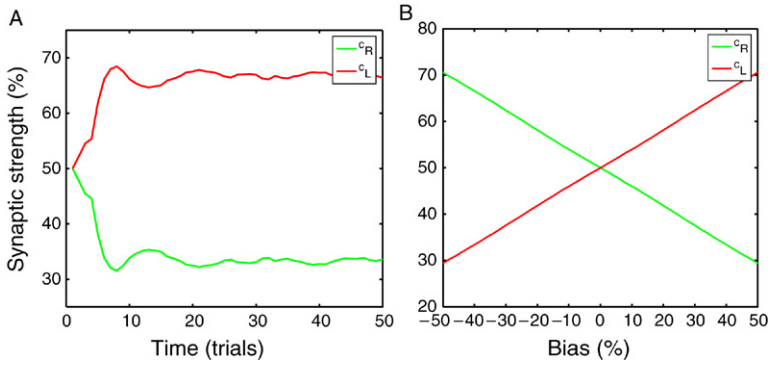

Fig. 11. Intrinsic bias can be compensated by plastic synapses. (A) Time course of the average synaptic strengths, c_R and c_L, in algorithm 1. Within about 10 trials the difference between the two synaptic strengths increases to compensate for the intrinsic bias. In this simulation the rightward choice receives an additional constant input which is equivalent to 40% difference in synaptic strengths. The average is computed over 1000 sessions. (B) The average synaptic strengths for different values of intrinsic bias. As the intrinsic bias increases the difference in synaptic strengths also increases. The averages are computed over 1000 sessions for each intrinsic bias. The model parameters are set to $q_r = 0.1$, $q_n = 0.2$, and $\sigma = 10\%$.

model. Although this model captured some key features of the animal's behaviour, the animal's choice behaviour was not stationary throughout the course of the experiment. Instead, the behaviour changed during and across different algorithms, as shown in Section 2.2. Here, we used our model with belief-dependent learning rule to fit the animal's choice behaviour for each day of the experiment separately. In this analysis, the value of σ value was fixed at 50%, since this made it easier to compare the learning rates across multiple days.

The resultant maximum likelihood estimates (Burnham & Anderson, 2002) of the learning rates are plotted for each day of experiment in Fig. 12. These estimates give some insight into the choice behaviour in different algorithms. For example, for monkey E in algorithm 0, q_r is substantially larger than q_n, which is in line with the results in Section 5.1 regarding the instability of the choice behaviour around $P_R = 0.5$. Furthermore, the value of learning rates changes from day to day, especially during algorithm 1 in monkey E. In this case, as the animal increased the use of WSLS strategy gradually, the learning rates also increased. When the algorithm 2 was introduced, the learning rates decreased, suggesting that random choice behaviour might result from adaptation of the learning rates to small values (slow learning). These results indicate that in order to provide a full account of the behaviour, an additional mechanism is required to adjust the learning rates. In the next section, we consider how this might be accomplished through an algorithm that modifies the learning rate in order to maximize the reward rate.

6. Meta-learning

The behavioural data and the estimates of learning rates described above suggest that in addition to the trial-to-trial dynamics of the choice behaviour, there is a much slower change in the behaviour which takes place across multiple days during the course of the experiment. This slow change was most noticeable, when the computer opponent switched to algorithm 2 (see Figs. 3 and 12). During this experiment, animals were not explicitly cued for the transitions in the algorithms used by the computer. Nevertheless, after algorithm 2 was introduced, they all experienced a transient reduction in the reward rate. Therefore, change in their choice behaviour might result from steps taken to restore the previous level of reward rate. In order to check this hypothesis we implemented a modified version of the meta-learning algorithm proposed by Schweighofer and Doya (2003), which is an extension of the stochastic real value units algorithm (Gullapalli, 1990).

6.1. Meta-learning algorithm

The goal of a meta-learning model is to maximize the long-term average of rewards, by using stochastic units and comparison between the medium-term and long-term running averages of the reward rate, denoted by $\overline{r(t)}$ and $\overline{\overline{r(t)}}$, respectively. Our model has three parameters which can be adjusted by this meta-learning algorithm (q_r, q_n, and σ). It is possible that signals related to these running average reward rate are encoded by the firing of neurons in specific brain areas. However, in our simulation, we do not explicitly model such neurons, since little is known about the candidate neural mechanism for such signals. Instead, we simply assume that learning rates of the plastic synapses, q_r and q_n, are controlled by the activity in two different sets of neurons. Similarly, we assume that the sensitivity of the network, σ, is controlled by another set of neurons. With these assumptions, we examine the effect of meta-learning on the model parameters directly.

During meta-learning, all model parameters (say Λ) are perturbed after every n trials ($n \gg 1$) as follows.

$$\Lambda'(t) = \Lambda_b + \delta_\lambda(t) \tag{15}$$

where Λ_b is the mean value of the parameter Λ, and δ_λ is the amount of perturbation which is drawn from a Gaussian distribution with a zero mean and variance ϵ_λ. These new parameters are then used to generate the choice behaviour. After a few hundred trials, the activity of modulating neurons are modified according to the difference between the mid-term and long-term averages of reward rate and as a result the mean value of model parameters (say Λ) are updated according to:

$$\Lambda'_b = \Lambda_b + \nu_\lambda(\overline{r(t)} - \overline{\overline{r(t)}})\delta_\lambda(t) \tag{16}$$

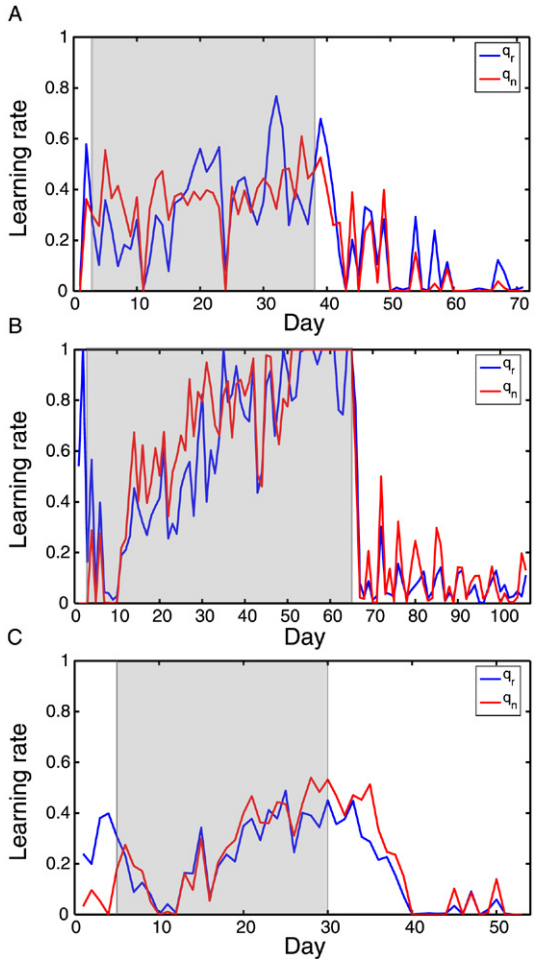

Fig. 12. Maximum likelihood estimate of the model parameters. These parameters are obtained from fitting the choice behaviour of three monkeys in each day of the experiment; (A) monkey C, (B) monkey E, (C) monkey F. The gradual change in the learning parameters during the experiment is another indication that monkeys changed their strategies continuously. Consistent with the results obtained in Section 5.1, q_r is larger than q_n in algorithm 0 which explains the observed unstable choice behaviour around $P_R = 0.5$. During algorithm 1, in all monkeys both learning rates increase which result in increase in the use of WSLS strategy. During algorithm 2 the learning rates decrease which shows that the only possible way to play randomly is to have slow learning. For these fittings, the value of σ is fixed at 50%.

where ν_λ is a meta-learning rate, and $\overline{r(t)}$ and $\overline{\overline{r(t)}}$ are the mid-term and long-term running averages of reward rate, respectively. The mid-term and long-term running average of reward rate are computed according to the following update rule.

$$\Delta\overline{r(t)} = \frac{1}{\tau_1}(-\overline{r(t)} + r(t))$$
$$\Delta\overline{\overline{r(t)}} = \frac{1}{\tau_2}(-\overline{\overline{r(t)}} + \overline{r(t)})$$

where τ_1 and τ_2 are the time constants for averaging the past rewards. Based on this algorithm if after a perturbation, $\overline{r(t)}$ becomes larger (smaller) than $\overline{\overline{r(t)}}$ (which means the perturbation was good (bad)) the mean learning parameter will be changed in the direction (in opposite direction) of the perturbation. In the next subsection, we show how this model

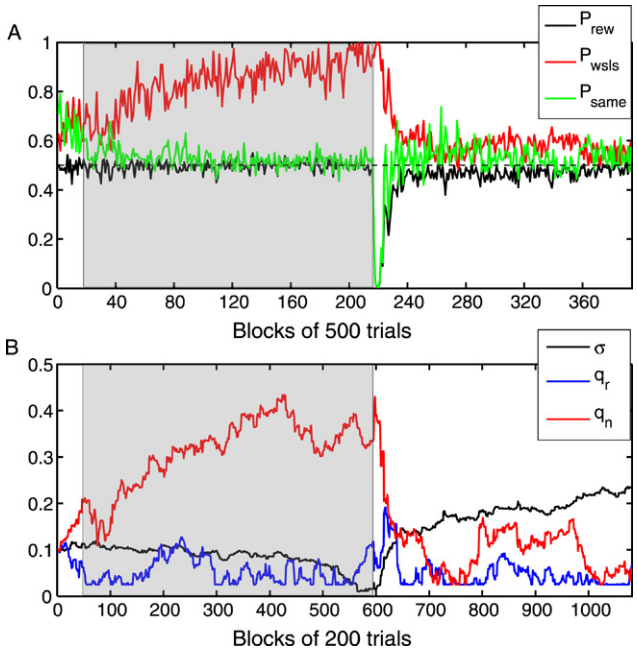

Fig. 13. An example of the model's average choice behaviour in 200 days of the experiment. When the meta-learning is active the model's choice behaviour is adjusted according to the algorithm used by the computer opponent. (A) Time courses of different measures of the model's choice behaviour (average over blocks of 500 trials). Blocks during algorithm 1 are shaded. (B) The model parameters are adjusted in each 200 trials according to meta-learning algorithm. The initial values for the model parameters are $q_r = q_n = 0.1$ and $\sigma = 10\%$ and meta- learning parameters used for updating the learning rates (q_r and q_n) are $\nu_q = 2$, and $\epsilon_q = 0.002$, and for updating the noise level σ, $\nu_s = 5$, $\epsilon_s = 0.005$. The time constants for averaging reward are set to $\tau_1 = 100$ and $\tau_2 = 400$ trials.

can qualitatively replicate the gradual changes in the animal's choice behaviour over the course of the experiment.

6.2. Choice behaviour with meta-learning

We simulated the behaviour of the model with the above meta-learning algorithm during the matching pennies game against the computer opponent for 200 days of experiment (9 days in algorithm 0, 100 days in algorithm 1 and 91 days in algorithm 2). Each day of experiment consisted of 1000 ± 200 trials and the model parameters were modified according to the meta-learning algorithm every 200 trials. We assumed that learning rates are not very small, so the minimum value for both learning rates (q_r and q_n) was set to 0.025. In addition, the maximum value for σ is set to 25%. These limits set the minimum change in the choice probability after any trial to be 2.5%.

An example of the model's choice behaviour over the course of the experiment is shown in Fig. 13. In Fig. 13A, averages of three quantities related to model's behaviour are plotted (P_{wsls}, P_{same}, P_{rew}). This plot shows that following the introduction of algorithm 2, the model obtained the maximum reward rate by exploring different parameters. Notice that sometimes fluctuations in the reward rate caused changes in the model parameters without influencing the reward rate (e.g. in algorithm 0). This example is qualitatively similar to the choice

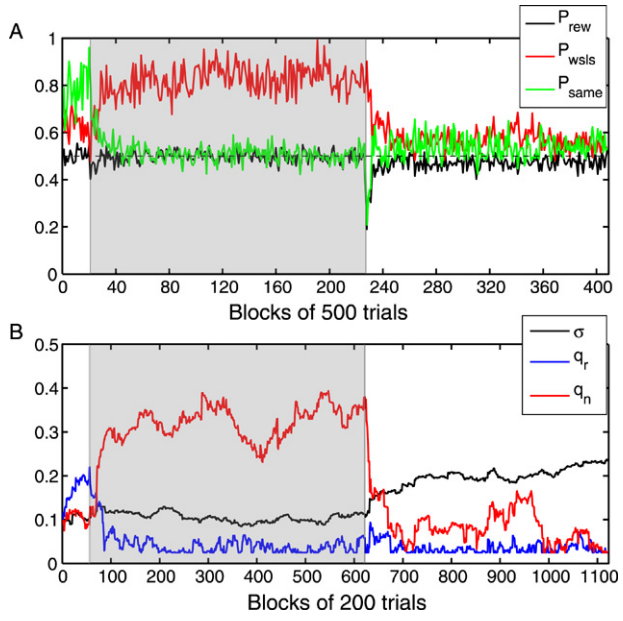

Fig. 14. Another example of the model's average choice behaviour in 200 days of the experiment. The model parameters are similar to those used in Fig. 13.

behaviour of monkey E, in that there was a gradual increase in the probability of WSLS strategy during algorithm 1, which diminished soon after the introduction of algorithm 2. During this simulated experiment, the model parameters changed from block to block, while maintaining the reward probability close to the maximum level (Fig. 13B). The learning rate for rewarded trials, q_r, was small and fluctuated around a relatively small value, whereas the learning rate for the unrewarded trials, q_n, increased during the course of algorithm 1. A value of q_n larger than q_r results in more use of WSLS strategy which can be seen from Fig. 13A. An interesting finding from this simulation is that during algorithm 1 the noise level does not need to be large, in order to obtain the maximum overall reward. This strategy is not viable during algorithm 2, as reflected in a transient reduction in reward rate following the introduction of algorithm 2. Consequently, q_n decreased quickly while σ increased.

In most of the results obtained without meta-learning algorithm, we fixed the value of σ which determines the noise level in the decision making process. A large value for σ gives rise to a random choice behaviour, but the behaviour resulting from a large σ is different from the experimental observation (e.g. P_{wsls} in all monkeys is significantly larger than 0.5). For this reason, we set a limit for σ value in order to prevent the model from adopting a trivial solution. For the model with meta-learning algorithm, it was important that both learning rates decreased and level of noise (σ) increased in order to obtain the maximum rate of reward during algorithm 2. In addition, the model with meta-learning produced a diverse pattern of results. If we rerun the same simulation with similar initial condition, we can observe other possible patterns of choice behaviour. The second example, shown in Fig. 14, is more similar to the choice behaviour of monkey C. Overall, these results show different possibilities for generating random choice behaviour. Not only the initial condition and meta-learning parameters, but also the probabilistic nature of this task can shape the time course of the choice behaviour. Therefore, the stochastic nature of meta-learning rule can be an underlying mechanism for generating a diverse repertoire of choice behaviour observed in various competitive games.

7. Discussion

One of the most important and influential models of decision making is reinforcement learning (Sutton & Barto, 1998). In this framework, desirability of each action is represented by a value function that estimates the expected amount of reward resulting from a particular action. Consequently, actions with high value functions are chosen more frequently. The outcome of each action is then compared to the previously expected outcome, and the resulting error is used to update value functions appropriately. Although the reinforcement model is plausible, its applicability to the brain remains to be firmly established in neurobiology. In particular, the detailed network and cellular mechanisms of reinforcement learning are still poorly understood and are currently the topic of active research (Lee, 2006; Sugrue, Corrado, & Newsome, 2005). In this paper, we proposed a biophysical implementation of a reinforcement learning algorithm based on reward-dependent stochastic Hebbian synaptic plasticity. Combined with a probabilistic decision-making network described previously (Wang, 2002), our model successfully accounted for several important features of the choice behaviour displayed by monkeys during a competitive game (for application of this model to another task see Soltani and Wang (2006)). The choice behaviour of our network model is determined by a softmax (i.e. logistic) function of the difference in the synaptic strengths, and accordingly the stochastic choice behaviour of our model results from ongoing fluctuation in neuronal activity and attractor dynamics of the decision-making network. Importantly, the plastic synapses in this model can temporally integrate information about the past rewards, such that the overall strengths of these synapses store information about value functions. This representation is bounded because the synaptic strengths can only take values between zero and one. Nevertheless, this bounded representation does not necessarily limit the ability of the model to generate a desired pattern of behavioural choices, such as a deterministic sequence of actions for problem solving or unbiased random behaviour during a competitive game. For example, to generate a deterministic behaviour with a relatively small synaptic strength, the input firing rates of presynaptic neurons can be increased, so that the difference in the overall currents injected to different neuronal populations in the network is large enough to introduce a strong bias in the choice behaviour of the network. In contrast, if the model is required to generate highly stochastic choice behaviour, as in the matching pennies game, the plastic synapses can be modified according to a specific learning rule to restore the network back to the probabilistic regime.

In our model, it is assumed that the plastic synapses determine the strengths of inputs to the neurons in the decision-making network. These plastic synapses must be modified

according to the chosen action and its outcome. The anatomical location of such plastic synapses, however, is not known. The decision-making network in our model was originally proposed as a simplified model for the lateral intraparietal cortex (LIP, a cortical area critical to controlling oculomotor behaviour (Wang, 2002)).

Although almost all areas of the primate cortex, including LIP, receive dopaminergic innervation (Lewis et al., 2001) and therefore in principle have access to the reward signals, it is not known whether plastic synapses modelled in our network are indeed localized in the LIP. It also has been proposed that basal ganglia provides a candidate circuit for action selection (Houk, Davis, & Beiser, 1995; Redgrave, Prescott, & Gurney, 1999; Reynolds & Wickens, 2002). In contrast to our model in which action selection is performed by competition through feedback inhibition, action selection in basal ganglia has been assumed to happen through multiple inhibitory pathways (Berns & Sejnowski, 1998) although the underlying neural mechanism for such action selection has not been fully understood. A recent study has shown that some neurons in the striatum indeed encode information about the value functions of different actions (Samejima, Ueda, Doya, & Kimura, 2005). It is also possible that the type of plastic synapses utilized in our model is widespread in a broad network of cortical and subcortical areas, including multiple regions of the prefrontal cortex.

We have showed that the model captures behavioural data from the matching pennies task of Barraclough et al. (2004), using a 'belief-dependent learning rule', which updates synapses onto neurons selective for both chosen and unchosen targets, in contrast to a 'choice-specific learning rule', according to which only those synapses onto neurons selective for the chosen target are modified. However, the belief-dependent learning rule we have used in the present study assumes that, in an unrewarded trial, synapses onto inactive neurons (those selective for the unchosen target) undergo potentiation, which may not be biologically plausible. In future studies it would be worth exploring variants of belief-dependent learning rules without this feature. Other kinds of mechanisms, perhaps involving a large network of multiple brain areas, are also conceivable.

A typical reinforcement learning algorithm only explains the change in choice behaviour that takes places according to the outcome of individual actions. In contrast, the choice behaviour of monkeys during the matching pennies displayed slow, gradual changes over a period of many days. Our model provides a biophysically plausible account of such behavioural changes by combining the learning rule operating on a trial-by-trial basis with an additional meta-learning algorithm. In reinforcement learning, meta-learning algorithms are commonly evoked to adjust parameters, such as the learning rate for updating value functions and the rate of temporal discounting for delayed rewards (Doya, 2002; Schweighofer & Doya, 2003). In contrast, a meta-learning algorithm in our model was used to set the learning rates for plastic synapses and the sensitivity of the network that controlled its randomness. We found that the model with a meta-learning algorithm reproduced several interesting features of the animal's choice behaviour during the matching pennies game, such as a gradual increase in the win-stay-lose-switch strategy against a partially exploitive opponent. These results suggest that even in a relatively simple dynamic decision-making task, such as matching pennies, animals continuously attempted to optimize their rate of reward on multiple timescales.

Acknowledgments

We are grateful to Dominic Barraclough, Michelle Conroy and Ben McGreevy for their help with the experiment. This study was supported by a grant MH073246 from the National Institute of Health.

References

Amit, D. J., & Fusi, S. (1994). Dynamic learning in neural networks with material synapses. *Neural Computation*, *6*, 957–982.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*, 428–454.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*, 404–410.

Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, *10*, 108–121.

Brunel, N., & Wang, X. -J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience*, *11*, 63–85.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

Camerer, C. F. (2003). *Behavioural game theory: Experiments in strategic interaction*. Princeton: Princeton Univ. Press.

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.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495–506.

Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biological Cybernetics*, *87*, 459–470.

Fusi, S., Drew, P. J., & Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, *45*, 599–611.

Gullapalli, V. (1990). A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks*, *3*, 671–692.

Herrnstein, R. J., Rachlin, H., & Laibson, D. I. (1997). *The matching law: Papers in psychology and economics*. Cambridge: Harvard Univ. Press.

Houk, J. C., Davis, J. L., & Beiser, D. G. (1995). *Models of information processing in the basal ganglia*. Cambridge: MIT Press.

Huang, Y. -Y., Simpson, E., Kellendonk, C., & Kandel, E. R. (2004). Genetic evidence for the bidirectional modulation of synaptic plasticity in the prefrontal cortex by D1 receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 3236–3241.

Jay, T. M. (2003). Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology*, *69*, 375–390.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 313–327.

Lee, D. (2006). Neural basis of quasi-rational decision making. *Current Opinion in Neurobiology*, *16*(2), 191–198.

Lee, D., Conroy, M. L., McGreevy, B. P., & Barraclough, D. J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research*, *22*, 45–58.

Lee, D., McGreevy, B. P., & Barraclough, D. J. (2005). Learning and decision making in monkeys during a rock-paper-scissors game. *Cognitive Brain Research*, *25*, 416–430.

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. *Journal of Comparative Neurology*, *432*, 119–136.

Liberman, B. (1962). Experimental studies of conflict in some two-person and three-person games. In J. H. Criswell, H. Solomon, & P. Suppes (Eds.), *Mathematical methods in small group processes* (pp. 203–220). Stanford: Stanford Univ. Press.

McCoy, A. N., & Platt, M. L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience*, *8*, 1220–1227.

Messick, D. M. (1967). Interdependent decision strategies in zero-sum games: a computer-controlled study. *Behavioural Science*, *12*, 33–48.

Mookherjee, D., & Sopher, B. (1994). Learning behaviour in an experimental matching pennies game. *Games and Economic Behaviour*, *7*, 62–91.

Neuringer, A. (1986). Can people behave "randomly?" the role of feedback. *Journal of Experimental Psychology: General*, *115*, 62–75.

O'Connor, D. H., Wittenberg, G. M., & Wang, S. S. -H. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 9679–9684.

Otani, S., Daniel, H., Roisin, M. -P., & Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cerebral Cortex*, *13*, 1251–1256.

Petersen, C. C., Malenka, R. C., Nicoll, R. A., & Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 4732–4737.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*, 352–363.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*, 1009–1023.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

Reynolds, J. N., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*, 507–521.

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

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Review Neuroscience*, *1*, 199–207.

Schultz, W. (2006). Behavioural theories and the neurophysiology of reward. *Annual Review Psychology*, *57*, 87–115.

Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks*, *16*, 5–9.

Soltani, A., & Wang, X. J. (2006). A biophysically-based neural model of matching law behavior: melioration by stochastic synapses. *Journal of Neuroscience*, *26*(14), 3731–3744.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton: Princeton Univ. Press.

Sugrue, L. P., Corrado, G. C., & Newsome, W. T. (2004). Matching behaviour and representation of value in parietal cortex. *Science*, *304*, 1782–1787.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Review Neuroscience*, *6*, 363–375.

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

Thorndike, E. L. (1911). *Animal intelligence; experimental studies*. New York: Macmillan.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton: Princeton Univ. Press.

Wang, X. -J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*, 955–968.