Model-based reinforcement learning under concurrent schedules of reinforcement in rodents

Namjung Huh, Suhyun Jo, Hoseok Kim, Jung Hoon Sul, and Min Whan Jung

Neuroscience Laboratory, Institute for Medical Sciences and Division of Cell Transformation and Restoration, Ajou University School of Medicine, Suwon 443-721, Korea

Reinforcement learning theories postulate that actions are chosen to maximize a long-term sum of positive outcomes based on value functions, which are subjective estimates of future rewards. In simple reinforcement learning algorithms, value functions are updated only by trial-and-error, whereas they are updated according to the decision-maker’s knowledge or model of the environment in model-based reinforcement learning algorithms. To investigate how animals update value functions, we trained rats under two different free-choice tasks. The reward probability of the unchosen target remained unchanged in one task, whereas it increased over time since the target was last chosen in the other task. The results show that goal choice probability increased as a function of the number of consecutive alternative choices in the latter, but not the former task, indicating that the animals were aware of time-dependent increases in arming probability and used this information in choosing goals. In addition, the choice behavior in the latter task was better accounted for by a model-based reinforcement learning algorithm. Our results show that rats adopt a decision-making process that cannot be accounted for by simple reinforcement learning models even in a relatively simple binary choice task, suggesting that rats can readily improve their decision-making strategy through the knowledge of their environments.

Animals must continually update their behavioral strategies according to changes in an environment in order to optimize their choices. Reinforcement learning (RL) models (Sutton and Barto 1998) provide a powerful theoretical framework for understanding choice behavior in humans and animals in a dynamic environment. In theories of RL, future actions are chosen so as to maximize a long-term sum of positive outcomes, and this can be accomplished by a set of value functions that represent the amount of expected reward that is associated with particular states or actions. The value functions are continually updated based on the reward prediction error, which is the difference between the expected and actual rewards. This way, even without prior knowledge about an uncertain and dynamically changing environment, an animal can discover the structure of the environment that can be exploited for optimal choice by trial-and-error. Not surprisingly, human and monkey choice behaviors in various tasks are well described by reinforcement learning algorithms (e.g., O’Doherty et al. 2003; Barraclough et al. 2004; Lee et al. 2004; Samejima et al. 2005; Daw et al. 2006; Pessiglione et al. 2006).

The updating of value functions can be achieved in two fundamentally different ways. In simple or direct RL algorithms, value functions are updated only by trial-and-error. In other words, only the value function that is associated with the chosen action is updated, and those that are associated with uncommitted actions remain unchanged. On the other hand, in indirect or model-based RL algorithms, the value functions might also change according to the decision-maker’s knowledge or model of the environment (Sutton and Barto 1998). It is likely that both updating processes are employed by humans and animals to different degrees in various circumstances. Unfortunately, however, the relative contributions of the two processes to value function update are largely unknown. Thus, to understand choice behavior in humans and animals fully, it is important to understand the extent to which direct RL models can explain choice behavior and the extent to which additional components are required, as in model-based RL algorithms.

In real life, expected outcomes of actions often change in a predictable manner. It would then be advantageous to incorporate this knowledge into the decision-making process. One notable example is the time-dependent increase of reward probability. In many natural settings, the availability of food at a particular patch decreases by foraging activity and then increases over time. Let us assume that an animal decides each morning where to go for foraging. As the number of consecutive visits to the other locations increases, the chance for obtaining food at the skipped location will increase. It would be advantageous for the animal to take this into consideration and choose foraging locations. In this regard, recent studies have shown that the history of past choices, as well as past rewards, has a strong effect on the current choice of monkeys (Lee et al. 2004; Lau and Glimcher 2005).

We trained rats in two different types of binary choice tasks to investigate this matter. In one version, a particular choice only counted toward the ratio requirement of that choice, so the probability of obtaining a reward is independent of the history of alternative choices. In the other version, either choice counted toward the ratio requirements of both choices, so the probability of obtaining a reward was dependent on the history of alternative choices. We compared a simple RL model with a modified model that incorporated the knowledge that the reward probability of the unchosen target increases. The results show that the behavioral data of the latter (dependent) task are better explained by the modified model, suggesting that rats can discover time-dependent increase of reward probability and use this information in making behavioral choices to maximize their income.

Results

Animal behavior

We employed in this study independent and dependent versions of the concurrent variable-ratio–variable-ratio (VR–VR) task (Fig. 1A;
Macdonall 1988). The probability to obtain a reward with a particular action was dependent on the history of alternative choices in the dependent task, but not in the independent concurrent VR-VR task. The independent and dependent versions of the task employed in the present study are equivalent to, and will be referred to as, the two-armed bandit (TAB) (Estes 1950) and dual assignment with hold (DAWH) tasks (Staddon et al. 1981; Lau and Glimcher 2005; Kennerley et al. 2006), respectively. They were free binary choice tasks with each choice associated with a different probability of reward that was constant within a block of trials (35–45), but changed across blocks. Although the mean probability of reward delivery was constant within a block of trials, the reward was delivered stochastically in each trial, and no explicit sensory information on reward probability was available to the animal. Hence, block transition could be discovered only by trial-and-error. In the TAB task, rats TAB1 and TAB2 performed a total of 12 and 14 sessions (one daily session consisting of four blocks of trials), respectively. Similarly, in the DAWH task, rats DAWH1 and DAWH2 performed a total of 15 and 14 sessions (one daily session consisting of four blocks of trials), respectively. In all sessions, rats discovered different reward probabilities and biased their choices accordingly within 10–20 trials after block transition (Fig. 1B,C).

Matching behavior
For the analysis of steady-state choice behavior, we analyzed the last 20 trials of each block for their conformity to the generalized matching law (Baum 1974, 1979), as follows:

\[
\frac{C_L}{C_R} = b \left( \frac{R_L}{R_R} \right)^a,
\]

where \(C_L\) (or \(C_R\)) and \(R_L\) (or \(R_R\)) are choice frequency and reinforcement frequency (i.e., the frequency of obtaining a reward) for the left (or right) goal, respectively. The coefficient \(a\) denotes the sensitivity to the reinforcement ratio and \(b\) is a bias term. Fitted data to the generalized matching law are shown in Figure 1D,E for the TAB and DAWH tasks, respectively. The experimental data showed good fits to the regression lines that were estimated by the generalized matching law irrespective of the task. The regression coefficient \(a\) (slope) was significantly positive in all cases (t-test, \(P < 0.001\) in all cases).

Dependence of choice behavior on run length in DAWH task
We then examined whether a rat’s choice behavior in the DAWH task depends on the latest run length (i.e., the number of consecutive alternative choices, \(n_{rle}\) note that this is not meant to be the distance the animal traveled). If an animal’s trial-by-trial choice behavior varies systematically as a function of run length, this would be an indication that the animals were aware of a run length-dependent increase in arming probability (Fig. 2) and used this information in choosing goals. We also used the last 20 trials of each block (i.e., steady-state data) for this analysis. Because an animal’s choice and reinforcement frequencies were stable at this stage, the animal’s choice behavior would be less compounded by other factors, such as potential variations in learning rate and degree of exploration. Essentially, the same conclusions were obtained, however, when we included all trials in the analysis (data not shown). As shown in Figure 3A, the probability to choose the left goal in the DAWH task tended to vary according to run length in all combinations of arming probabilities (i.e., 0.72:0.12, 0.63:0.21, 0.21:0.63, and 0.12:0.72) in both rats. Here, consecutive left and right goal choices were indicated as negative and positive values of \(n_{rle}\) respectively. Logistic regression analyses indicated a significant positive relationship between \(n_{rle}\) and \(P_L\) for all cases (\(P < 0.005\)) except one. See Table 1 for the magnitude of regression coefficient and \(P\)-value for each combination of arming probabilities for each rat. These results suggest that the animals were aware of run length-dependent changes in arming probability and used this information for action selection.

Dependence of choice behavior on relative arming probability in DAWH task
We also examined how an animal’s choice behavior changed according to relative arming probabilities of the two goals. Run length-dependent arming probability in the DAWH task will be referred to as “stacked arming probability,” because the arming probability of the unchosen target was “stacked” during the alternative choices. We related an animal’s choice to stacked arming probability \(S_{action}(n_{rle})\) as follows:

\[
P_L(n_{rle}) = c_1 \left( \frac{S_L(n_{rle})}{\sum_{k=1}^{2} S_k(n_{rle})} \right)^{c_2},
\]

where \(P_L(n_{rle})\) or \(P_R(n_{rle})\) is the probability to choose the left (or right) goal at a particular value of the latest run length and \(c_1\) and \(c_2\) are constants. The stacked arming probability in the DAWH task is given by a function of the block arming probability \(a_{action}\) and the latest run length \(n_{rle}\) as follows:

**Figure 1.** Behavioral task. (A) Rats were tested on a modified figure-eight-shaped maze to choose between two locations (circles) that delivered water reward with different probabilities. Arrows indicate movement directions of the animal. (B, C) Examples of choice behaviors of two rats performing the (B) TAB or (C) DAWH tasks. The probability to choose the left goal \((P_L)\) is plotted (moving average of 20 trials) across four blocks with different combinations of reward probability [black curve: actual choice of the animal, gray curve: choice of the (B) RW and (C) SP models, respectively (average of 100 simulations)]. Block transitions are marked by vertical lines. Numbers in each block indicate block arming probabilities associated with the left and right goal choices. (D, E) Conformity of steady-state choice behaviors to the generalized matching law. Steady-state behavioral data (last 20 trials of each block) were analyzed for each animal in the (D) TAB and (E) DAWH tasks. Five sessions were combined for noise reduction in an overlapping manner for the calculation of each data point. The lines were determined by logistic regression relating animal’s goal choice and the log reinforcement ratio.
Model-based reinforcement learning in rodents

Model comparison

Simulated choice behavior

We examined whether simulated choice behaviors of two different RL models, with or without considering stacked arming probability, show similar run length dependence as in the experimental data. The first model was a simple RL model that updated its action value functions according to the Rescorla–Wagner rule (RW model) (Rescorla and Wagner 1972), and the second model added “stack parameter” (see Materials and Methods) to the RW model (stack probability model or SP model). Both models generated the same number of trials with the same block sequence as in the behavioral data of the animal DAWH1 using parameters α (learning rate) and β (inverse temperature, see Materials and Methods) that were estimated from the behavioral data of the animal DAWH1. Trial-by-trial estimates of α and β were averaged for each session and then across the entire sessions to yield single values (0.2 and 1.9, respectively). The simulation was repeated 100 times for each model, and, as in the analysis of experimental data, only the last 20 trials in each block (steady state) were analyzed.

Figure 5 shows analysis results of the simulated behavioral data that were generated by the RW or SP model. To facilitate comparisons with the experimental data, we also show regression lines that were obtained from the behavioral data of the animal DAWH1. The regression lines in Figure 5A,B are identical to those in Figure 3A,B (left) and those in Figure 5C are identical to the lines in the left panel in Figure 1E. As expected, the simulated choice data generated by both models showed global matching behavior.
that was similar to that of the experimental data shown in Figure 1 (Fig. 5C). However, only trial-by-trial choice behavior of the SP model showed dependence on run length (and hence on the stacked arming probability) as in the experimental data, whereas that of the RW model was independent of run length. There was a significant positive relationship between run length and $P_L$ (mean of 100 simulations) in all block arming probability conditions for the data generated by the SP model, but not for the data generated by the RW model (Fig. 5A). See Table 2 for the magnitude of regression coefficient and $P$-value for each block arming probability condition. Furthermore, the squared difference between the simulation data ($P_L$, mean of 100 simulations; symbols in Fig. 5A) and the experimental data (lines in Fig. 5A) was smaller for the SP model (mean ± SD, 0.002 ± 0.002) than for the RW model (0.009 ± 0.004; Wilcoxon rank sum test, $z = –2.88, P = 0.040$). Also, significant linear relationships were observed between log stacked arming probability ratio and log choice ratio (mean of 100 simulations) for each block arming probability condition for the data generated by the SP model. On the other hand, only one of the regression slopes was significant for the data generated by the RW model, and even in this case, the slope was negative (Fig. 5B). The magnitudes of regression coefficients and $P$-values for each block arming probability condition are summarized in Table 2. Again, the squared difference between the simulation data (the log choice ratio, mean of 100 simulations; symbols in Fig. 5B) and the experimental data (solid line in Fig. 5B) was smaller for the SP model (mean ± SD, 0.042 ± 0.035) than for the RW model (0.276 ± 0.176; Wilcoxon rank sum test, $z = –3.84, P < 0.001$). Similar results were obtained when the behavioral data obtained from the rat DAWH2 ($\alpha = 0.2, \beta = 1.6$) were analyzed in the same manner (data not shown). These results indicate that an animal’s choice behavior in the DAWH task is well captured by the SP, but not the RW model, further corroborating the conclusion that the animals actively updated their choice preference, incorporating run length information in the DAWH task.

**Model comparison: Predictive performance**

We also compared how well the two models predict the next choice of the animal based on the previous behavioral data (predictive performance) by calculating the likelihood of model prediction ($L(i)$) as follows:

$$L(i) = P(a(i)|[a(j)], r(j)|1 \leq j \leq i – 1),$$

where $P(a(i)|[a(j)], r(j)|1 \leq j \leq i – 1)$ denotes the probability of a model to generate the actual action chosen by the animal in the $i$th trial given the history of the animal’s choices $[a(j)]$ and their outcomes $r(j)$ up to the $(i–1)$-th trial. This process was repeated from trial 10 to the last trial of each session and then for all behavioral sessions.

Figure 6A shows the mean likelihood of model prediction per trial at run lengths between 1 and 4 in the DAWH task (total numbers of trials: 2194, 989, 461, and 232, respectively). The predictive performance of the RW model decreased linearly with the latest run length, indicating that the RW model’s prediction of an animal’s action in the DAWH task progressively deviates from the actual choice of the animal as the run length increases (linear

---

**Table 1. Results of regression analyses of the experimental data**

| Run length vs. log($C_i/C_0$) | Rat   | DAWH1 | DAWH2 | TAB1 | TAB2 |
|-------------------------------|-------|-------|-------|------|------|
|                               | 0.12  | 0.28* | 0.22  | 0.092| 0.052|
|                               | 0.72  | 0.011 | (0.11) | 0.556| (0.658)|
|                               | 0.63  | 0.171 | (0.040)| –0.211| –0.206|
|                               | 0.21  | 0.37* | (0.001)| (0.139)| (0.222)|
|                               | 0.63  | 0.017 | (0.035)| 0.004| 0.389|
|                               | 0.21  | 0.340 | (0.003)| 0.978| (0.203)|
|                               | 0.63  | 0.284*| (0.020)| 0.463| –0.348*|

| log($S_i/S_a$) vs. log($C_i/C_0$) | Rat   | DAWH1 | DAWH2 | TAB1 | TAB2 |
|-----------------------------------|-------|-------|-------|------|------|
|                                  | 0.12  | 0.720*| (0.015)| (0.129)| 0.409| (0.084)|
|                                  | 0.21  | 0.357 | (0.129)| 0.985*| (0.002)| 0.805*| (0.014)|
|                                  | 0.63  | 0.946*| (0.002)| 0.985*| (0.002)| 0.805*| (0.014)|
|                                  | 0.21  | –0.009| (0.107)| –0.924| (0.946)| 0.346| (0.464)|
|                                  | 0.110 | –0.501| (0.218)| 0.867| (0.203)| –1.013*| (0.003)|

Summary of the coefficients for logistic regressions that relate animal’s goal choice and run length (top) and those that relate animal’s choice and log stacked arming probability ratio (bottom) of the experimental data. The coefficients indicate the slopes of the lines relating log choice ratio and run length (top) or log choice ratio and log stacked arming probability ratio (bottom). Each row represents analysis results obtained from one rat. Numbers in the parentheses indicate $P$-values. Columns 2–5 represent different block arming probability conditions. The significance of coefficients is indicated by asterisks (*: $P < 0.05$).
regression analysis, \( r = -0.010, P < 0.001 \). On the other hand, the performance of the SP model showed no such tendency (linear regression analysis, \( r = 0.000, P = 0.978 \)). Also, the mean likelihood of prediction was significantly larger for the SP model at run lengths \( >1 \) (mean likelihood for all trials with run length 2–4, RW model: \( L = 0.515 \pm 0.003 \), SP model: \( L = 0.527 \pm 0.003 \); Wilcoxon signed rank test, \( z = -4.39, P < 0.001 \)), indicating a superior performance of the SP to the RW model under long run length conditions. On the other hand, in the TAB task (total numbers of trials with run length 1–4: 1269, 607, 397, and 279, respectively), the mean likelihood of prediction was significantly larger for the RW model (mean likelihood for all trials with run length 1–4, RW model: \( L = 0.581 \pm 0.004 \), SP model: \( L = 0.545 \pm 0.004 \); Wilcoxon signed rank test, \( z = -13.6, P < 0.001 \); Fig. 6B), indicating a superior performance of the RW compared to the SP model. There was a significant positive relationship between the predictive performance of the RW model and the latest run length (linear regression analysis, RW model: \( r = 0.016, P < 0.001 \); SP model: \( r = 0.004, P = 0.306 \)).

For further confirmation, we generated simulated choice behavior using the SP or RW model and applied the same analyses to the simulated data. The SP (or RW) model generated the same numbers of trials with the same block sequences as in the behavioral data of the animals DAWH1 and DAWH2 (or TAB1 and TAB2) 10 times using parameters separately estimated from each animal (DAWH1: \( \alpha = 0.2, \beta = 1.9 \); DAWH2: \( \alpha = 0.2, \beta = 1.6 \); TAB1: \( \alpha = 0.2, \beta = 1.8 \); TAB2: \( \alpha = 0.2, \beta = 2.2 \)), and the generated data were combined for analysis. As shown in Figure 6C, for the data generated by the SP model, the performance of the RW model, but not the SP model, decreased linearly with the latest run length (linear regression analysis applied to mean likelihood per trial, RW model: \( r = -0.007, P = 0.047 \); SP model: \( r = 0.001, P = 0.913 \)), and the mean likelihood of prediction per trial was significantly larger for the SP model at run lengths \( >1 \) (mean likelihood for all trials with run lengths 2–4, RW model: \( L = 0.528 \pm 0.001 \), SP model: \( L = 0.542 \pm 0.001 \); Wilcoxon signed rank test, \( z = -15.4, P < 0.001 \)). These results further indicate that an animal’s choice behavior at run lengths \( >1 \) in the DAWH task is better captured by the SP than RW model. Likewise, for the data generated by the RW model (Fig. 6D), the predictive performance of the RW model was superior to the SP model (mean likelihood for all trials with run lengths 1–4, RW model: \( L = 0.550 \pm 0.001 \), SP model: \( L = 0.537 \pm 0.001 \); Wilcoxon signed rank test, \( z = -19.1, P < 0.001 \)). Here, the predictive performance of the SP model decreased linearly with the latest run length (linear regression analysis applied to mean likelihood per trial, RW model: \( r = 0.005, P = 0.065 \); SP model: \( r = -0.006, P = 0.007 \)).

**Trial–by-trial dynamics of model parameters**

The reason why the two models generated different choices under a long run length situation is illustrated in an example behavioral session in Figure 7. It shows trial-by-trial behavioral choice data of one DAWH task session, along with the estimated block arming probability (\( \beta^{\text{stim}} \)) and the estimated stack arming probability (\( \beta^{\text{action}} \); see Materials and Methods) for each goal choice, and the probability to choose the left goal (\( P_L \)) of the SP and RW models. Although the overall tendency (i.e., low frequency component) of choice probability over the entire session was similar between the SP and RW models, local behaviors of choice probability (i.e., high-frequency

---

**Table 2. Results of linear regression analyses of the simulated data**

| Run length vs. \( P_L \) | Model | \( 0.12 \) | \( 0.21 \) | \( 0.63 \) | \( 0.72 \) |
|--------------------------|------|--------|--------|--------|--------|
| SP | 0.043* (0.027) | 0.055* (0.001) | 0.052* (0.004) | 0.046* (0.035) |
| RW | -0.004 (0.357) | -0.009 (0.050) | -0.007 (0.199) | -0.004 (0.080) |
| \( \log(S_L/S_R) \) vs. \( \log(C_L/C_R) \) | Model | \( 0.12 \) | \( 0.21 \) | \( 0.63 \) | \( 0.72 \) |
|--------------------------|------|--------|--------|--------|--------|
| SP | 0.654* (0.002) | 0.809* (0.001) | 0.818* (0.001) | 0.676* (0.001) |
| RW | -0.065 (0.153) | -0.179 (0.229) | -0.080 (0.143) | -0.122* (0.003) |

Summary of the coefficients for linear regressions that relate run length and the probability for left goal choice (\( P_L \); mean of 100 simulations, top) and those that relate log stacked arming probability ratio and log choice ratio (mean of 100 simulations, bottom) of the simulated data generated by the RW and SP models. Table is set in the same format as Table 1. The significance of coefficients is indicated by asterisks (*: \( P < 0.05 \)).
Figure 6. Predictive performance of the RW and SP models. Each model’s predictive performance was assessed by calculating the likelihood of prediction per trial at different run lengths for the behavioral data obtained from the (A) DAWH and (B) TAB tasks. Predictive performances of the two models were also compared using simulated choice data generated by the (C) SP and (D) RW models. The graphs show mean likelihood per trial at different run lengths. Error bars denote SEM. Filled and open circles indicate the results predicted by the SP and RW models, respectively.

Discussion

We examined in this study behavioral strategies of rats in two binary choice tasks with different reinforcement updating rules. The two tasks were different only in that arming probability stayed constant or increased with the alternative choices. Behavioral, as well as modeling, data provide converging evidence that rats discovered the reinforcement updating rules and used this information in deciding which goal to visit. First, choice behavior in the DAWH task showed dependence on the latest run length; as the animals chose one goal consecutively, they were more likely to choose the other goal. Importantly, run length dependence was not observed in the TAB task, indicating that run length dependence is not due to an innate tendency of the animal to alternate, but to a correctly updated arming probability of the unvisited goal. In this regard, a recent study has shown that pigeons tended to switch their choices more often than expected by chance in a behavioral task that is similar to the DAWH task employed in the present study (Jensen and Neuringer 2008). Also, compared with a simple RL algorithm, a model-based RL algorithm that incorporated run length-dependent increase in arming probability better predicted an animal’s choice in high-run length (>1) trials and generated simulated choice data that are consistent with the experimental data in the DAWH task. These results clearly show that a standard RL model cannot account for a rat’s choice behavior in the DAWH task.

The simple (direct) and model-based (indirect) RL models update value functions in fundamentally different ways. Both models have advantages as well as disadvantages. For example, the model-based RL algorithm is advantageous in that it allows the maximal use of limited information about the environment when the amount of exploration is small. In such a circumstance, the simple RL algorithm cannot do much, because meaningful value functions have yet to be constructed. On the other hand, the model-based RL algorithm is prone to biases. If a model of the environment is incorrect or biased, then behavioral choices based on the model will be suboptimal. The direct RL model does not suffer from this problem because it updates value functions only by trial-and-error (Sutton and Barto 1998). Thus, depending on the problem at hand, one algorithm could be advantageous over the other. In this regard, it would be important to understand circumstances in which humans or animals employ simple vs.  

Figure 7. An example illustrating divergent behavior between simple and model-based RL algorithms. Predictions of goal choice based on the RW and SP models were compared. (A) Trial-by-trial estimated block arming probability ($A_{\text{est}}$) and estimated stack parameter ($X_{\text{est}}$) are shown across four blocks of trials. Solid and dotted lines indicate SP model parameters for the left and right choices, respectively. Tick marks indicate actual choices of the animals (left choice [upward] and right choice [downward]) with the long and short ones denoting rewarded and unrewarded trials, respectively. (B) The probabilities to choose the left goal ($P_L$) estimated by the (B) SP and (C) RW models are shown. Dashed boxes indicate the epochs where the preferred choice probability is differentiated between the SP and the RW models.
model-based RL algorithms in the process of decision making. Currently, the relative contributions of the two processes to value function update are largely unknown. In general, the relative importance of deliberate cognitive control vs. trial-and-error-based behavioral control has been debated for a long time, but is far from being clear. In this regard, recent studies have shown that reward prediction error signals in monkey midbrain dopamine neurons in a one-direction-rewarded task are context dependent, so their activities are better accounted for by a model-based RL algorithm (Nakahara et al. 2004), and that human choice behavior in a simple reversal learning task is better explained by a model-based RL algorithm (Hampton et al. 2006), and that human choice behavior in a one-armed bandit task with variable uncertainty cannot be well accounted for by a simple RL algorithm (Behrens et al. 2007). In line with these reports, our results suggest that even rats adopt a decision-making process similar to a model-based RL algorithm in a relatively simple binary choice task. Combined, these studies suggest that animals and humans possess a strong capability to exploit available information about the environment for adaptive decision making, and that such a capability might be a general characteristic across different animal species. If so, it is likely that the vast majority of choice behaviors cannot be explained by simple RL algorithms, but requires model-based RL algorithms to be fully accounted for.

Assuming that the animals made choices in the DAWH task based on an SP model-like process, it is unknown how they computed stacked arming probability. In our SP model, the estimated block arming probability ($A_{\text{est}}^{\text{block}}$) and the estimated stack parameter ($X_{\text{est}}^{\text{stack}}$) were calculated in an interactive manner. The estimated stack parameter was used in calculating the estimated block arming probability and vice versa (Fig. 8). It is unknown whether the brain computes stacked arming probability as in our SP model. It would be important in future studies to test whether and where neural signals that are related to the estimated block arming probability, stack parameter, and stacked arming probability are found. As another possibility, which is not mutually exclusive to the scenario above, the estimate of reward-based arming probability (i.e., action value function estimated according to a simple RL algorithm) and the latest run length might be separately computed before being combined to estimate the final stacked arming probability. Physiological studies have found neural signals that are related to action value functions that were computed based on a simple RL algorithm (Barralough et al. 2004; Samejima et al. 2005; Seo and Lee 2007) and neural signals that are related to animal’s previous choice (Barralough et al. 2004; Kim et al. 2007; Seo and Lee 2008) or the number of self-executed actions (Sawamura et al. 2002) in cortical and subcortical brain structures. The latter may represent the latest run length in the DAWH task. These results suggest separate processing of reward-based estimation of action value functions (a simple RL-like process) and run length estimation in the brain. It remains to be determined, however, how and where such signals are combined to generate stacked arming probability.

Choice behavior of the animals was consistent with the generalized matching law in both TAB and DAWH tasks, even though the animals did not show run length-dependent choice behavior in the TAB task. In general, making choices according to relative arming probabilities (probability matching) (Bitterman 1965) is frequently observed in a TAB task in various animal species including monkeys (Morris et al. 2006) and humans (Vulkan 2000). It is also well known that animals tend to alternate between two rewarding choices (Richman et al. 1986; Lalonde 2002). It is unclear why such a behavioral strategy, which is suboptimal in the TAB task, is adopted across different animal species. It is possible that arming probability conditions, as in the TAB task, are extremely rare so that alternating choices is, in general, advantageous to maximize income in natural settings. On the other hand, animals might alternate their choices to explore potential changes in arming probabilities because reward probabilities often change dynamically. Other accounts have also been proposed for probability matching behavior of humans (see Vulkan 2000). The reason for choice alternation perhaps depends on a number of factors including the structure of a specific task at hand. In our study, there was a need for exploration because reward probabilities changed dynamically. Hence, it is possible that the TAB animals in our study showed generalized matching behavior largely because of the need for exploration. In the RL models used in the present study, exploration was implemented by the randomness in action selection (parameter $\beta$). That choice behavior of the animals was closely approximated by RL models (Fig 1B,C) suggests that alternating choice behavior of the animals could be largely accounted for by the exploration term of the RL models in the present study.

Materials and Methods

Subjects

Experiments were performed with young male Sprague-Dawley rats ($\approx$9–11 wk old, 250–330 g, n = 4). Animals were individually housed in a colony room and initially allowed free access to food and water. They were then handled extensively while adapting to water deprivation (maintained at >80% ad libitum body weight) for 1 wk, and, once behavioral training began, restricted to 30 min of access to water after finishing one behavioral session per day. Experiments were performed in the dark phase of a 12-h light/dark cycle. The experimental protocol was approved by the Ethics Review Committee for Animal Experimentation of the Ajou University School of Medicine, Korea.

Apparatus

The overall dimension of the maze was 70 × 60 cm, and the width of the track was 8 cm (Fig. 1A). It was elevated 30 cm from the floor with 3–7-cm-high walls along the entire track except on the elevating central arm. Water (~30 μL) was delivered at two goal locations by solenoid valves in a probabilistic manner. Navigation of the animal was monitored by three sets of infrared beam detectors that monitored the animal’s arrival at the goal locations and return to the central section. Monitoring of animal behavior and water delivery were automatically
controlled by a personal computer using LabView software (National Instruments).

**Behavioral tasks**

Animals were trained in an independent or a dependent concurrent VR–VR task on the same maze. Each trial began by holding the animal in the central section of the maze by elevating the central arm that prevented the animal from accessing the goal locations. After a 2-sec delay, the central arm was lowered and the animal was allowed to navigate to one of the two goal locations. After visiting a goal, the animal was required to return to the starting location at the center of the maze by completing the remaining track. A new trial began as the animal returned to the central section of the maze. In the independent version of the concurrent VR–VR task, a particular choice only counted toward the ratio requirement of that choice, and the ratio requirement of the other goal did not change. Hence, the probability for obtaining a reward was independent of the history of alternative goal choices, and the task was equivalent to a TAB task (Estes 1950). Water was delivered only to the visited goal in this task and animals never failed to collect the reward when it was delivered. On the other hand, in the dependent version, either choice counted toward the ratio requirements of both choices, so that water was delivered to neither, either, or both goal locations in a given trial. If water was delivered to the unvisited goal, it was held there until the animal’s next visit with further delivery of water being prevented. This type of reinforcing schedule was named as the DAWH task (Staddon et al. 1981). Unlike the TAB task, the probability for obtaining a reward with a particular choice increased with the number of alternative choices in the DAWH task. Again, animals never failed to collect the reward when it was available at the visited goal. Change-over-delay was used in neither task.

In both versions of the task, each choice was associated with a different probability of reward that was constant within a block of trials, but changed across blocks. The number of trials in each block was 35 plus a random number of trials that were drawn from a geometric distribution with a mean of five trials, with the maximum number of trials set at 45.

All animals were tested across four blocks of trials on each day that consisted of the following arming probability ratios (left vs. right goals): 0.72:0.12, 0.12:0.72, 0.63:0.21, and 0.21:0.63. The sequence of reward ratios was randomly selected with the constraint that the richer alternative always switched its spatial location across block transitions (e.g., 0.72:0.12 to 0.21:0.63 or 0.12:0.72) (Lau and Glimcher 2005).

**Training procedure**

Following 1 wk of handling and water deprivation, each animal was allowed to individually explore the maze for 1 h/d without water reward on the maze for 2 d and then with water delivered at both reward sites for two more days. For the next 3 d, they were guided to navigate from the central arm to either goal location to obtain water reward (arming probability = 1) and then to come back to the central arm via the lateral alley for 1 h/d. The animals were then divided into two groups (two animals in each group) and trained to perform either the TAB or DAWH task. For the initial 3 d, they were trained under 0.72:0.12 and 0.12:0.72 arming probability conditions with the minimum of two blocks per day. For the next 3 d, they were trained to perform four blocks with arming probability conditions for the first two blocks 0.72:0.12 and 0.12:0.72 (in random order) and those for the last two blocks 0.63:0.21 and 0.21:0.63. They were then trained to run four blocks of trials in random order as described above on each day. Data were collected from the fourth (or seventh) day of training according to the final protocol in the TAB (or DAWH) task.

**Determination of steady state**

For each animal, all trials in 0.72: 0.12 and 0.12:0.72 arming probability blocks were pooled together and the rest (0.63:0.21 and 0.21:0.63 arming probability blocks) were pooled together aligning the trials to the end of each block. The probability to choose the higher rewarding goal (i.e., 0.72 or 0.63) was calculated for each trial and smoothed with moving average of five trials. The maximum length of trials (from the last trial) in which the animal chose the higher rewarding goal more frequently than chance (one-tailed binomial test, P < 0.01) in all trials was determined as the steady state for each arming probability condition for each animal (n = 8). The number of trials in the steady state ranged between 20 and 31. To be consistent, we used the last 20 trials of each session for the analysis of steady-state choice behavior.

**Modeling**

Two different models were tested in the present study. A simple RL model was compared with a modified model that added “stacked arming probability” to the simple RL model.

**Rescorla–Wagner (RW) model**

We constructed a simple RL model in which the estimate of arming probability \( A_{\text{est}}(i) \) was updated according to Rescorla–Wagner rule as follows:

\[
A_{\text{est}}^{\text{est}}(i+1) = (1 - \alpha) A_{\text{est}}^{\text{est}}(i) + \alpha (i) \quad \text{if action = action}(i),
\]

\[
A_{\text{est}}^{\text{est}}(i+1) = A_{\text{est}}^{\text{est}}(i) \quad \text{else},
\]

where \( \alpha \) is the learning rate, \( r(i) \) represents the reward in the \( i \)th trial (1 if rewarded and 0 otherwise), and “action” indicates the selected action (left or right goal choice). The difference in local arming probabilities was represented as:

\[
R_D(i) = A_{\text{est}}^{\text{est}}(i) - A_{\text{est}}^{\text{est}}(i),
\]

where “L” and “R” denote left and right goal choices, respectively. Choice of action was determined by the softmax action selection rule (Bridle 1990) in which choice probabilities varied as a graded function of the estimated difference in local arming probabilities. Thus, the probability for selecting the left goal \( P_L(i) \) was defined as:

\[
P_L(i) = \frac{1}{1 + \exp(-\beta R_D(i))},
\]

where \( \beta \) is the inverse temperature that defines the degree of exploration in action selection. Parameters \( \alpha \) and \( \beta \) were estimated using the Bayesian inference method as previously described (Samejima et al. 2004, 2005).

**Stack probability (SP) model**

In the DAWH task, the arming probability changed trial-by-trial according to the number of consecutive alternative choices (run length). In other words, the arming probability of the unchosen target was “stacked” during the alternative choices. Thus, the simple RL model above (RW model) was modified to estimate run length-dependent arming probability (stacked arming probability) in the DAWH task. Because stacked arming probability is a function of the block arming probability and the latest run length, the estimated stacked arming probability associated with an action (i.e., left or right goal choice) in the \( i \)th trial \( S_{\text{est}}(i) \) was divided into two parts—estimated block arming probability \( A_{\text{est}}^{\text{est}}(i) \) and estimated stack parameter \( X_{\text{est}}^{\text{est}}(i) \), which increases with the number of consecutive alternative choices as follows:

\[
S_{\text{est}}(i) = A_{\text{est}}^{\text{est}}(i) X_{\text{est}}^{\text{est}}(i),
\]

where \( X_{\text{est}}^{\text{est}}(i) = \sum_{m=0}^{n} (1 - A_{\text{est}}^{\text{est}}(i))^{m} \) if action = L

\[
X_{\text{est}}^{\text{est}}(i) = \sum_{m=0}^{n} (1 - A_{\text{est}}^{\text{est}}(i))^{m} \quad \text{else},
\]

The estimated block arming probability was updated according to the Rescorla–Wagner rule as follows:

\[
P_L(i) = \frac{1}{1 + \exp(-\beta R_D(i))},
\]

where \( \beta \) is the inverse temperature that defines the degree of exploration in action selection. Parameters \( \alpha \) and \( \beta \) were estimated using the Bayesian inference method as previously described (Samejima et al. 2004, 2005).
After the block arming probability was estimated, the stack parameter was updated as follows:

\[
A_{\text{action}}^\text{est}(i+1) = (1 - \alpha)A_{\text{action}}^\text{est}(i) + \frac{\alpha}{X^\text{est}_{\text{action}}(i)} \cdot r(i) \text{ if action = action}(i)
\]

\[
A_{\text{action}}^\text{est}(i+1) = A_{\text{action}}^\text{est}(i) \text{ else.}
\]

The estimated difference in local arming probabilities was represented as:

\[
X^\text{est}_{\text{action}}(i+1) = 1 \text{ if action = action}(i)
\]

\[
X^\text{est}_{\text{action}}(i+1) = 1 + X^\text{est}_{\text{action}}(i) \cdot (1 - A_{\text{action}}^\text{est}(i)) \text{ else.}
\]

Choice of action was determined by the softmax action selection rule as in the RW model.

Statistical analysis

Linear and logistic regression analyses were employed to test the statistical significance of the relationship between various factors (e.g., run length vs. choice probability), and the Wilcoxon signed rank test was used to compare the predictive performance of the models. A P-value < 0.05 was used as the criterion for a significant statistical difference unless noted otherwise. Data are expressed as mean ± SEM unless noted otherwise.

Acknowledgments

We thank Daeyeol Lee for helpful comments on the initial version of the manuscript. This work was supported by the Korea Research Foundation grant KRF-2005-216-E00058 (N.H.); a 2007 grant from the Department of Medical Sciences, the Graduate School, Ajou University; Korea Science and Engineering Foundation grant KRF-2005-216-E00058 (N.H.); a 2007 grant from the Ministry for Health, Welfare and Family Affairs (A080742); and the Cognitive Neuroscience Program of the Korea Ministry of Science and Technology (M.W.J.).

References

Barraclough, D.J., Conroy, M.L., and Lee, D. 2004. Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. 7: 404–410.

Baum, W.M. 1974. On two types of deviation from the matching law: Bias and undermatching. J. Exp. Anal. Behav. 22: 231–242.

Baum, W. 1979. Matching, undermatching, and overmatching in studies of choice. J. Exp. Anal. Behav. 32: 269–281.

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10: 1214–1221.

Bittman, M.E. 1965. Phylectic differences in learning. Am. Psychol. 20: 396–410.

Bridle, J.S. 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In Proceedings of the 1989 Conference on Advances in Neural Information Processing Systems (ed. D.S. Touretzky), pp. 211–217. Morgan Kaufmann, San Mateo, CA.

Daw, N.D., O’Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. 2006. Cortical substrates for exploratory decisions in humans. Nature 441: 876–879.

Estes, W.K. 1950. Toward a statistical theory of learning. Psychol. Rev. 57: 94–107.

Hampton, A.N., Bossaerts, P., and O’Doherty, J.P. 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. 26: 8360–8367.

Herrnstein, R.J. 1982. Melioration as behavioral dynamism. In Quantitative analyses of behavior: Matching and maximizing accounts, Vol. 2 (eds. M.L. Commons et al.), pp. 433–458. Ballinger, Cambridge, MA.

Herrnstein, R.J. and Vaughan, W. 1980. Melioration and behavioral allocation. In Limits to action (ed. J.E.R. Staddon), pp. 143–176. Academic, New York.

Jensen, G. and Neuringer, A. 2008. Choice as a function of reinforcer “hold”: From probability learning to concurrent reinforcement. J. Exp. Psychol. 34: 437–460.

Kennerley, S.W., Walton, M.E., Behrens, T.E., Buckley, M.J., and Rushworth, M.F. 2006. Optimal decision making and the anterior cingulate cortex. Nat. Neurosci. 9: 940–947.

Kim, Y.B., Huh, N., Lee, H., Baeg, E.H., Lee, D., and Jung, M.W. 2007. Encoding of action history in the rat ventral striatum. J. Neurophysiol. 98: 3548–3556.

Lalonde, R. 2002. The neurobiological basis of spontaneous alternation. Neurosci. Biobehav. Rev. 26: 91–104.

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

Lee, D., Conroy, M.L., McCleery, R.P., and Barraclough, D.J. 2004. Reinforcement learning and decision making in monkeys during a competitive game. Brain Res. Cogn. Brain Res. 22: 45–58.

Macdonell, J.S. 1988. Concurrent variable-ratio schedules: Implications for the generalized matching law. J. Exp. Anal. Behav. 50: 55–64.

Morris, G., Nevet, A., Arkadit, D., Vaadia, E., and Bergman, H. 2006. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 9: 1057–1063.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. 2004. Dopamine neurons can represent context-dependent prediction error. Neuron 41: 269–280.

O’Doherty, J.P., Dayan, P., Pritson, K., Critchley, H., and Dolan, R.J. 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38: 329–337.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442: 1042–1045.

Rescorla, R.A. and Wagner, A.R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory (eds. A.H. Black and W.F. Prokasy), pp. 64–99. Appleton-Century-Crofts, New York.

Richman, C.L., Dember, W.N., and Kim, P. 1986. Spontaneous alternation behavior in animals: A review. Curr. Psychol. Res. Rev. 5: 358–391.

Samejima, K., Doya, K., Ueda, Y., and Kimura, M. 2004. Estimating internal variables and parameters of a learning agent by a particle filter. In Advances in neural processing systems, Vol. 16 (eds. S. Thrun et al.), pp. 1335–1342. MIT Press, Cambridge, MA.

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

Sawamura, H., Shima, K., and Tanji, J. 2002. Numerical representation for action in the parietal cortex of the monkey. Nature 415: 918–922.

Seo, H. and Lee, D. 2007. Temporal filtering of reward signals in the dorsolateral anterior cingulate cortex during a mixed-strategy game. J. Neurosci. 27: 8366–8377.

Seo, H. and Lee, D. 2008. Cortical mechanisms for reinforcement learning in competitive games. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363: 3845–3857.

Staddon, J.E., Hinson, J.M., and Kram, R. 1981. Optimal choice. J. Exp. Anal. Behav. 35: 397–412.

Sutton, R.S. and Barto, A.C. 1998. Reinforcement learning: An introduction. MIT Press, London, UK.

Vulkan, N. 2000. An economist’s perspective on probability matching. J. Econ. Surv. 14: 101–118.

Received December 3, 2008; accepted in revised form March 5, 2009.
Model-based reinforcement learning under concurrent schedules of reinforcement in rodents

Namjung Huh, Suhyun Jo, Hoseok Kim, et al.

Learn. Mem. 2009, 16:
Access the most recent version at doi:10.1101/lm.1295509

References

This article cites 28 articles, 4 of which can be accessed free at:
http://learnmem.cshlp.org/content/16/5/315.full.html#ref-list-1

License

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.