How Do Humans Handle the Dilemma of Exploration and Exploitation in Sequential Decision Making?

Naoya Namiki
Tokyo Denki University
Ishizaka, Hatoyama, Hiki,
Saitama, JAPAN
14rmd17[at]ms.dendai.ac.jp

Kuratomo Oyo
Tokyo Denki University
Ishizaka, Hatoyama, Hiki,
Saitama, JAPAN
kuratomo.oyo[at]gmail.com

Tatsuji Takahashi
Tokyo Denki University
Ishizaka, Hatoyama, Hiki,
Saitama, JAPAN
tatsujit[at]mail.dendai.ac.jp

ABSTRACT
In an uncertain environment, decision-making meets two opposing demands. One is to explore new information, while the other is to exploit already acquired information. The opposition is long called the exploration-exploitation dilemma. In brain science, it is known that human brain estimates options comparatively, and the average behavior correlates to the Softmax action selection rule. Softmax randomly chooses options with the selection probability that is a monotonous function of the estimated value. However, it needs a kind of pseudo-random number generator in human's mind. In cognitive psychology, it is indicated that recognition and generation of random sequence by human are quite biased, generally very unfaithful. Then, is it possible that humans adopt the Softmax policy while they are that bad at generating and recognizing random numbers? In this study, we analyzed how humans behave in face of the exploration-exploitation dilemma through experiments of the *N*-armed bandit problems and compared some policies commonly used in reinforcement learning modeling, from a viewpoint of whether humans really choose options randomly.

Keywords
Exploration-exploitation dilemma, N-armed bandit problems, Win-shift, Reinforcement learning

1. INTRODUCTION
In an uncertain environment, decision-making meets two opposing demands. One is gathering new information and the other is exploiting already known information. The opposition of the two are called the exploration-exploitation dilemma. Exploration is to search for a better option from the given options, and exploitation is to stick to the (subjectively) best option, utilizing the acquired experiences. This dilemma is a crucial factor for achieving the acquisition of maximum rewards. *N*-armed bandit problems form a class of the most basic problems of reinforcement learning, and represent the exploration-exploitation dilemma in a simplest way. Some algorithms are suggested for this problem [1].

In brain science, it is known that human's brain estimates options relatively and the behavior correlates to the Softmax policy that randomly chooses options with the selection probability proportional to the estimated value [2]. However, can humans utilize random numbers well? In cognitive psychology, it is indicated that humans are bad at recognizing the randomness of number sequences (e.g., see [3]). For example, imagine a toss-up. If some heads have come up successively, a person may consider that tails come up in the next toss, although each toss is independent. This tendency is called "gambler's fallacy". Humans tend to "discover" some rules in random sequences. This applies not only to recognition but also to generation of random sequences.

With the difficulty for humans to generate random sequences, do humans really tend to choose options randomly? Additionally, Softmax only correlates to the aggregated, averaged data, not to individual data. So we cannot be sure that Softmax well represents the policy that humans employ in sequential decision-making under uncertainty.

The exploration-exploitation dilemma is deeply related to sequential learning and humans' decision-making. Elucidating how humans handle the dilemma should promote understanding of their activities. By application of these characters, it might be able to make AI or robots to learn more autonomously. In this study, we analyze how humans behave in face of the exploration-exploitation dilemma through two experiments with *N*-armed bandit problems and compare the data with policies commonly used in reinforcement learning, focusing on whether humans really tend to randomly choose options.

2. THE EXPLORATION-EXPLOITATION DILEMMA
In order to maximize rewards, a decision-maker necessarily has to find out the best option and continue choosing it (exploitation). However, in an uncertain environment, the decision-maker necessarily has to try choosing various options (exploration) to find out the best option because it cannot be known without actual trials. The option with the highest expected value in reward is not necessarily the best option, since the trials can be not enough to objectively compare all the options. If she only concentrates on exploitation, she may overlook the best option, and hence fails to obtain maximum accumulated rewards. If a decision-maker only concentrates on exploration, she tends to fail to acquire rewards sufficiently higher than the average, so finally obtained rewards should be much lower than the maximum. If she could repeat the choices for infinite times, the exploration-exploitation dilemma would dissolve. However, in real environments, many factors (e.g.,

BICT 2014, December 01-03, Boston, United States

DOI 10.4108/icst.bict.2014.258045

time and other resources, and also other important things to do) may prevent her from infinite choices. So, to obtain as much rewards as possible, she has to somehow treat the dilemma well.

3. *N*-ARMED BANDIT PROBLEMS

In a *N*-armed bandit problems, there are *N* slot machines with different probabilities of hit assigned to each. A player does not know the probability of hit for each machine, but needs to choose one of the machines at a time. In order to maximize the acquisition of rewards, the player needs to search for the best slot machine (exploration), and once found, she should keep choosing it (exploitation). Thus, playing a *N*-armed bandit problem involves two kinds of actions, exploration and exploitation, and the problem represents the exploration-exploitation dilemma. Thus, in this study, we adopted the two-armed bandit problems as experiment tasks for observing how humans behave in face of the exploration-exploitation dilemma. We made the tasks two-armed bandit problems, considering the cognitive loads such as on working memory.

4. HOW DO HUMANS HANDLE THE EXPLORATION AND EXPLOITATION DILEMMA?

The exploration-exploitation dilemma has been studied as one of the most fundamental topics in reinforcement learning [1]. Recently, the dilemma has begun to be studied in brain science [2]. Particularly, it is studied how human's brain handles the dilemma by observing it with fMRI. Daw et al. observed the brain activity of participants playing a four-armed bandit problem. They studied neural substrates related to exploration and switching between exploration and exploitation. As a result, they found that the ventral medial prefrontal cortex (vmPFC) encodes the magnitude of rewards in a relative manner, and the frontal pole activates at exploration. Boorman observed participants' brains when they were playing a two-armed bandit problem, and studied the relationship of the exploration-exploitation dilemma and activation of brain regions [4]. They showed that the activity of vmPFC correlates to the relative value of the chosen option, and that a correlation between the frontal pole (FPC) signal and the relative unchosen probability. They suggested that the calculation in the frontal pole is significant for flexibility in humans' behavior under uncertainty.

From the above, it is found that human evaluates actions in a relative manner, rather than in an absolute manner, in face of the exploration-exploitation dilemma. It is supported by that behavior of human is often modeled by the Softmax action selection method which is based on a kind of relative evaluation [2]. However, because of the biased recognition and generation of random sequences [3], it should be quite difficult for humans to perform a probabilistic policy such as Softmax. Softmax is used only for modeling the aggregated data, not the individual sequences of actions. It is not known if individual behaviors correlate with Softmax.

5. MISCONCEPTION OF RANDOM SEQUENCES

It is difficult for humans to recognize random sequences correctly, and they may find out some rules out of contingencies in random sequences, which is called misconception of random sequences. For example, in sequential toss-up, humans may think that head is more likely to come up after a succession of heads (hot hand fallacy). Humans may also think, inversely, that tail is more likely to come up after a succession of heads (gambler's fallacy). As far as heads and tails come up at the equal probability, and the flips are independent, these tendencies are biases.

There is also the "law of small numbers" in the perception of random sequences by humans [3]. Humans tend to think that the samples extracted from a population have the typical characteristics of the population, which called the law of small numbers. For example, in toss-up, humans may think that the same number of heads and tails are more likely to come up as a result of random trials than it actually is. An extreme case is that, for a sequence of two toss-ups, "head, tail" and "tail, head" look significantly more typically a result of random process than "head, head" or "tail, tail". Humans have such misconception in recognizing randomness, and it is similar in generation, too. Therefore the doubt arises if human beings can really select actions like Softmax that is a policy premised on random choice of actions.

6. EXPERIMENTS

To test how humans handle the exploration-exploitation dilemma, especially in relation to the Softmax policy, we conducted two experiments. We focused on a type of action switch that we call *win-shift*, described below. Win-shift should be frequently enacted if humans follow a policy close to Softmax.

6.1 Policies in Reinforcement Learning

We introduce some representative policies in reinforcement learning for comparison with human behavior. The value of a slot machine is calculated as the expected value of reward, where hit and miss are coded as 1 and 0, (hence identical to the conditional probability of hit given the slot machine).

6.1.1 Greedy Policy

The greedy policy is the most basic strategy to choose always the subjectively best action (*greedy action*; one with the highest expected value among all the slot machines) at the time.

6.1.2 Epsilon-Greedy Policies

The epsilon-greedy policies are strategies that clearly separate exploration and exploitation. ϵ is a parameter in $[0,1]$. It chooses an option randomly at the probability of ϵ, or chooses the subjectively best option at the probability of $1-\epsilon$. There are several types of epsilon-greedy policy. In this study, we adopted the following three policies.

6.1.2.1 Epsilon Constant

The epsilon constant policy is a strategy with which ϵ is constant throughout the whole steps. In the simulation $\epsilon = 0.5$ and it means that the greedy action is chosen at the probability of 0.75 with two slot machines.

6.1.2.2 Epsilon Decreasing

In the epsilon decreasing policy, ϵ decreases gradually as a function of step, in this study with

$$\varepsilon = \frac{1.0}{1.0 + \tau t}, \quad (1)$$

where τ is the parameter for decay speed (set at 0.5) and t is the current step.

6.1.2.3 Epsilon First

The epsilon first policy is a strategy to select totally randomly until the determined step. The step τ satisfies $\tau = \epsilon n$, where n is the total number of steps. ϵ is set at 0.3.

6.1.3 The Softmax Action Selection Rule

Softmax is a policy that selects an action in a random manner. It somehow keeps the balance between exploration and exploitation. In this study, we used a version of Softmax method as in [5]. $P(X)$ is the selection probability of some option X, $M(1|X)$ is the value of some option X, τ is the parameter of speed of decreasing (analogous to temperature), t is the current step. τ is set at 0.5.

$$P(X) = \frac{\exp(M(1|X) \cdot \tau t)}{\Sigma_{X' \in \{A,B\}} \exp(M(1|X') \cdot \tau t} \quad (2)$$

6.2 Experiment 1

In this experiment, 39 participants (students of Tokyo Denki University) played two-armed bandit problems on a computer. There are two tasks; one is an easier task (the difference between the probabilities of giving hit of the two slot machines is larger), and the other is a more difficult task (the difference is smaller). The participants were randomly assigned to two groups. One group played the easier task, and then the more difficult (ED[1] group, N=17). The other played the more difficult task, then the easier (DE group, N=22). In the easier task, the hit probabilities of the two slot machines were 0.8 and 0.2. In the harder task, they were 0.6 and 0.4. The number of plays (steps) was 20 in the easier task and 40 in the more difficult task. Each slot machine gave only two kinds of reward, hit or miss. The participants were instructed that the problem is stationary: the hit probabilities assigned to the slot machines stay the same throughout a task.

In order to promote intuitive decision-making of humans, participants were not given most of information (e.g. the number of trials and hits on each arm so far, and the total number of steps in the task), while in most of previous studies, subjects know that kind of information (e.g., [6]).

6.3 Result of Simulation 1

We used the index of *win-shift* to observe whether people randomly select a slot machine to play. Win-shift is the proportion of trials/participants where the action to choose was switched even though the previous reward was hit. It is likely that "lose-shift" occurs very often, which is that a participant switches the machine to choose after getting negative feedback (miss). Win-shift should occur basically only when people select options in a somewhat random manner. It never occurs under the greedy policy, for instance. We classified participants based on the data of individual by win-shift as the main result, instead of the aggregated, average behavior.

The typical time development of the proportion of win-shift for the policies we described in section 6.1 is shown in Fig. 1. The greedy policy never causes win-shift. The epsilon constant, decreasing, and first policies have a predictable time development of win-shift proportions. Softmax is similar to epsilon decreasing. With the patterns of win-shift percentage change, we could model human behavior in terms of win-shift. People may have different policies from each other. Thus, we have classified human data in terms of steps when win-shift occurred. We equally divided the steps (20 or 40, depending on the problem) into the first, the middle, and the final terms. The behavior of the participants is classified according to in which phase a win-shift occurred at least

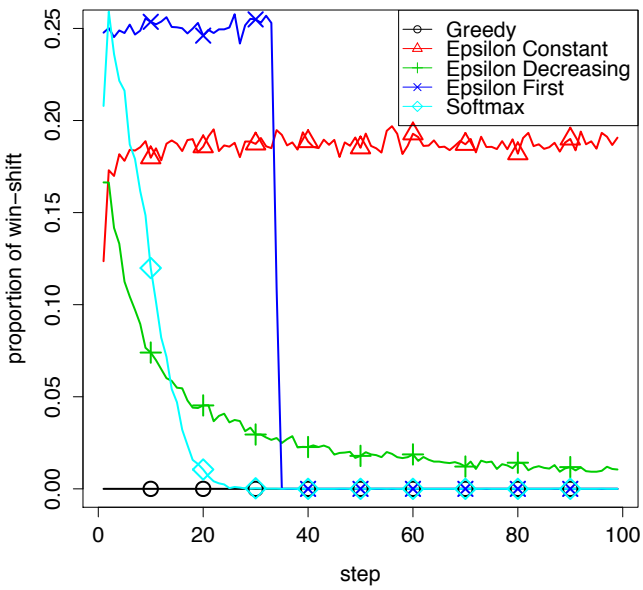

Figure 1. Time development of proportion of win-shift by reinforcement learning policies

once. Terms in which win-shift occurred (did not occur) are labeled as S (N). There are eight possible combinations: NNN, NNS, NSN, NSS, SNN, SNS, SSN, and SSS. For example, for a sequence of actions of type NNS, win-shift occurred only in the final term. The average accuracy (the percentage of trials/participants that selected the best action) and the percentage of each type are also shown in Table 1–4.

As in Table 1, in the ED group in the first easier task, the most common types are NNN and NNS. In Table 2, in the ED group in the last, more difficult task, the modal is NNN. As in Table 3, in the DE group in the second, easier task, the modal is NNN. The modal type in the first, harder task for the DE group is SSS, as in Table 4.

Table 1. Accuracy and percentages of types in the ED group and result of the policies in the first, easier task

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	89	35
SNN	93	12
NNS	80	6
SSN	73	35
SNS	60	12
Greedy	93	
Epsilon first	83	
Epsilon constant	93	
Epsilon decreasing	84	
Softmax	77	

[1] Easy-hard (hence EH and HE groups) contrast may be more natural than easy-difficult (ED and DE groups). However, to avoid redundancy with 'H' (high), we adopted easy-hard pair.

Table 2. Accuracy and percentages of types in the ED group and result of the policies in the last, more difficult task

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	63	35
SNN	79	23
NSN	74	12
SSN	34	12
SNS	55	6
NSS	63	6
SSS	48	6
Greedy	72	
Epsilon first	72	
Epsilon constant	69	
Epsilon decreasing	73	
Softmax	69	

Table 3. Accuracy and percentages of types in the DE group in the last, easier task

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	89	45
SNN	80	14
NSN	58	9
SSN	72	14
SNS	58	9
NSS	70	5
SSS	50	5

Table 4. Accuracy and percentages of types in the DE group in the first, more difficult task

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	64	27
SNN	63	14
NNS	86	9
SNS	60	14
NSS	88	5
SSS	56	32

6.4 Experiment 2

In the second experiment, 25 participants (students of Tokyo Denki University) played two-armed bandit problems on a computer. There are two problems; one is a high probabilities problem of 0.7 and 0.8, and the other is a low probabilities problem of 0.2 and 0.3. Similarly to Experiment 1, the participants were randomly assigned to two groups. The ones in the HL (LH) group played the high (low) probabilities problem first, and then the low (high) one. The number of participants is 11 for HL and 14 for LH. The number of steps people can play is 50 in both problems. Only differences from Experiment 1 are hit probabilities and the number of steps.

6.5 Results 2

Similarly to Experiment 1, we classified the data by the terms where win-shift occurred. For the HL group, in both problems, the modal type is SSS, as shown in Table 5 and 6. As for the LH group, both in the first, low probabilities problem, and in the last, high probabilities problem, the modal type was NNN.

Table 5. Accuracy and percentages of types in the HL group in the first, high probabilities problem

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	73	36
SNN	58	9
SSS	43	55

Table 6. Accuracy and percentages of types in the HL group in the last, low probabilities problem

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	79	18
NSN	46	18
NNS	36	9
SNS	58	9
SSS	51	45

Table 7. Accuracy and percentages of types in the LH group in the first, high probabilities problem

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	91	50
SNN	78	7
SSN	74	7
SSS	51	36

Table 8. Accuracy and percentages of types in the LH group in the last, low probabilities problem

Types / Policies	Accuracy (%)	Percentage of types (%)
NNN	56	36
SNN	45	14
NSN	56	7
NNS	10	7
SSN	42	7
SNS	35	21
SSS	60	7

7. DISCUSSION

In Experiment 1, except for the DE group in the difficult task, the most common type of people was NNN, in which win-shift is not observed at all. In the difficult task in the DE group, NNN was the second most common. In particular, win-shift did not appear in the half of people of the DE group in the easy task. Thus, in general, the result suggests that many people do not select an action in a random manner as in Softmax or epsilon-greedy policies.

In Experiment 2, in the HL group, the most common type was SSS in which win-shift occurred in all three phases, in both of the two problems with the high and low hit probabilities, with the second most common type NNN. In the LH group, the most common type was NNN in both of the two problems. In particular, it is remarkable in the high probabilities problem, the half of the participants was of NNN type. Thus, although not in all the problems, it is confirmed that there are many people who do not perform win-shift at all, and it suggests they generally do not probabilistically select actions. However, in both of the HL and LH groups, the percentage of type NNN was lower in the low probabilities problem than in the high one. It could be because the participants tended to use another policy than in other problems. The two problems given to the HL and LH groups were both difficult, which is that the difference in the hit probabilities of the two slot machines is hard to see with a small number of steps. It could be that the participants tried to see the difference by a kind of heuristics of giving a bunch of trials like, say, five times of trying a slot machine, and then trying the other for five times. If they did not find a significant difference after the ten times of plays, they could continue to do the same kind of action sequences. This is probably what we should look further into from the data and from more experiments.

From two experiments, in most of the problems, we see that people generally do not select actions randomly. People do not shift their action when they are given positive feedback (hit) to their action, and they shift only when people have been given negative feedback (miss).

In both Experiment 1 and Experiment 2, there were types win-shift occurred in the final term. Typical policies of reinforcement learning like the ones introduced in 6.1 do not cause win-shift much in the final term. It is possible that this is what we should consider modeling the policies of humans. The reason for this might be emotional factors like weariness and fatigue. In particular, in these experiments, participants were not informed the number of steps they could play. It could lead participants to be distracted from the task. This feature may be valuable for a non-stationary problem, where the hit probabilities of the slot machines mutate in the middle of the play. Although the participants were instructed that the problems are stationary, we could say that they could be unconsciously adaptive for non-stationary problems. Typical bandit-like problems in the wild like foraging and mating are not stationary nor Markovian, because of the real non-stationary character of the environment or incomplete observation and identification of states and actions. This aspect of possible adaptability to non-stationary problems or incomplete perception should be studied in future.

8. CONCLUSION

We investigated the behavior of people in face of the exploration-exploitation dilemma. As a result, it is found that they show properties different from the probabilistic policies like epsilon-greedy policies that separate (and switch between) exploration and exploitation, and the Softmax policy that balances the two strategies by probabilistically mixing them. Although it is generally considered that Softmax fits the behavior of people in a bandit problem, people do not act in a random manner as Softmax dictates; they switch action generally only when negative feedback is given (lose-shift). Furthermore, it was found that people have a tendency of making a win-shift, in the final term, which is different from most policies in reinforcement learning. Our results may help formalize the behavior of people in face of the exploration-exploitation dilemma which is inevitable in sequential decision making under uncertainty.

9. ACKNOWLEDGMENTS

This work was partially carried out under the supports by Grant-in-Aid for Scientific Research (KAKENHI) 25730150 from JSPS, the Cooperative Research Project Program H25/A12 of the Research Institute of Electrical Communication, Tohoku University, and Research Institute for Science and Technology of Tokyo Denki University Grant Number Q11K-02 and Q13K-03 / Japan.

10. REFERENCES

[1] Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

[2] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879, 2006.

[3] Tversky, A. and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases, *Science*, 185(4157), 124-1131, 1974.

[4] Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth M.F. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence n Favor of Alternative Courses of Action. *Neuron*, 62(5), 733-743, 2009.

[5] Oyo, K., Takahashi, T. A cognitively inspired heuristic for two-armed bandit problems: The loosely symmetric (LS) model. *Procedia Computer Science* 24 (2013) 194–204, 2013.

[6] Zhang, S. and Yu, A.J. Cheap but Clever: Human Active Learning in a Bandit Setting. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2013.