Human behavior in contextual multi-armed bandit problems

Hrvoje Stojic[1] (hrvoje.stojic@upf.edu),
Pantelis P. Analytis[2] (analytis@mpib-berlin.mpg.de), **Maarten Speekenbrink**[3] (m.speekenbrink@ucl.ac.uk)
[1]Department of Economics and Business, Universitat Pompeu Fabra
[2]Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development
[3]Department of Experimental Psychology, University College London

Abstract

In real-life decision environments people learn from their direct experience with alternative courses of action. Yet they can accelerate their learning by using functional knowledge about the features characterizing the alternatives. We designed a novel contextual multi-armed bandit task where decision makers chose repeatedly between multiple alternatives characterized by two informative features. We compared human behavior in this contextual task with a classic multi-armed bandit task without feature information. Behavioral analysis showed that participants in the contextual bandit task used the feature information to direct their exploration of promising alternatives. Ex post, we tested participants' acquired functional knowledge in one-shot multi-feature choice trilemmas. We compared a novel function-learning-based reinforcement learning model to a classic reinforcement learning. Although reinforcement learning models predicted behavior better in the learning phase, the new models did better in predicting the trilemma choices.

Keywords: decision making; reinforcement learning; exploration–exploitation trade-off; contextual multi-armed bandits; function learning

Introduction

George, an early-career American academic, has just accepted a new position at a European university. Somewhat of a culinary fanatic, he is determined to enjoy the local cuisine as much as possible. As there are over 1,000 restaurants in the area, he is spoiled for choice. George soon starts to try out different restaurants, sometimes leaving ecstatic and sometimes close to nauseous. Keen to avoid the latter, he notices that the quality of the food on offer is related to various pieces of information, such as the facade of the restaurant, the number of patrons, and the distance to the local market. Using this knowledge, George manages to eat out every day, never leaving disappointed.

George's story captures the essential characteristics of numerous widely encountered decision-making problems, where (a) individuals repeatedly face a choice between a large number of uncertain options, the value of which can be learned through experience, and (b) there are various cues such that they can form an expectation about the value of an option without having tried it previously. These two characteristics are related to two learning problems that have been explored extensively in psychology and cognitive science, yet mostly in isolation. These are how people learn to make decisions from experience (Barron & Erev, 2003; Hertwig et al., 2004) and how they learn to make predictions from multiple noisy cues (Nosofsky, 1984; Speekenbrink & Shanks, 2010). The structure of "decisions from experience" problems can be formally represented in a multi-armed bandit (MAB) framework (Sutton & Barto, 1998). MAB problems involve a fine balance between taking the action that is currently believed to be the most rewarding ("exploitation") and taking potentially less rewarding actions to gain knowledge about the expected rewards of other alternatives ("exploration"). MAB problems have proven to be a useful framework to study how people tackle this exploration–exploitation trade-off (e.g. Barron & Erev, 2003; Cohen et al., 2007; Speekenbrink & Konstantinidis, 2015; Steyvers et al., 2009).

Decision situations in real life typically contain more information than classic MAB problems, as alternatives usually have many features that are potentially related to their value. In other words, there is a function relating features of the alternatives to their value, and we assume people can learn this function. In our example, after enough visits to various restaurants, George has learned the function and with one look at the restaurant's features can estimate the quality of the food. Strictly speaking, feature information is not needed to make good decisions. People who try an alternative many, many times have no need to engage in function learning to estimate its value. However, function learning can be very useful. There might not be time to try out alternatives many times, especially when the number of alternatives is large or the choice sets change frequently. Also, choosing a previously untried alternative might cost the decision maker dearly. In such situations it becomes important to be able to appraise an alternative's worth without actually trying it.

More subtle questions arise in a MAB problem with function learning. For example, exploration choices can now be made with the goal to learn more about the function, not just to estimate the value of a particular alternative. Indeed, choosing an alternative that is believed to be particularly bad may improve one's knowledge of the function to such an extent that the future benefit of being able to better predict the value of alternatives outweighs the current loss.

Decision-making problems that include both function learning and direct experiential learning can be captured formally in the theoretical framework of contextual multi-armed bandits (CMABs). This paradigm has received a lot of attention recently in the domain of machine learning due to the numerous applications in autonomous machine decision making (e.g. Li et al., 2010; Agrawal & Goyal, 2012). Although the optimal decision policy for CMAB problems is generally intractable, several heuristic strategies, such as upper confidence bounding (Auer, 2003), have been developed to tackle the problem in a reasonable manner, balancing the search for new high-quality alternatives (exploration) and the use of the most promising alternative discovered so far (ex-

ploitation). While these algorithms give reasonable results in practice, they rely on extensive memory and processing capacity, and their performance is often evaluated under the assumption of an infinite or very distant time horizon, which makes their applicability to human decision making in these problems unclear.

In the present study we aimed to shed light on how human decision makers allocate decisions among alternatives in contexts involving both function learning and direct experiential learning. Although theoretically not necessary to make decisions in a given situation, because of its usefulness for generalizing to new situations we expect that people nevertheless engage in function learning. Moreover, we developed a new function-based reinforcement learning model that offers novel predictions on how people tackle the exploration–exploitation trade-off in CMAB problems. We tested these predictions in an experiment where people made choices between a relatively large number of alternatives. By showing only some participants informative cues to the value of the alternatives, we were able to assess the relative benefit of contextual information in decision making in MAB problems. In a later test phase, we also assessed how people generalize their contextual knowledge to decisions between new alternatives.

Methods

We investigated the influence of function learning on decision making in a stationary MAB task. There were three versions of the task: (1) a classic MAB task where feature values were not visually displayed (we refer to this as the *classic* condition), (2) a CMAB task where feature values were visible and participants were instructed that features might be useful for their choices (*explicit contextual* condition), and (3) a CMAB task where feature values were visible but participants were not informed about the relation between features and the value of alternatives (*implicit contextual* condition). The contextual conditions had an additional test phase with new alternatives, where we examined whether participants had learned the function and could use the acquired knowledge to make better choices when facing new alternatives.

Participants

In total, 193 participants (94 female), aged 18–73 years (M = 32.5 years, SD = 11.4), took part in this study on a voluntary basis. Participants were recruited via Amazon's Mechanical Turk (`mturk.com`) and were required to be based in the United States and have an approval rate of 95% or above.[1] Participants in the experiments earned a fixed payment of US$0.30 and a performance-dependent bonus of US$0.50 on average. Participants were randomly assigned to one of the three experimental groups: the classic ($N = 66$), explicit contextual ($N = 64$), and implicit contextual ($N = 63$) conditions.

As Amazon's Mechanical Turk is an online environment that offers less control than laboratory experiments, we excluded participants who did not pay due attention to the experimental task. At the end of the instructions, participants answered four questions to check whether they recalled basic information from the instructions. Excluding participants who failed to answer all four questions correctly would have left us with too small a sample, so we excluded participants who failed to answer at least two of these correctly. Importantly, this exclusion was done before we looked at further results. In total, 47 participants were excluded from the analysis.

[1] This means that in at least 95% of cases they were paid for the work they had done—a rough measure of the quality of the work done on Mechanical Turk.

Figure 1: Screenshots from the experiment. **A**. Alternatives in the classic multi-armed bandit (MAB) task were presented as simple red boxes without features. **B**. Alternatives in the CMAB tasks were presented as the same red boxes but now with lengths of horizontal and vertical yellow lines to represent features. Here we have illustrated only 2 alternatives; participants actually faced 20 in the training and 3 in the test phase.

Task

Training phase The task consisted of a training and a test phase. The training phase comprised 100 trials and in each trial participants were presented with the same 20 alternatives (bandit arms) and asked to choose one. After making a choice in trial t, they were informed of the payoff $R(t)$ associated with their choice. For each arm $j = 1, \ldots, 20$, the payoffs $R_j(t)$ on trial t were computed according to the following equation:

$$R_j(t) = w_1 x_{1,j} + w_2 x_{2,j} + \varepsilon_j(t).$$

The two feature values, $x_{1,j}$ and $x_{2,j}$, of each alternative j were drawn from a uniform distribution $U(0.1, 0.9)$, for each participant at the beginning of the training phase. Weights were set to $w_1 = 2$ and $w_2 = 1$ for all participants. The error term, $\varepsilon_j(t)$, was drawn randomly from a normal distribution $N(0,1)$, independently for each arm. The difference between conditions was that the feature values, $x_{1,j}$ and $x_{2,j}$, were visually displayed in the contextual conditions but not in the classic condition, as illustrated in Figure 1.

Test phase The structure of the task was similar in the test phase, but now participants were presented with three new alternatives with randomly drawn feature values on each trial. Weights of the function were kept the same, $\mathbf{w} = (2, 1)$. As participants faced a new decision problem on each trial in the test phase, there was no longer an exploration–exploitation trade-off, and participants were expected to always choose the alternative they deemed best. There were five types of trials, specifically designed so that participants would exhibit whether they had learned the functional form and the weights, w_1 and w_2. Two of the types were easy and difficult interpolation trials, where feature values were drawn from the same

interval, $U(0.1, 0.9)$, as in the training phase. Two others were easy and difficult extrapolation trials, with feature values drawn from $U(0, 0.1)$ and $U(0.9, 1)$. Trials consisted of a dominating, a middle, and a dominated alternative. In easy trials the difference in function values between the alternatives was larger than in the difficult trials. The fifth type of trial was designed so we could examine whether participants learned which feature had the greater weight. Here a trial consisted of one alternative that had a large value on a feature with higher weight and a small value on the other feature, one alternative with the opposite pattern, and one alternative that was clearly dominated. There were 70 trials in the test phase. Only participants in the contextual conditions completed this phase.

Procedure[2]

After providing informed consent, participants started the experiment by reading the instructions and completing a brief sociodemographic questionnaire, followed by comprehension questions with which we checked how much attention they paid to the instructions. Participants were told that they would be presented with 20 alternatives, that their task was to select between them, and that for each choice they would receive experimental points that would at the end be converted to money, with an exchange rate of US$1.00 for 400 experimental points. The goal of the game was to win as many experimental points as possible. Participants were informed that they would see the same alternatives in every round but that the rewards associated with each alternative might vary from round to round.

After reading the instructions and completing the questionnaires, participants started the experimental task. On each trial, they were presented with 20 alternatives in the form of simple square-shaped buttons. They selected an alternative via a mouse click. The number of points won or lost was then displayed below the alternative until they pressed the ENTER key, which would display the next trial. Buttons in the classic condition were empty, while in the contextual conditions feature values were displayed on each button in the form of one horizontal and one vertical line, both starting from the lower left corner of the square. We randomized whether a certain feature was represented as a vertical or a horizontal line across participants. Throughout the task, a counter displayed the total points received thus far, the number of the current trial, and the total number of trials in the phase. In the training phase, participants completed only a single MAB problem. After finishing it, participants in the contextual conditions read the instructions for the test phase. We told them they would face new alternatives in every trial, would not see any feedback in the second phase, and would no longer see the running total but that their payoff would still be affected by their choices.

[2]Readers can try out the experiment at the following URL: `experimentnext.com/CMABvsMABexp1`. Raw data from the experiments are also publicly available on Figshare: `http://dx.doi.org/10.6084/m9.figshare.1314099`

Behavioral Results

Training phase

Performance in the training phase is illustrated in Figure 2. Over the course of the training phase participants in both the classic and contextual MAB conditions were able to improve their performance by choosing more promising alternatives. This is evident in the downward slopes of linear fits of the average rankings of the chosen options as a function of trial. As the training phase progressed, participants discovered alternatives that yielded higher earnings on average, and the average ranking of the alternatives they had chosen decreased as a result. Although the increase in returns was similarly steep, the participants in the CMAB conditions had a head start and identified better alternatives already in the first rounds. This seems to have been the case especially for the explicit contextual condition where participants were instructed that the features could be used to improve their decisions. Such an increase early on may have been due to a strong prior expectation for positive linear relationships, as often found in the function learning literature (Busemeyer et al., 1997).

We analyzed choice performance with a generalized linear mixed-effects model. Trials were aggregated into four blocks of 25 trials each. We included experimental condition and block as fixed effects and subject-specific random intercepts. The main effect of condition was significant, $\chi^2(2) = 10.91$, $p = 0.004$, where differences stem from the classic condition, for which the intercept estimate was significantly higher, indicating worse performance overall. The main effect of block was also significant, $\chi^2(3) = 91.34$, $p < 0.001$, reflecting a general decrease in average ranking of selected alternatives from the first to the fourth block. Thus, participants learned to make better choices and the choice performance improved over time in all three conditions. The interaction between condition and block was not significant. The same conclusion was reached when we analysed expected earnings instead of rankings of the chosen alternative. We report the results for the rankings because, due to the random selection of feature values, potential earnings differed between participants.

To get a sense of the improvement made possible by the presence of features, it is instructive to examine the overall earnings. The range of possible expected earnings was from 0.3 to 2.7 experimental points per trial. Empirically, the lowest ranking arm had a value of 0.6 points on average, while the highest ranking alternative had the average value of 2.4 points. Participants in contextual conditions earned more on average per trial ($M = 1.8$ points, $SD = 0.52$, both contextual conditions combined) than participants in the classic condition ($M = 1.64$ points, $SD = 0.54$), $t(123.9) = 3.896$, $p < 0.001$, 95% CI [0.08, 0.23]. The possibility to use function learning enabled the participants to reach about 10% higher earnings.

In terms of exploration, participants in the contextual conditions tried 10.4 alternatives, and in the classic condition they tried 11.2 alternatives on average. For the remaining analyses we decided to pool the results for the two contextual

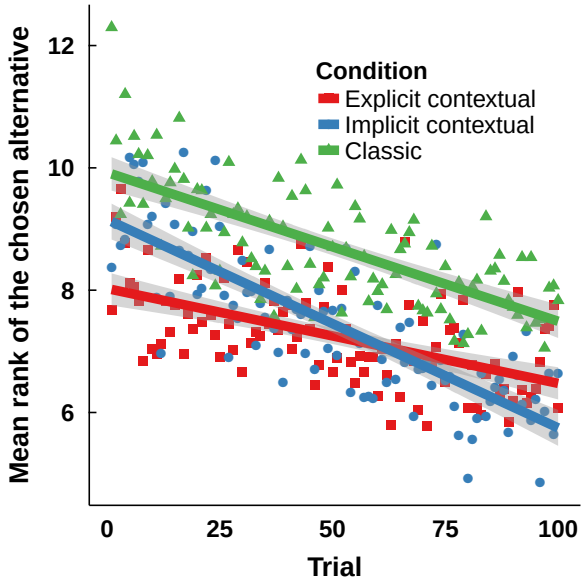

Figure 2: Average ranking of the chosen alternative in the training phase as a function of trial. Averages are across the participants and the lines were obtained by linear regression. Shaded areas are 95% confidence intervals.

conditions, since the performance in them was very similar.

Test phase

While behavior in the training phase showed evidence of function learning, the true test of function learning is performance on new, previously unseen items. If participants in the contextual conditions did not learn the function, we would expect that participants would choose randomly between the new sets of three alternatives. Results are presented in Table 1. On easy trials, participants selected the alternative with the higher expected value almost 50% of the time, while the middle and dominated alternatives were selected much less frequently (approximately 25% of the time each). On difficult trials, in contrast, participants selected the dominating and the middle alternative equally often, approximately 37% of the time. The dominated alternative was still selected approximately 25% of the time. Extrapolation trials are crucial for establishing the extent of function learning (Busemeyer et al., 1997). In our case, performance in interpolation and extrapolation trials was similar, indicating that participants extrapolated relatively well. Performance on the "weight test" trials gives a clue as to why they chose the middle alternatives as often as they did—on average, participants seem not to have learned the feature weights correctly, which may have been because of the level of noise in the alternative values (the error term $\varepsilon_j(t)$ in the function value). Participants seem to have learned that both feature weights were positive, but not that they differed.

Modeling

In addition to the behavioral results, we used computational modeling to further assess whether participants based their decisions on knowledge of the functional relationship between the feature values and the alternative value. To explain the behavior in CMAB problems, we developed a new reinforcement learning model based on function learning and pitted it against reward-only reinforcement learning models employed to explain behavior in MAB problems.

Table 1: Choice allocation between alternatives with high, medium, or low expected earnings in the test phase. Each row of the table corresponds to a different type of trial. The high, medium, and low columns refer to the dominating, middle, and dominated items, respectively.

Type of trial	# Trials	High	Medium	Low
Easy interpolation	15	0.47	0.28	0.25
Difficult interpolation	25	0.38	0.37	0.25
Easy extrapolation	10	0.49	0.27	0.24
Difficult extrapolation	10	0.39	0.35	0.26
Weights test	10	0.36	0.37	0.27

Reward-only reinforcement learning models do not take into account the feature values and update the expected value of an alternative only on the basis of rewards received after making a choice. We call this type of learning *mean learning*. In our novel feature-based model, a participant observes the feature values and uses the knowledge of the functional relationship between features and value to compute the expected value of a particular alternative. Instead of updating the expected value of an alternative directly, participants update the parameters of the functional relationship. We call this type of learning *function learning*. To provide the cleanest comparison, we used the same choice rules in both types of models. The main difference was in whether the expected values were computed by mean learning or function learning and then passed as inputs to the choice rules. Overall, we evaluated a factorial combination of 4 Learning rules (2 mean learning + 2 function learning) $\times$ 2 Choice probability rules, producing a total of eight models. The models were assessed in two ways. First we examined how the models fit the training data; second, we used the parameters from the training phase and let the models predict the choices in the test phase.

Mean learning

We assumed that after receiving a reward $R_j(t)$ on trial t for a chosen alternative j, participants would update the expected value $E_j(t+1)$ of choosing alternative j on trial $t+1$. We considered two learning mechanisms: the delta rule and the decay rule.

Delta learning The delta rule is a popular model-free learning rule:

$$E_j(t) = E_j(t-1) + \delta_j(t)\eta[R_j(t) - E_j(t-1)],$$

where $\delta_j(t)$ is an indicator variable, being 1 if alternative j was chosen on trial t, and 0 otherwise. We opted for a simple fixed learning rate, $\eta \geq 0$.

Decay learning The decay rule (e.g. Ahn et al., 2008) is another popular model-free learning rule, according to which

expected values of the unchosen alternatives decay toward 0:

$$E_j(t) = \eta E_j(t-1) + \delta_j(t) R_j(t),$$

with decay parameter $0 \leq \eta \leq 1$.

Function learning

Least mean squares network model The least mean squares (LMS) network model (e.g. Speekenbrink & Shanks, 2010) is essentially a linear regression model that updates the weights from trial to trial. Feature values $\mathbf{x}_j$ are inputs, and the expected value of an alternative is the function output, $E_j(t) = \mathbf{x}_j \hat{\mathbf{w}}(t)$, where $\hat{\mathbf{w}}(t)$ is a vector of estimated connection weights (identical for each alternative). Weights are updated through the delta rule

$$\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \delta_j(t)\eta(R_j(t) - E_j(t))\mathbf{x}_j^T,$$

where η is a vector of feature-specific learning rate parameters. Starting weights were initialized to $\hat{\mathbf{w}}(0) = (0,0)^T$. We considered two versions of the LMS model: one with an intercept (LMS_i) and without it (LMS).

Choice rules

ε-greedy A heuristic rule for balancing exploitation and exploration (e.g. Sutton & Barto, 1998) exploits the alternative with the maximum expected value with probability $1-\varepsilon$, and with probability ε chooses randomly from the remaining arms:

$$P(C(t) = j) = \begin{cases} (1-\varepsilon)/K_{\max} & \text{if } E_j(t) > E_k(t), \quad \forall k \neq j \\ \varepsilon/(K - K_{\max}) & \text{otherwise} \end{cases}$$

where K is the number of arms and $K_{\max}$ is the number of arms with the same maximum value. If all the values are the same, $P(C(t) = j) = 1/K$.

Softmax The "softmax" choice rule varies gradually between pure exploitation and pure exploration through a temperature parameter $\theta \geq 0$:

$$P(C(t) = j) = \frac{\exp(\theta E_j(t))}{\sum_{k=1}^{K} \exp(\theta E_k(t))}$$

Model estimation and inference

We estimated the model parameters for each participant by maximum likelihood using the Nelder–Mead simplex algorithm implemented in the `optim` function in R. For model selection purposes in the training phase, we computed the Bayesian information criterion (BIC), reported as difference scores between a baseline model[3] and the model of interest, Δ(BIC). For these difference scores, negative values of Δ(BIC) indicate that the model fitted worse than the baseline model, while increasing positive values indicate better fit. We also reported BIC weights, w(BIC) that approximate the posterior probability of the models assuming equal prior probability (Wagenmakers & Farrell, 2004). To model the behavior in the test phase, we used models with parameters estimated on the training data to predict choices in the test phase. For model selection we used Mean absolute deviation (MAD) of choices from model predictions.

[3]The baseline model was a parameter-free random choice model with probability of choosing an alternative equal to $1/K$.

Modeling results

We fitted the models to the training data of the contextual conditions. Because the behavioral and modeling results were similar for explicit and implicit contextual conditions, we collapsed the results into a single contextual condition. Table 2 shows the fit measures. The BIC scores show, contrary to our expectation, that the best fitting models are reward-only reinforcement learning models. This holds in terms of both average ΔBIC, BIC weights and number of participants best fitted by the models. Decay learning with the softmax choice rule is a clear winner. Participants often repeated their previous choices, and the decay rule is able to capture that tendency better (Ahn et al., 2008). Among function learning models, LMS_i version with the intercept is able to learn the average earning in the task so this model can be thought of as a hybrid between mean and function learning. However, LMS_i did not fare much better than the version without intercept. The softmax rule also worked much better than ε-greedy—people's response probabilities were obviously sensitive to expected values and the softmax choice rule captures this aspect better.

The modeling results thus contrast with the behavioral results, which showed evidence of function learning. One reason for this discrepancy might be that the LMS model is not the appropriate function learning model. Hence, in future work there is scope for examining more complex associative function learning models (Busemeyer et al., 1997) and generalized context models (Nosofsky, 1984). Another reason might be that the LMS model learns too well, that is, weights learned by the LMS model tend to the objective weights too fast. Participants were not giving their predictions on values of chosen alternatives and without them it is difficult to properly calibrate the function learning part of the model. Hence, obtained weights might not reflect participants' actual beliefs about feature weights. Indeed, results from the test phase, shown in Table 1, indicated that participants on average did not learn which feature had a larger weight.

Even though LMS models did not fit the training phase best, the true value of function learning should become obvious when new alternatives appear in the choice set. This was the logic behind having the test phase with new alternatives. Importantly, the reward-only models cannot predict anything other than random choice. Test trials were single-shot decisions and reward-only models have no means of estimating the expected values of arms without sampling them first. In the test phase, the only way to distinguish the alternatives was through their feature values. We used the LMS models with parameters fitted in the training phase to predict choices in the test phase. Table 3 shows that the LMS models indeed did better than the reward-only models, which here

Table 2: Modeling results of the training phase. Values of Δ(BIC) and *w*(BIC) are averages and the standard deviation is given in parentheses. Values of *N* are the total number of participants best fit by the corresponding model according to the BIC.

Learning	Choice	*N*	Δ(BIC)	*w*(BIC)
Decay	Softmax	42	127.56 (150.35)	0.36 (0.45)
Delta	Softmax	22	106.85 (139.47)	0.19 (0.36)
Decay	ε-greedy	12	100.86 (140.72)	0.06 (0.19)
Delta	ε-greedy	9	80.34 (120.52)	0.02 (0.08)
LMS_i	Softmax	11	56.73 (80.05)	0.11 (0.23)
LMS	Softmax	10	55.78 (80.51)	0.08 (0.18)
LMS	ε-greedy	8	19.49 (53.94)	0.03 (0.14)

Note. BIC, Bayesian information criterion; LMS_i, least mean squares with intercept; LMS, least mean squares without intercept.

would perform as well as the baseline random choice model. On average LMS models predicted the choices of the participants with 50%. According to the MAD criterion, the majority of participants were best predicted by one of the LMS softmax models, 13 participants were best predicted by a random choice model and 19 by ε-greedy LMS model.

Table 3: Modeling results of the test phase. Values of MAD are averages and the standard deviation is given in parentheses. Values of *N* are the total number of participants best fit by the corresponding model according to the MAD.

Learning	Choice	*N*	MAD
LMS_i	Softmax	54	0.50 (0.14)
LMS	Softmax	44	0.54 (0.13)
LMS	ε-greedy	19	0.65 (0.02)
RCM		13	0.67 (0)

Note. MAD, mean absolute deviation; LMS_i, least mean squares with intercept; LMS, least mean squares without intercept; RCM, random choice model.

Discussion and Conclusion

We developed a novel experimental paradigm that can be theoretically framed as a CMAB problem. In contextual conditions in our experiment each alternative had two features that were linearly related to the value of the alternative. In reward-only reinforcement learning models contextual information is ignored—mean returns are estimated directly from the sequence of past rewards without a demanding function-learning mechanism. We argued that in decision-making problems encountered in everyday life, people cannot afford to sample alternatives enough times to get reliable estimates. Moreover, choice sets change often, and estimating the value of new alternatives without trying them is a useful ability. Under these circumstances function learning seems to be an indispensable mechanism, even if unnecessary or prohibitively expensive in a single decision situation.

In the experiment we compared contextual and classic bandit problems, with feature information presented and not presented, respectively. We showed that with a large enough choice set, people engage in function learning—participants in the contextual conditions performed better even in the training phase. More importantly, function learning enabled them to generalize their knowledge in the test phase, where they faced one-shot trilemmas with new alternatives. We also developed a novel function-learning-based reinforcement learning model. Our simple model did not work as well as expected in the training phase, but it performed better in terms of predicting choices in the test phase where reward-based reinforcement learning models cannot do better than chance level. Other, more complex function-learning modeling approaches are left for future work.

Acknowledgments

We would like to thank Robin Hogarth and Gael Le Mens for comments and Doug Markant for practical advice with implementing the experiment on Amazon's Mechanical Turk.

References

Agrawal, S., & Goyal, N. (2012). Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*.

Ahn, W.-K., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, *32*(8), 1376–1402.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, *3*, 397–422.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215–233.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts and categories. studies in cognition.* (pp. 408–437). Cambridge, MA, US: MIT Press.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1481), 933–942.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 661–670).

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, 1–17.

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266–298.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

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

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.