Using Heuristic Models to Understand Human and Optimal Decision-Making on Bandit Problems

Michael D. Lee (mdlee@uci.edu)
Shunan Zhang (szhang@uci.edu)
Miles Munro (mmunro@uci.edu)
Mark Steyvers (msteyver@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Abstract

We study bandit problems in which a decision-maker gets reward-or-failure feedback when choosing repeatedly between two alternatives, with fixed but unknown reward rates, over a short sequence of trials. We collect data across a number of types of bandit problems to analyze five heuristics—four seminal heuristics from machine learning, and one new model we develop—as models of human and optimal decision-making. We find that the new heuristic, known as τ-switch, which assumes a latent *search* state is followed by a latent *stand* state to control decision-making on key trials, is best able to mimic optimal decision-making, and best account for the decision-making of the majority of our experimental participants. We show how these results allow human and optimal decision-making to be characterized and compared in simple, psychologically interpretable ways, and discuss some theoretical and practical implications.

Keywords: Bandit problems, heuristic models, reinforcement learning, human decision-making, optimal decision-making

Introduction

In Bandit problems, a decision-maker chooses repeatedly between a set of alternatives. They get feedback after every decision, either recording a reward or a failure. They also know that each alternative has some fixed, but unknown, probability of providing a reward each time it is chosen. The goal of the decision-maker is to obtain the maximum number of rewards over all the trials they complete. In some bandit problems, known as infinite horizon problems, the number of trials is not known in advance, but there is some probability any trial will be the last. In other bandit problems, known as finite horizon problems, the number of trials is fixed, known, and usually small.

Because bandit problems provide a simple task that addresses fundamental issues of learning and optimality in decision-making, they have been widely studied in the machine learning (e.g., Berry & Fristedt, 1985; Gittins, 1979; Kaebling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1988) and cognitive science (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Steyvers, Lee, & Wagenmakers, in press) literatures. In particular, bandit problems provide an interesting formal setting for studying the balance between exploration and exploitation in decision-making. In early trials, it makes sense to explore different alternatives, searching for those with the highest reward rates. In later trials, it makes sense to exploit those alternatives known to be good, by choosing them repeatedly. How exactly this balance between exploration and exploitation should be managed, and should be influenced by factors such as the distribution of reward rates, the total number of trials, and so on, raises basic questions about adaptation, planning, and learning in intelligent systems.

In this paper, we focus on finite-horizon bandit problems. We also restrict ourselves to the most basic, and most often considered, case where of there being only two alternatives to choose between. For this class of bandit problems, there is a well known optimal decision process that can be implemented using dynamic programming (see, for example Kaebling et al., 1996, p. 244). The basic approach is that, on the last trial, the alternative with the greatest expected reward should be chosen. On the second-last trial, the alternative that leads to the greatest expected total reward should be chosen, given that the last trial will be chosen optimally. By continuing backwards through the trial sequence in this way, it is possible to establish a recursive process that makes optimal decisions for the entire problem.

A motivating challenge for our work involves interpreting, evaluating and potentially improving human decision-making. Using the optimal benchmark, it is possible to evaluate how well a person solves bandit problems. The conclusion might be something like "you got 67% rewards, but optimal behavior would have given you 75% rewards, so you are falling short". This seems like only a partial evaluation, because it does not explain *why* their decisions were sub-optimal, and it is not clear how to relate the recursive algorithm to their data to provide this information.

Instead, to help us understand human and optimal decision-making on bandit problems, we evaluate a set of heuristic models. These include several heuristics from the existing machine learning literature, as well as a new one we develop. The attraction of the heuristic models is that they provide simple process accounts of how a decision-maker should behave, depending on a small set of parameters. We choose heuristic models whose parameters have clear and useful psychological interpretations. This means that, when we fit the models to data, and estimate the parameters, we obtain in-

terpretable measure of key aspects of decision-making. Instead of just telling people they are falling short of optimal, we now aim also to tell them "the problem seems to be you are exploring for too long: the optimal thing to do is to stop exploring at about the 5th trial", or "you are not shifting away quickly enough from a choice that is failing to reward you: the optimal thing to do is to leave a failed choice about 80% of the time."

The structure of this paper is as follows. First, we introduce the five heuristics used in this study. We then evaluate their ability to mimic optimal decision-making, and their ability to fit human data we collected for this study. Having found some heuristics that are able to describe human and optimal behavior, we finish by discussing the psychological characteristics of optimal behavior in bandit problems, and the properties of human decision-making we observed.

Five Heuristics

Win-Stay Lose-Shift

Perhaps the simplest reasonable heuristic for making bandit problem decisions is the Win-Stay Lose-Shift (WSLS) heuristic. In its deterministic form, it assumes that the decision-maker continues to choose an alternative following a reward, but shifts to the other alternative following a failure to reward. In the stochastic form we use, the probability of staying after winning, and the probability of shifting after losing, are both parameterized by the same probability γ.

Psychologically, the win-stay lose-shift heuristic does not require a memory, because its decisions only depend on the presence or absence of a reward on the previous trial. Nor is the heuristic sensitive to the horizon (i.e., the finite number of trials) in the bandit problem version we consider, because its decision process is the same for all trials.

ε-Greedy

The ε-greedy heuristic is a standard approach coming from reinforcement learning. It assumes that decision-making is driven by a parameter ε that controls the balance between exploration and exploitation. On each trial, with probability $1-\varepsilon$ the decision-maker chooses the alternative with the greatest estimated reward rate (i.e., the greatest proportion of rewards obtained for previous trials where the alternative was chosen). This can be conceived as an 'exploitation' decision. With probability ε, the decision-maker chooses randomly. This can be conceived as an 'exploration' decision.

Psychologically, the ε-greedy heuristic does require a limited form of memory, because it has to remember counts of previous successes and failures for each alternative. It is not, however, sensitive to the horizon, and uses the same decision process on all trials.

ε-Decreasing

The ε-decreasing heuristic is a variant of the ε-greedy heuristic, in which the probability of an exploration move decreases as trials progress. In its most common form, which we use, the ε-decreasing heuristic starts with an exploration probability ε_0 on the first trial, and then uses an exploration probability of ε_0/i on the ith trial. In all other respects, the ε-decreasing heuristic is identical to the ε-greedy heuristic.

This means the ε-decreasing heuristic does more exploration on early trials, and focuses on its estimate of expected reward more on later trials. Psychologically, the innovation of the ε-decreasing heuristic means it is sensitive to the horizon, making different decisions over different trials.

π-First

The π-first heuristic is usually called the ε-first heuristic in the literature. It is, however, quite different from the ε-decreasing and ε-greedy heuristics, and we emphasize this with the different name. The π-first heuristic assumes two distinct stages in decision-making. In the first stage, choices are made randomly. In the second stage, the alternative with the greatest currently observed reward rate is chosen. The first stage can be conceived as 'exploration' and the second stage as 'exploitation'. In our implementation, a discrete parameter π determines the number of exploration trials, so that the π-th trial marks the last trial of exploration.

Psychologically, the π-first requires both the memory of previous successes and failures needed in the exploration stage, and has a clear sensitivity to the horizon. The notion of two decision-making stages is a psychologically plausible and interesting approach to capturing how a decision-making might balance the tradeoff between exploration and exploitation.

τ-Switch

The τ-switch is a new heuristic, motivated by the idea of latent decision-making stages used by the π-first heuristic. The τ-switch heuristic also assumes an initial 'search' stage, followed by a later 'stand' stage. The trial number at which the change in stages takes place is determined by the parameter τ, similarly to the role of the parameter π. The different decision-making strategies employed in each stage in the τ-switch heuristic, however, rely on an analysis of different possible states in bandit problems.

Figure 1 provides a graphical representation of three possible cases. In Case I, both alternatives have the same reward history. The τ-switch heuristic assumes both alternatives are chosen with equal probability when confronted with this state. In Case II, one alternative has more successes and the same or fewer failures than the other alternative (or, symmetrically, it has fewer failures and the same or more successes). This means one alternative is clearly 'better', because it dominates the other in terms of successes and failures. The τ-switch heuristic assumes the 'better' alternative with (high) probability γ.

The crucial situation is Case III, in which one alternative has more successes but also more failures, when compared to the other alternative. This means neither

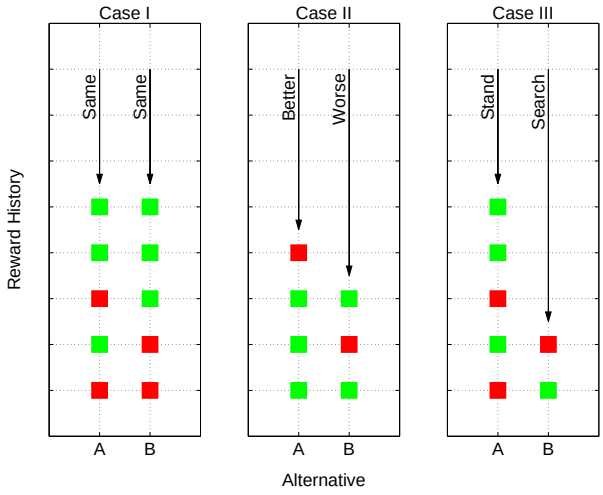

Figure 1: The three different possible cases for a bandit problem considered by the τ-switch heuristic. Green (lighter) squares correspond to previous rewards, while red (darker) squares correspond to previous failures.

alternative can clearly to be preferred. Instead, the alternative chosen more often previously can be conceived as a 'stand' choice, because it is relatively well known. The alternative chosen less often can be conceived as an 'search' choice, because it is relatively unknown. The τ-switch assumes that, faced with an observed State III, the decision-maker chooses the 'search' alternative when they are in the initial latent 'search' stage, with the same (high) probability γ. But, the decision-maker is assumed to choose the 'stand' alternative once they have switched to the latent 'stand' stage.

Psychologically, the τ-switch heuristic has the same memory requirements as the ε-greedy, ε-first and π-first heuristics. The τ-switch heuristic also takes into account the horizon, using the same latent stage approach as the π-first heuristic. It is the detail of the decisions it makes, depending on how its internal state relates to the state of reward history observed, that makes the τ-switch heuristic new and interesting.

Human and Optimal Decision Data

Subjects Data were collected from 10 naive participants (6 males, 4 females).

Stimuli There were six different types of bandit problems, all involving just two alternatives. These six conditions varied two trial sizes (8 trials and 16 trials) and three different environmental distributions ('plentiful', 'neutral' and 'scarce'). Following Steyvers et al. (in press), the environments were defined in terms of Beta(α,β) distributions, where α corresponds to a count of 'prior successes' and β to a count of 'prior failures'. The plentiful, neutral and scarce environments used, respectively, the values $\alpha=4$, $\beta=2$, $\alpha=\beta=1$, and $\alpha=2$, $\beta=4$. Within each condition, the reward rates for each alternative in each problem were sampled independently from the appropriate environmental distribution.

Procedure Within-participant data were collect for 50 problems for all six bandit problem conditions, using a slight variant of the experimental interface shown in Steyvers et al. (in press). The order of the conditions, and of the problems within the conditions, was randomized for each participant. All $6\times 50=300$ problems (as well as 5 practice problems per condition) were completed in a single experimental session, with breaks taken between conditions.

Optimal Decision Data We generated decision data for the optimal decision-process on each problem completed by each participant. In generating these optimal decisions, we used the true α and β values for the environment distribution. Obviously, this gives the optimal decision process an advantage, because participants must learn the properties of the reward environment. However, our primary focus is not on measuring people's shortcomings as decision-makers, but in characterizing what people do when making bandit problem decisions, and comparing this to the best possible decision. From this perspective, it makes sense to use an optimal decision process with environmental knowledge. It would also be interesting, in future work, to develop and use an optimal decision process that optimally *learns* the properties of its environment.

Analysis With Heuristic Models

We implemented all five heuristic models as probabilistic graphical models using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). All of our analyses are based on 1,000 posterior samples, collected after a burn-in of 100 samples, and using multiple chains to assess convergence using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997).

Characterization of Optimal Decision-Making

We applied the heuristics to behavior generated by the optimal decision process. Table 1 shows the expected value of the inferred posterior distribution for the key parameter in each heuristic model (we observed all of the 'accuracy of execution' parameters were close to 1, as expected). These key parameter values constitute single numbers that characterize optimal decision-making within the constraints of each heuristic. They are shown for each of the plentiful, neutral and scarce environments for both 8 and 16 trial problems.

For WSLS, the parameter values shown in Table 1 correspond to the optimal rate at which a decision-maker should stay if they are rewarded, and shift if they are not. The patterns across environments and trial sizes are intuitively sensible, being higher in more plentiful environments and for shorter trial sizes.

For ε-greedy probability of choosing the most rewarding alternative is high, and very similar for all environments and trial sizes. For ε-decreasing, the starting prob-

Table 1: Expected posterior values for the key parameter in each heuristic model, based on inferences from optimal decision-making, for plentiful, neutral and scarce environments, and 8 and 16 trial problems.

	Plentiful		Neutral		Scarce	
Heuristic	8	16	8	16	8	16
WSLS (γ)	.87	.85	.85	.78	.72	.65
Greedy (ε)	.09	.07	.05	.05	.06	.07
Decrease (ε_0)	.62	.76	.57	.75	.56	.63
First (π)	1.0	1.0	1.0	1.0	1.0	1.0
Switch (τ)	5.1	7.0	4.1	5.0	2.0	2.0

ability of random exploration ε_0, which decreases as trials progress, is higher for more rewarding environments, and also for problems with more trials.

The π-first parameter is the trial at which the switch from random exploration to choosing the most rewarding alternative. This is always the first trial in Table 1, which is essentially a degenerate result. We interpret this as suggesting not that the notion of an exploration followed by an exploitation stage is ineffective, but rather that initial random decisions in a problem with few trials is so sub-optimal that it needs to be minimized.

Finally, the results for the τ-switch heuristic detail the optimal trial to switch moving from 'standing' to 'searching' in the Case III scenario described in Figure 1. This optimal switching trial becomes earlier in a problem as the environment becomes less rewarding, which makes sense. More plentiful environments should be searched more thoroughly for high yielding alternatives. The number of searching trials generally extends moving from 8 to 16 trial problems, but not by much. This also makes sense, since in the fixed environments we consider, longer sequences of exploitation will give many rewards, as long as sufficient exploratory search has been conducted.

All of these optimal parameter settings make sense, and demonstrate how a heuristic can give a straightforward psychology characterization of optimal decision-making for bandit problems. For example, in a neutral environment with 8-trial problems, an optimal decision-maker constrained in their cognitive processing capabilities to applying WSLS should win-and-stay or lose-and-shift 85% of the time. Alternatively, a more cognitive elaborate decision-maker, able to apply the two-stage τ-shift heuristic, should switch from searching to standing after the fourth trial.

How Optimal Are the Heuristics?

Of course, knowing what constitutes optimal behavior within the bounds of a heuristic does not take into account how well decisions will match unboundedly optimal decision-making.

To analyze this aspect of the heuristics' performance,

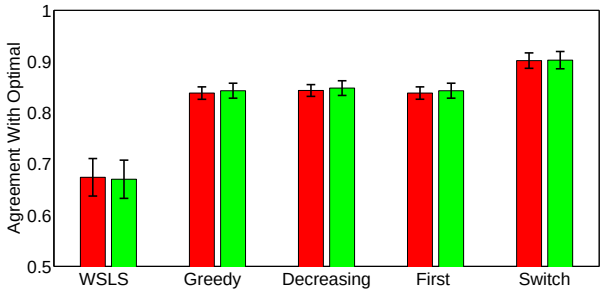

Figure 2: Posterior predictive average agreement of the heuristic models with the optimal decision process for 40 training problems (red, darker) and 10 test problems (green, lighter).

Figure 2 shows the posterior predictive average agreement of the heuristic models with the optimal decision process. The red bars correspond to a training set of the first 40 problems seen by all participants in which the parameters of the heuristic models were inferred by observing the optimal decisions. The green bars correspond to a test set of the final 10 problems seen by all participants, where the inferred parameters for the heuristic models were directly applied with observing the optimal decisions. The relative results between the heuristics are consistent over environments and trial sizes, and so are averaged to give a simple and general conclusion, but include error bars showing one standard error caused by the averaging.

It is clear that training and test performance are very similar for all of the heuristics. This is because the agreement is measured by a complete posterior predictive, which averages across the posterior distribution of the parameters. This means the measure of agreement—unlike measures of fit based on optimized point-estimates for parameters—automatically controls for model complexity. Thus, it is not surprising test performance is essentially the same as training performance.

Most importantly, Figure 2 shows that the WSLS heuristic is not able to mimic optimal decision-making very well, that the ε-greedy, ε-decreasing and π-first heuristics are able to do much better, and that the new τ-switch heuristic is clearly is the best performed.

Heuristics Modeling of Human Performance

We now apply the heuristics to the human data, and explore their ability to account for the way people solve bandit problems. Figure 2 shows the posterior predictive average agreement of the heuristic models with the human decisions. As before, the red bars correspond to a training set of the first 40 problems completed by each participant, and were used to infer posterior parameter distributions for each heuristic. The green bars correspond to agreement on the test set of the final 10 problems, integrating over the already inferred posterior distributions, and without knowing the participants' behav-

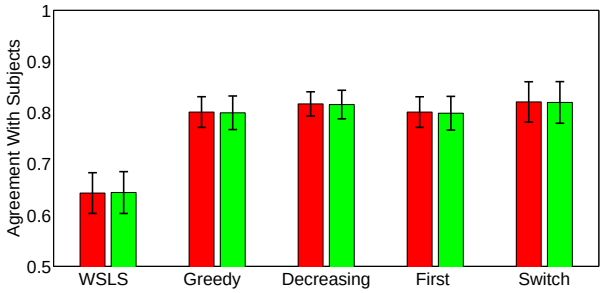

Figure 3: Posterior predictive average agreement of the heuristic models with human decision-making for 40 training problems (red, darker) and 10 test problems (green, lighter).

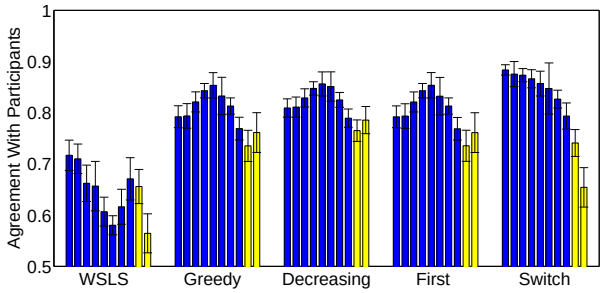

Figure 4: Posterior predictive average agreement of the heuristic models with each individual participant. Two 'outlier' participants, not modeled well by any of the heuristics, are highlighted in yellow (lighter).

ior on the test problems.

Figure 2 shows the ability of the heuristics to model human decision-making follows the same ordering as their ability to mimic optimal decision-making. WSLS is the worst, followed by the three reinforcement learning heuristics, which are approximately the same, and then slightly (although not significantly) improved by the new τ-first heuristic.

Figure 4 examines the ability of the heuristics to account for human decision-making at the level of the individual participants. Each participant is shown as a bar against each of the heuristics. For the first 8 of the 10 participants shown (in blue), the overall pattern seen in Figure 3, holds at the individual level. That is, the τ-switch heuristic provides the greatest level of agreement. For the last 2 of the 10 participants shown (in yellow), this result is not observed, but it is clear that none of the heuristics is able to model these participants well at all. We speculate that these participants may have changed decision-making strategies significantly often to prevent any single simple heuristic from providing a good account of their performance.

In any case, our results show that, for the large majority of participants well described by any heuristic, the τ-switch heuristic is the best. And the complexity control offered by the posterior predictive measure, and verified by the training and test sets, shows that this conclusion takes into account the different model complexity of the heuristics.

Characterization of Human Decision-Making

The analysis in Figure 2 shows the τ-switch heuristic can closely emulate optimal decision-making for bandit problems, and the analysis in Figure 4 shows it can also describe most participants' behavior well. Taken together, these results let us use the τ-switch heuristic to realize our original motivating goal of comparing people's decisions to optimal decisions in psychologically meaningful ways. The key psychological parameters of a well-performed heuristic like τ-switch provide a measure that relates people to optimality.

Figure 5 gives a concrete example of this approach. Each panel corresponds to one of the 8 participants from Figure 4 who were well modeled by the τ-switch heuristic. Within each panel, the large green curves show the switch trial (i.e., the expected posterior value of the parameter τ) inferred from optimal decision-making. These optimal parameter values are shown for each of the plentiful, neutral and scarce environments, for both 8- and 16-trial problems. Overlayed in each panel, using smaller black curves, are the patterns of change in this parameter for the individual participants.

The commensurability of the switch point parameter between people and optimality, and its ease of interpretation, allows for insightful analyses of each participant's performance. Participants like B and F are choosing near optimally, especially in the 8-trial problems, and seem sensitive to the reward rates of the environments in the right ways. Their deviations from optimality seem more a matter of 'fine tuning' exactly how early or late they switch away from exploratory search behavior. Participants like A and D, in contrast, are reacting to the changes in environment in qualitatively inappropriate ways. Participants like C, E, and H seem to perform better on the 8- than the 16-trial problems, and do not seem to be adjusting to the different environments in the 16-trial case. But C is switching at roughly the optimal trial on average, while E is switching too early, and H is too early for the shorter problems and too late for the longer ones. Finally, participant G seems to be employing a 'degenerate' version of the τ-switch heuristic that involves no initial search, but simply stands on the highest success rate alternative throughout the problem.

This analysis is not intended to be complete or exact. Potentially, the other heuristics could provide alternative characterizations with some level of justification. What the sketched analysis does provide a concrete illustration of the way human and optimal performance can be characterized by parametric variation using our best-fitting heuristic model.

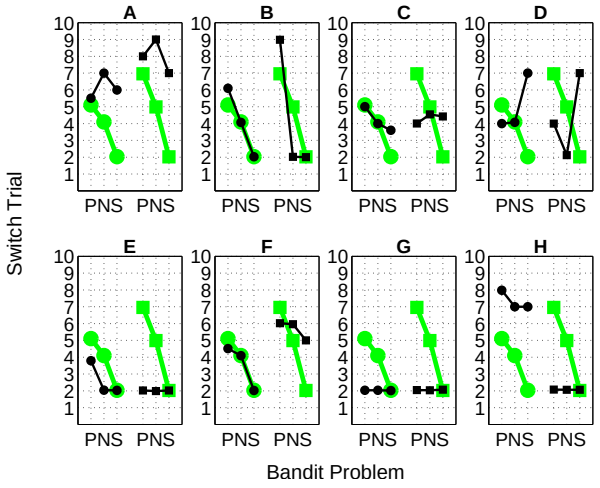

Figure 5: Relationship between the optimal switching point under the τ-first heuristic in (larger, green markers) and inferred switch points for 8 subjects A–H in (smaller, black markers). Comparisons are shown for P=plentiful, N=neutral and S=scarce environments, and 8-trial (circle) and 16-trial (square) environments.

Discussion

One finding from our results is that the τ-switch heuristic is a useful addition to current models of finite-horizon two-arm bandit problem decision-making. Across the three environments and two trial sizes we studied, it consistently proved better able to mimic optimal decision-making than classic rivals from the statistics and machine learning literatures. It also provided a good account of human decision-making, for the majority of the participants in our study.

A potential theoretical implication of the success of the τ-switch heuristic is that people may use latent states to control their search behavior, and manage the exploration versus exploitation trade-off. We think these sorts of models deserve as much attention as those, like ε-greedy, based more directly on reinforcement learning.

One potential practical application of the τ-switch heuristic is to any real-world problem where a short series of decisions have to made be made with limited feedback, and with limited computational resources. The τ-switch heuristic is extremely simple to implement and fast to compute, and may be a useful surrogate for the optimal recursive decision process in some niche applications. A second, quite different, potential practical application, relates to training. The ability to interpret optimal and human decision-making using one or two psychologically meaningful parameters could help instruction in training people to make better decisions. It would be an interesting topic of future research to take the sorts of analysis accompanying Figure 5, for example, and see whether feedback along these lines could improve their decision-making on future bandit problems.

More generally, we think our results illustrate a useful general approach to studying decision-making with heuristic models. Three basic challenges in studying any real-world decision-making problem are to characterize how people solve the problem, characterize the optimal approach to solving the problem, and then characterize the relationship between the human and optimal approach. Our results show how simple heuristic models, using psychologically interpretable decision processes, and based on psychologically interpretable parameters, can aid in all three of these challenges. While our specific results are for short-horizon two-alternative bandit problems, and involve a small set of heuristics, we think the basic approach has more general applicability. We think heuristic models, and their inferred parameter values, are useful for understanding and comparing human and optimal decision-making.

Acknowledgments

This work was funded by award FA9550-07-1-0082 from the Air Force Office of Scientific Research.

References

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments* . London: Chapman & Hall.

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

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

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, *41*, 148–177.

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* , *4*, 237–285.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* , *10*, 325–337.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on evolutionary computation*, *2*(1), 2-22.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (in press). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*.

Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: An introduction* . Cambridge, MA: MIT Press.