Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 12 (2011) 164–174
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Psychological models of human and optimal performance in bandit problems

Action editor: Andrew Howes

Michael D. Lee *, Shunan Zhang, Miles Munro, Mark Steyvers

Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA

Received 7 December 2009; accepted 9 July 2010
Available online 11 August 2010

Abstract

In bandit problems, a decision-maker must choose between a set of alternatives, each of which has a fixed but unknown rate of reward, to maximize their total number of rewards over a sequence of trials. Performing well in these problems requires balancing the need to search for highly-rewarding alternatives, with the need to capitalize on those alternatives already known to be reasonably good. Consistent with this motivation, we develop a new psychological model that relies on switching between latent *exploration* and *exploitation* states. We test the model over a range of two-alternative bandit problems, against both human and optimal decision-making data, comparing it to benchmark models from the reinforcement learning literature. By making inferences about the latent states from optimal decision-making behavior, we characterize how people should switch between exploration and exploitation. By making inferences from human data, we begin to characterize how people actually do switch. We discuss the implications of these findings for understanding and measuring the competing demands of exploration and exploitation in sequential decision-making.

Keywords: Bandit problem; Exploration versus exploitation; Heuristic models; Latent state models; Reinforcement learning

1. Introduction

1.1. Bandit problems

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, the number of trials is not known in advance, but there is some probability any trial will be the last. These are known as 'infinite horizon' bandit problems. In other bandit problems the number of trials is fixed, known, and usually small. These are known as 'finite-horizon' bandit problems.

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. 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. For these reasons, bandit problems have been widely studied in machine learning (Berry & Fristedt, 1985; Gittins, 1979; Kaebling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1998) and cognitive science (Cohen, McClure, & Yu, 2007; Daw, O'Doherty,

* Corresponding author.
E-mail address: mdlee@uci.edu (M.D. Lee).

1389-0417/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2010.07.007

Dayan, Seymour, & Dolan, 2006; Steyvers, Lee, & Wagenmakers, 2009), and many models of decision-making strategies have been proposed.

1.2. Research goals

A first motivation for our work is to refine and extend one existing theoretical idea that seems especially relevant to understanding human decision-making on bandit problems. This is the idea of latent state modeling, in which behavior is treated as a mixture of different processes, controlled by unobserved states. Latent state models are well suited to situations, where two or more qualitatively different types of decision-making are needed to explain performance as a whole. The general latent state approach has been successful in many areas of the cognitive sciences, ranging from all-or-none theories of learning (Batchelder, 1970), to models of language (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), to models of the roles of guessing and other contaminant behavior in simple decision-making (Vandekerckhove & Tuerlinckx, 2007). The latent state approach seems a particularly natural account of human decision-making in bandit problems, given the requirement to choose between the competing, and qualitatively different, demands of exploration and exploitation.

A second motivating challenge for our work involves interpreting, evaluating and potentially improving human decision-making. Using the optimal decision process (Kaebling et al., 1996), 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. Instead, to help us understand human and optimal decision-making on bandit problems, we use simple heuristic models. These include several benchmark models from the existing machine learning literature, as well the new latent state model we develop. The attraction of these models is that they provide simple process accounts of how a decision-maker should behave, depending on a small set of parameters. We choose 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 interpretable measures 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."

1.3. Overview

With these motivations in place, the outline of this paper is as follows. First, we describe an experiment in which human and optimal decision-making data for a variety of bandit problems was collected. We then describe four existing benchmark heuristics, before developing our new model. We test all of these models as accounts of the human and optimal decision data, using Bayesian methods that balance both goodness-of-fit and model complexity in model evaluation, and find that our new model performs better than the existing ones. Finally, we demonstrate how the psychological interpretability of the heuristics can help characterize and compare human and optimal decision-making on bandit problems.

2. Human and optimal decision data

2.1. Bandit problem conditions

We considered six different types of bandit problems, all involving just two-alternatives, which is the most commonly studied case in the literature, and all having short fixed horizons. The six conditions varied in a 2×3 design, manipulating how many trials there were in a problem, and how the reward rates for the alternatives were chosen. Specifically, there were two trial sizes (8-trial and 16-trial), and three different environmental distributions ('plentiful', 'neutral' and 'scarce') controlling the reward rates.

The basic idea of environmental distributions is to manipulate whether reward rates tend to have high or low values. Following (Steyvers et al., 2009), 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 $\alpha = 2$, $\beta = 4$. Reward rates for each alternative in each problem were sampled independently, for a total of 50 problems in each condition, from the appropriate environmental distribution.

2.2. Human data

Data were collected from 10 naive participants (6 males, 4 females). A representation of the basic experimental interface is shown in Fig. 1. The two large panels correspond to the alternatives, either of which can be chosen on any trial by pressing the button below. Within the panel, the outcomes of previous choices are shown as count bars, with successes on the left, and failures on the right. At the top of each panel, the proportion of successes, if defined, is shown. The top of the interface provides the success count, the current trial number, the total number of trials, and a count of how many problems out of the entire set have been completed.

Using this interface, within-participant data were collect for all 50 problems for all six bandit problem conditions. 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 five practice problems per condition) were completed in a single experimental session, with breaks taken between conditions.

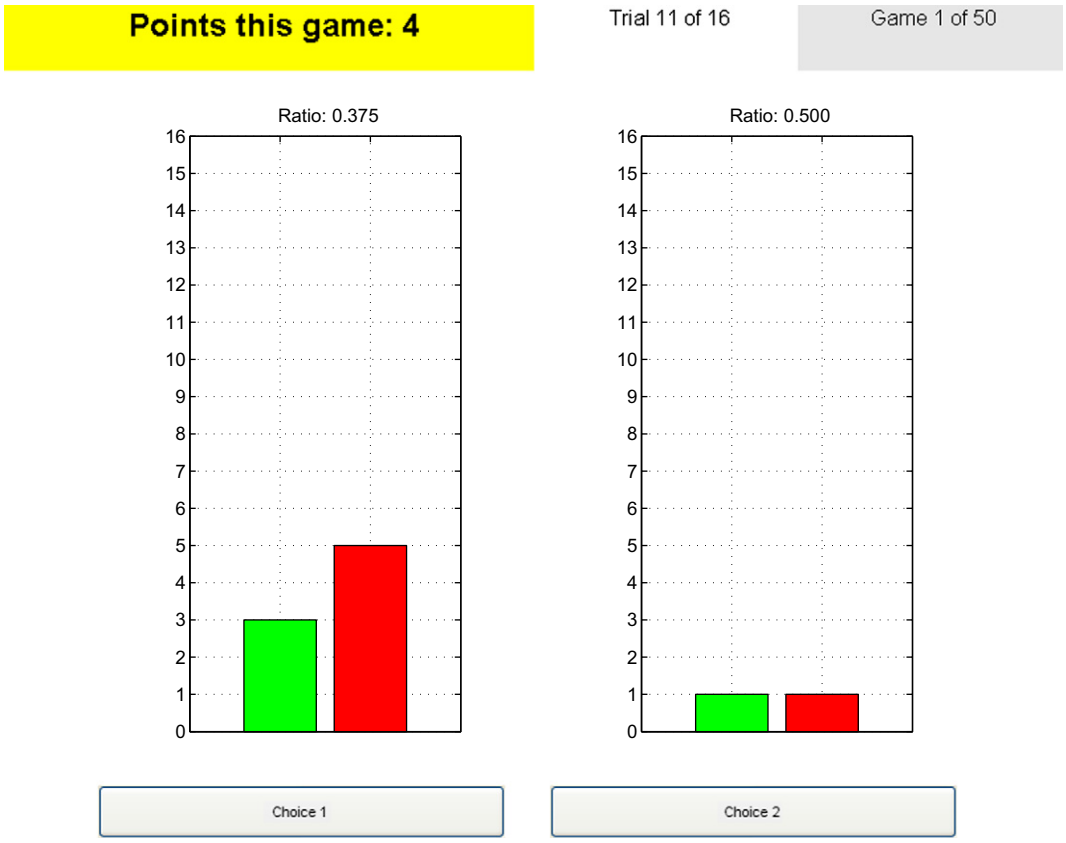

Fig. 1. The experimental interface for a sample bandit problem, with two-alternatives and 16 total trials. After 10-trials, the first alternative on the left has two successes (lighter, green bar) and five failures (darker, red bar), while the alternative on the right has one success and one failure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Optimal data

For finite-horizon bandit problems, there is a well known optimal decision process, able to be implemented using dynamic programming (Kaebling et al., 1996). Intuitively, this optimal approach recognizes 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 entire problem.

We generated decision data using 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-maker an advantage, because participants have to learn the properties of the reward environment. However, our primary focus is not on measuring people's shortcomings as decision-makers, but rather 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 perfect environmental knowledge. It would obviously also be interesting to develop and use an optimal decision process that optimally learns the properties of its environment.

3. Four benchmark models

In this section, we describe four prominent models of bandit problem decision-making, to serve as benchmark models for human and optimal decision-making.

3.1. Win-stay lose-shift

Perhaps the simplest model for making bandit problem decisions is the Win-Stay Lose-Shift (WSLS) heuristic (Sutton & Barto, 1998). 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.

3.2. ϵ-Greedy

The ϵ-greedy heuristic is a standard approach coming from reinforcement learning (Sutton & Barto, 1998). It assumes that decision-making is driven by a parameter ϵ that controls the balance between exploration and exploitation. On each trial, with probability $1-\epsilon$ 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.

3.3. ϵ-Decreasing

The ϵ-decreasing heuristic is a variant of the ϵ-greedy heuristic, in which the probability of an exploration decision decreases as trials progress (Sutton & Barto, 1998). 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.

3.4. π-First

The π-first model is usually called the ϵ-first model in the literature (Sutton & Barto, 1998). It is, however, quite different from the ϵ-decreasing and ϵ-greedy models, and we emphasize this with the different name. The π-first model assumes two distinct stages in decision-making. In the first stage, choices are made randomly. In the second stage, the alternative with the greatest estimated reward rate is always 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 πth trial marks the last trial of exploration. In addition, we include a parameter γ that we call the 'accuracy of execution' to make the exploitation behavior probabilistic. Formally, π-first chooses the alternative with greatest estimated reward with probability γ in the second stage. Intuitively, γ quantifies how perfectly the deterministic decision rule is followed in producing behavior, allowing for occasional lapses caused by factors outside the core cognitive decision processes of interest.

Psychologically, the π-first model 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.

4. A new latent state model

In this section, we develop a new model of bandit problem decision-making, motivated by the idea of latent states used by the π-first model. The development comes in two parts. We first implement and evaluate a 'full' latent state model, that allows for switching between exploration and exploitation at any trial in a bandit problem. We show, however, by applying this model to the human and optimal decision data, that it is possible to simplify the model significantly. Accordingly, we finish this section defining a simple latent state model that can subsequently be compared to the simple benchmark models already described.

4.1. A full latent state model

In the full model we allow each trial to have a latent state, introducing the possibility of switching flexibly between exploration and exploitation to solve bandit problems. It is possible, for example, to begin by exploring, then exploit, and then return for an additional period of exploration before finishing by exploiting. Indeed, any pattern of exploration and exploitation, changing trial-by-trial if appropriate, is possible. This is an obvious extension of the single switch assumption in the π-first model.

The second difference between our latent state model and π-first is that we implement exploration and exploitation behavior using a more subtle mechanism. Recall that π-first model just uses random search followed by deterministic responding. Our model, in contrast, controls exploration and exploitation by distinguishing between three different situations.

These situations are explained by the example in Fig. 2. In the *same* situation, both alternatives have the same number of observed successes and failures. In the *better–worse* situation, one alternative has more successes and fewer failures than the other alternative (or more successes and equal failures, or equal successes and fewer failures). In this situation, one alternative is clearly better than the other. In the *explore–exploit* situation, one alternative has been chosen more often, and has more successes but also more failures than the other alternative. In this situation, neither alternative is clearly better, and the decision-maker faces the explore–exploit dilemma. Choosing the better-understood

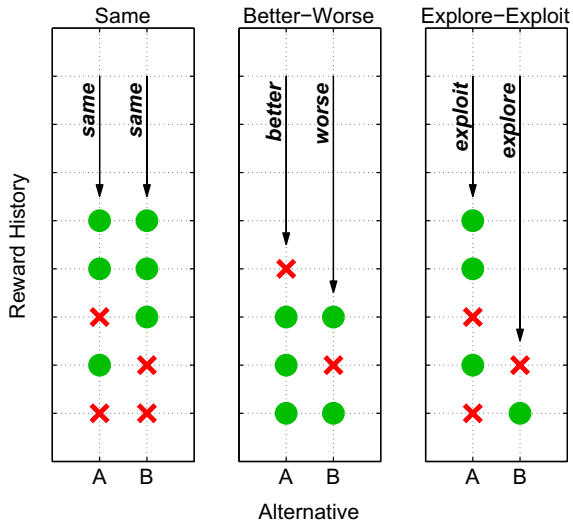

Fig. 2. The three different possible cases for a bandit problem considered by the new latent state model. Green (lighter) circles correspond to previous rewards, while red (darker) crosses correspond to previous failures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

alternative corresponds to exploiting, while choosing the less well-understood alternative corresponds to exploring.[1]

Within our model, which alternative is chosen depends on the situation, as well as the latent exploration or exploitation state. For the *same* situation, both alternatives have an equal probability of being chosen. For the *better–worse* situation, the better alternative has a high probability, given by a parameter γ, of being chosen. The probability the worse alternative is chosen is $1-\gamma$. The crucial situation is the third one shown in Fig. 2, which we call the *explore–exploit* situation. In this situation, our model assumes the exploration alternative will be chosen with the high probability γ if the decision-maker is in a latent 'explore' state, but the exploitation alternative will be chosen with probability γ if the decision-maker is in the latent exploit state. In this way, the latent state for a trial controls how the exploration versus exploitation dilemma is solved at that stage of the problem.

Overall, therefore, our new model has a binary latent state parameter for each trial in a bandit problem, and a single accuracy of execution parameter. The model operates by examining, at each trial, the previous history of rewards and failures for both alternatives. Using this information, it determines whether the current decision is being made in a same, better-worse, or explore–exploit situation. It then executes the decision-making strategy appropriate to the situation, as just described, using the latent state and accuracy of execution parameters as necessary.

4.2. Analysis of the full latent state model

We implemented the full latent state model as a probabilistic graphical model in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which makes it easy to do fully Bayesian inference using computational methods based on posterior sampling. Examples and tutorials for implementing cognitive models as graphical models are provided by Lee (2008) and Shiffrin et al. (2008), and details of implementing this specific model are provided in Zhang, Lee, and Munro (2009).

The key result coming from the modeling analysis involves the pattern of change between latent exploration and exploitation states over the trials, which are summarized in Fig. 3. This figure shows whether the model is an exploration or exploitation state as it accounts for both the human and optimal data, over all six experimental conditions. The experimental conditions are organized into the panels, with rows corresponding the plentiful, neutral and scarce environments, and the columns corresponding to the 8- and 16-trials problems. Each bar graph shows the probability of an exploitation state for each trial, beginning at the third trial (since it is not possible to encounter the explore–exploit situation until at least two choices have been made). The larger bar graph, with darker blue bars, in each panel is for the optimal decision-making data. The 10 smaller bar graphs, with lighter green bars, corresponds to the 10 subjects within that condition.

The most striking feature of the pattern of results in Fig. 3 is that, to a good approximation, once the optimal or human decision-maker first switches from exploration to exploitation, they do not switch back. There are some exceptions—both participants RW and BM, for example, sometimes switch from exploitation back to exploration briefly, before returning to exploitation—but, overall, there is remarkable consistency. Most participants, in most conditions, begin with complete exploration, and transition at a single trial to complete exploitation, which they maintain for all of the subsequent trials. This general finding is remarkable, given the completely unconstrained nature of the model in terms of exploration and exploitation states. All possible sequences of these states over trials are given equal prior probability, and all could be inferred if the decision data warranted.

4.3. τ-Switch: A simpler latent state model

The fact that both optimal and human data lead to a highly constrained pattern of exploration and exploitation states across trials suggests an obvious simplification of the latent state model. Specifically, the pattern of change across the latent states can be well modeled using only a single parameter, controlling when exploration switches to exploitation. That is, rather than needing a latent state

[1] Note that, by its construction, neither choice in the explore–exploit situation is clearly better. In this way, our operational definition of 'exploration' and 'exploitation' is very narrow, and is just meant to characterize how decisions are made in this one case. A bit of care is needed in interpretation, since our definition of, say, 'exploration' is different from that of π-first or ϵ-greedy, and none of the operational definitions match the scope of the natural language use of the words.

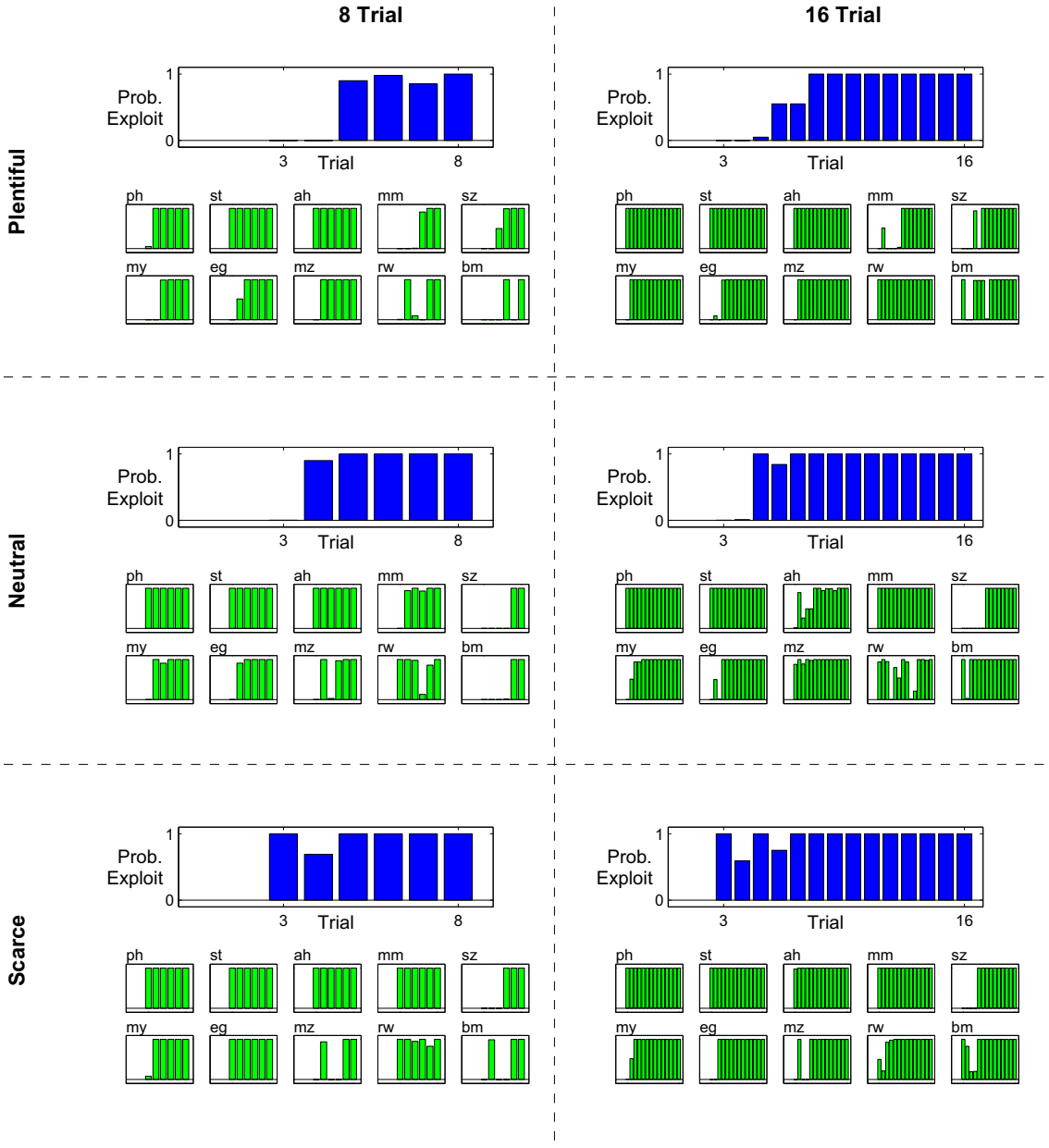

Fig. 3. Each bar graph shows the inferred probabilities of the exploitation state over the trials in a bandit problem. Each of the six panels corresponds to an experimental condition, varying in terms of the plentiful, neutral or scarce environment, or the use of 8 or 16-trials. Within each panel, the large blue (darker) bar graph shows the exploitation probability for the optimal decision process, while the 10 smaller green (lighter) bar graphs correspond to the 10 participants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

parameter for each trial, only a single switch-point parameter is needed, with all earlier trials following the exploration state, and all later trials following the exploitation state.

Thus, in the final version of our new model we use a parameter τ to control the trial at which the switch takes place, and so call this the τ-switch model. Psychologically, the τ-switch model has the same memory requirements as the ϵ-greedy, ϵ-first and π-first heuristics. The τ-switch model also takes into account the horizon, using the same latent stage approach as the π-first model. But τ-switch is fundamentally different from π-first, because of the way it makes exploration and exploitation decisions, using the approach summarized by Fig. 2. Recall, for example, that π-first chooses randomly in its initial exploration state, whereas τ-switch makes 'same', 'better' or 'explore' choices, depending on the successes and failures for each alternative on each trial. It is the detail of the decisions τ-switch makes, depending on how its internal state relates to the state of reward history observed, and its justification in terms of the detailed trial-wise analysis presented here, that makes the τ-switch model new and interesting.

5. Evaluation of the new and benchmark models

We implemented all five models—the existing WSLS, ϵ-greedy, ϵ-decreasing, and π-first models, and our new τ-switch model—again as probabilistic graphical models using WinBUGS. Using these implementations, we first test the ability of the models to account for optimal decision-making, then their ability to account for individual participant decision-making. Finally, we use the inferred parameter values of the models to begin to characterize optimal decision-making, human decision-making, and the relationship between the two.

We evaluate the ability of the models to produce both human and optimal decisions by calculating their posterior predictive average agreement with the optimal decision data. Posterior prediction is a standard Bayesian approach to assessing how well a model describes data, widely used in statistics (Gelman, Carlin, Stern, & Rubin, 2004; Bernardo & Smith, 2000; Congdon, 2006), and being adopted in cognitive science (Shiffrin, Lee, Kim, & Wagenmakers, 2008). It takes the behavior that a model predicts at *all* possible parameter settings of the model, and weights those predictions by the posterior probability of each parameter setting. This produces a posterior predictive distribution over the data space, which can be compared to the data actually observed. An important property of posterior predictive methods is that they automatically balance goodness-of-fit with model complexity in their evaluation. This is because complicated models are ones that can produce a wide range of behavior—and so fine-tune parameters to fit whatever is observed—but posterior prediction considers every possible model behavior in its overall evaluation. Intuitively, posterior predictions characterize the average, rather than the best-case, behavior of the model, and this averaging controls model complexity.[2]

5.1. How close to optimal are the models?

The posterior predicted average agreement between all five heuristics and optimal decisions is shown in Fig. 4, for all six bandit problem conditions. It is clear WSLS model is not able to capture optimal decision-making very well in any condition, but that the ϵ-greedy, ϵ-decreasing and π-first models are able to do much better. It is also clear and that our new τ-switch model is able to make optimal decisions most often. This ordering of the models in terms of their optimality holds for all six of the bandit problem conditions, although the absolute level of agreement generally changes systematically. In particular, agreement improves for the 16-trial problems, and is better for plentiful and scarce environments than for neutral environments.

5.2. Modeling human performance

Fig. 5 examines the ability of the models to account for human decision-making, showing the posterior predictive average agreement of each model to each individual participant. Participants are shown as bars against each of the models. We conduct analysis at the level of individual participants to allow for the possibility of individual differences. This intuition seems to be borne out. For the first 8 of the 10 participants (shown in darker blue), the τ-switch models provides the greatest level of agreement. For the last 2 of the 10 participants (shown in lighter yellow), this result is not observed, but it is clear that none of the models is able to model these participants well. One possibility is that these participants may have changed decision-making strategies during completing the 50 problems, and this prevents any single model from providing a good account of their performance.

Overall, however, our results show that, for the large majority of participants well described by any model, the τ-switch model is the best. In fact, Fig. 5 suggests that the ability of the model 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 models, which are approximately the same, and then slightly improved by the new τ-first model.

5.3. Characterization of optimal decision-making

We applied the models to optimal decision process data, and inferred posterior distributions over the model parameters. Table 1 shows the expected value of the inferred posterior distribution for the key parameter in each model. These parameter values constitute single numbers that characterize optimal decision-making within the constraints of the decision-making processes assumed by each model. They are shown for each of the plentiful, neutral and scarce environments for both 8- and 16-trials 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 probability of random exploration ϵ_0, which decreases as trials progress, is higher for more rewarding environments, and also for problems with more trials.

[2] It is worth noting that computationally simpler methods for model evaluation, like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), would be inappropriate for comparing our models, because they equate model complexity with a parameter count (Pitt, Myung, & Zhang, 2002). Many of our heuristics have the same number of parameters, but differ in complexity in the way those parameters interact with each other to control decision-making. The AIC and BIC would not detect these difference, but posterior predictive methods do.

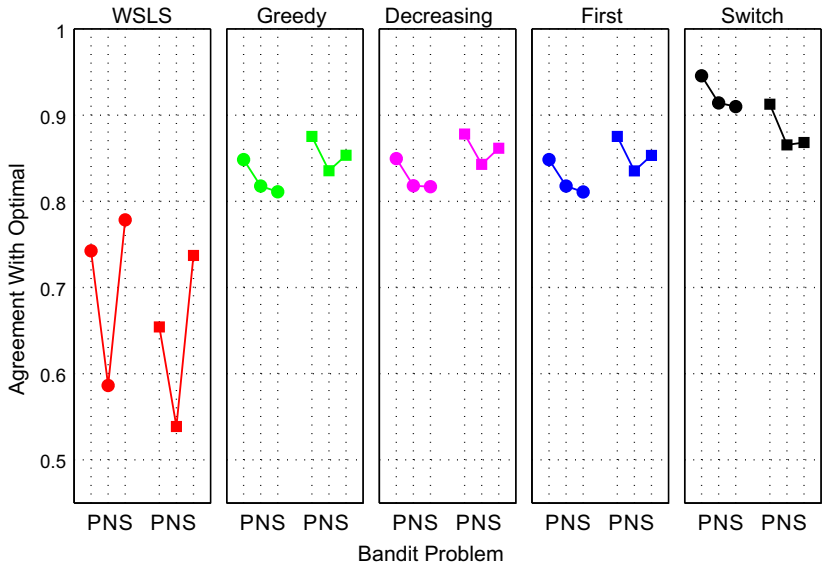

Fig. 4. Posterior predictive average agreement of the models with the optimal decision data. Agreement is shown for every model (panels), for the P = plentiful, N = neutral and S = scarce environments, and for the 8-trial (circle, left) and 16-trial (square, right) problems.

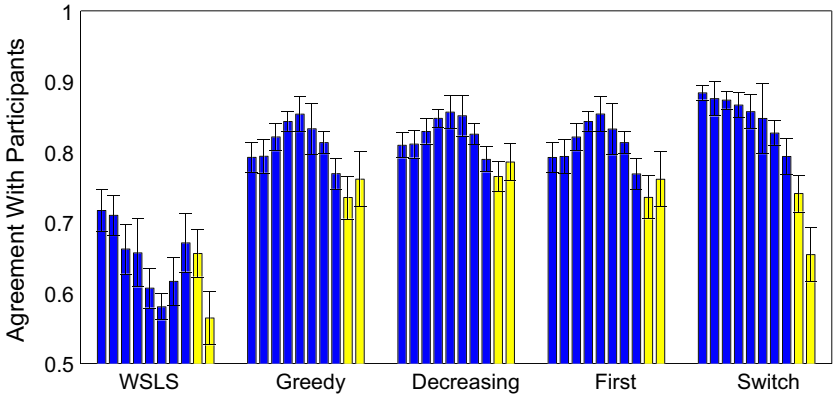

Fig. 5. Posterior predictive average agreement of the heuristic models with each individual participant. Two 'outlier' participants, not modeled well by any of the heuristics, are highlighted in lighter yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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

	Plentiful		Neutral		Scarce	
Heuristic	8	16	8	16	8	16
WSLS (γ)	0.87	0.85	0.85	0.78	0.72	0.65
Greedy (ϵ)	0.91	0.93	0.95	0.95	0.94	0.93
Decreasing (ϵ_0)	0.62	0.76	0.57	0.75	0.56	0.63
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

The π-first parameter is the trial at which the switch from random exploration, to choosing the most rewarding alternative, takes place. 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 making 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 model detail the optimal trial to switch from exploring to exploiting. 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-trials 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.

5.4. Characterization of human decision-making

The analysis in Fig. 4 shows the τ-switch model can closely emulate optimal decision-making for bandit problems,

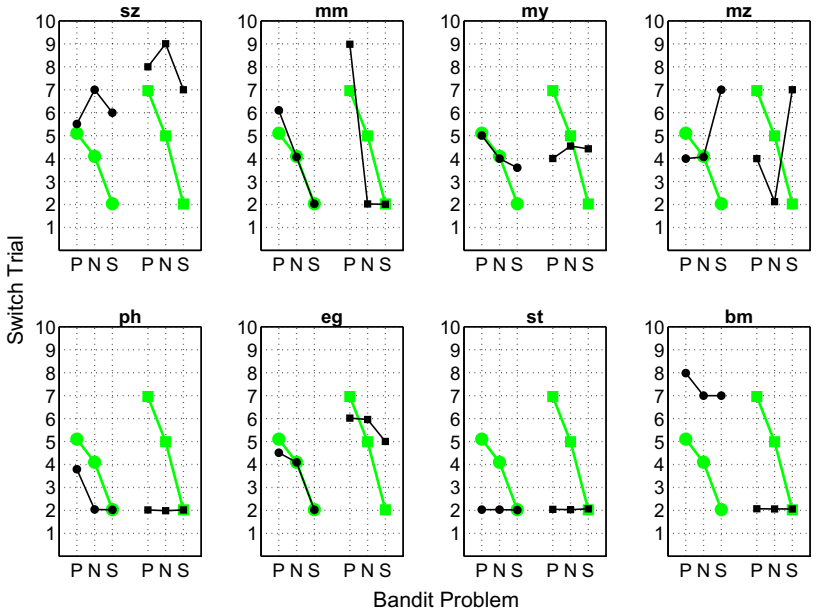

Fig. 6. Relationship between the optimal switching point under the τ-first heuristic in (larger, ligher, and green markers) and inferred switch-points for eight subjects (smaller, darker, and black markers). Comparisons are shown for P = plentiful, N = neutral and S = scarce environments, and 8-trial (circle, left) and 16-trial (square, right) environments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and the analysis in Fig. 5 shows it can also describe most participants' behavior well. Taken together, these results let us use the τ-switch model to address our motivating goal of comparing people's decisions to optimal decisions in psychologically meaningful ways. The key psychological parameters of a model like τ-switch provide a measure that relates people to optimality.

Fig. 6 gives a concrete example of this approach. Each panel corresponds to one of the eight participants from Fig. 5 who were well modeled by the τ-switch heuristic. Within each panel, the large lighter 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-trials problems. Overlayed in each panel, using smaller darker 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, allow for quick and insightful analyses of each participant's performance. For example, participants like MM and EG are choosing near optimally, especially in the 8-trial problems, and are sensitive to the reward rates of the environments, with appropriately larger numbers of trials devoted to exploration for the 16-trial problems. Their deviations from optimality seem more a matter of 'fine tuning' exactly how early or late they switch away from exploratory search behavior. Participants SZ and MZ, in contrast, are reacting to the changes in environment in qualitatively inappropriate ways. Participants MY, PH, and BM seem to perform better on the 8- than the 16-trials problems, and do not seem to be adjusting to the different environments in the 16-trial case. But MY is switching at roughly the optimal trial on average, while PH is switching too early, and BM is too early for the shorter problems and too late for the longer ones. Finally, participant ST seems to be employing a 'degenerate' version of the τ-switch heuristic that involves no initial search, but simply chooses the highest success rate alternative throughout the problem.

This analysis is intended to give an example of a basic approach, in which a model, like τ-switch, that is able to mimic optimal behavior well, and describe many people's behavior well, can then be used to understand how people are meeting or falling short of the demands of optimality. The variety of mismatches in Fig. 6 suggests that people are not optimal, and deviate in potentially many complicated ways. The key point is that using heuristics that can behave like people and optimal decision-makers provides an approach for building an understanding of the variety of behavior.

Potentially, of course, the other heuristics could also provide alternative characterizations with some level of justification. And there may be more that could be learned by jointly examining the accuracy of execution parameter for the τ-switch heuristic together with the key trial switch parameter. What the sketched analysis does provide a concrete illustration of the way human and optimal performance can be characterized by parametric variation using our new model.

6. Discussion

One finding from our results is that the τ-switch model 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 approximate 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. To this end, the model comparisons we have done have theoretical implications for understanding the nature and limitations of human decision-making. Most obviously, it is clear heuristics that include some sort of memory outperformed the one that did not. But, it would be interesting to ask a harder theoretical question about the nature of memory, and compare heuristics that did or did not, for example, give greater weight to more recent information. Similarly, a potential theoretical implication of the success of the τ-switch model is that people may use something like latent states to control their search behavior, and manage the exploration versus exploitation tradeoff. We think these sorts of models deserve as much attention as those, like ϵ-greedy, based more directly on reinforcement learning.

There are a number of potential extensions to the types of models we have considered. We noted in identifying two outlier participants that some sort of strategy shift might have been responsible for their incompatibility with the current models, which assume the same decision process is applied throughout. The latent state approach provides a natural means of allowing for such shifts. More generally, our models do not tackle the central issue of internal adaptation, learning or self-regulation in sequential decision-making. Their behavior is driven by the application of simple pre-determined rules to a changing environment. It seems likely that a more complete account of how people solve bandit problems would include the possibility of tuning existing strategies, or learning new ones. A good concrete issue to start the extension to learning might be developing a theory of how parameters for current models change systematically across different task environments. Our results show, for example, that τ-switch uses different switch-points in plentiful, neutral and scarce environments, but does not say how those points are determined. A more complete psychological account should model, rather than measure, these adaptations.

On a different front, one potential practical application of the τ-switch model 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 model 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 Fig. 6, 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 using simple heuristic cognitive 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 the use of 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 small-horizon two-alternative bandit problems, and involve a small set of models, we think our basic approach has much more general applicability. Heuristic models can be assessed in terms of their ability to model human or optimal decision-making, and their inferred parameter values can be used to understand and compare how those decisions are made.

Acknowledgments

This work is was supported by an award from the Air Force Office of Scientific Research (FA9550-07-1-0082) to Michael Lee and Mark Steyvers. We thank Jun Zhang for suggesting we evaluate the type of latent state model developed and tested in this paper.

References

Batchelder, W. H. (1970). An all-or-none theory for learning on both the paired-associate and concept levels. *Journal of Mathematical Psychology, 7*, 97–117.

Bernado, J. M., & Smith, A. F. M. (2000). *Bayesian theory*. Chichester, UK: Wiley.

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

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? Exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*, 933–942.

Congdon, P. (2006). *Bayesian statistical modeling*. Chichester, UK: Wiley.

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.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (second ed.). Boca Raton, FL: Chapman & Hall.

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

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in Neural Information Processing Systems*, 17.

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

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*(1), 1–15.

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.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*(8), 1248–1284.

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

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

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*, 1011–1026.

Zhang, S., Lee, M. D., & Munro, M. N. (2009). Human and optimal exploration and exploitation in bandit problem. In Proceedings of the 9th International conference on cognitive modeling (ICCM), Manchester, UK.