Exploitative and Exploratory Attention in a Four-Armed Bandit Task

Adrian R. Walker (adrian.walker@unsw.edu.au)
School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia

Mike E. Le Pelley (m.lepelley@unsw.edu.au)
School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia

Tom Beesley (t.beesley@unsw.edu.au)
School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia

Abstract

When making decisions, we are often forced to choose between something safe we have chosen before, and something unknown to us that is inherently risky, but may provide a better long-term outcome. This problem is known as the Exploitation-Exploration (EE) Trade-Off. Most previous studies on the EE Trade-Off have relied on response data, leading to some ambiguity over whether uncertainty leads to true exploratory behavior, or whether the pattern of responding simply reflects a simpler ratio choice rule (such as the Generalized Matching Law (Baum, 1974; Herrnstein, 1961)). Here, we argue that the study of this issue can be enriched by measuring changes in attention (via eye-gaze), with the potential to disambiguate these two accounts. We find that when moving from certainty into uncertainty, the overall level of attention to stimuli in the task increases; a finding we argue is outside of the scope of ratio choice rules.

Keywords: Reinforcement Learning; Attention; Decision-Making; Exploitation/Exploration Trade-Off; Bandit Task.

Introduction

In everyday decision-making, we often have to choose between trying something new, or sticking with what we know. For example, when deciding what to eat at a restaurant, we can choose to order our regular "safe" meal (e.g., spaghetti bolognese), or try a new "risky" meal (e.g., steak tartare). By ordering the risky meal, we learn about how tasty it is. If it is tastier than our regular meal, we may become more likely to order it on subsequent visits. However, if it is worse than the regular meal, we wasted an opportunity to sample our regular meal. This problem is known as the *Exploitation/Exploration Trade-Off* (or "EE Trade-Off") (Cohen, McClure, & Yu, 2007; Knox et al., 2012; Mehlhorn et al., 2015).

One common method used to study the EE Trade-Off is the Multi-Armed Bandit Task (e.g., Daw et al., 2006; Gittins, 1979; Knox et al., 2012; Speekenbrink & Konstantinidis, 2015). On each trial, participants are presented several "arms" and are asked to pick one arm to receive some reward (e.g., points). Each arm provides a different amount of reward, with the goal of the participant being to maximize the amount of reward they receive. Participants are not told the value of each arm at the outset, and must learn these values through sampling each arm. The reward structure is generally stochastic, with the value of each arm changing gradually over time (e.g., Daw et al., 2006; Laureiro-Martinez et al., 2015). The key measurement of this task is how often participants choose the arm which gives the highest observed pay-off. Generally, a participant is considered to be "exploiting" an arm if they choose the arm with the highest observed pay-off, while they are considered to be "exploring" when making any other choice (Knox et al., 2012).

Explanations for Exploration

Work by Daw et al. (2006) found evidence that exploitation is the "default" for human behavior, while exploration is a high-level decision *not* to exploit on a given trial. Subsequent research with the multi-armed bandit task has primarily focused on determining what parameters induce exploratory responding over exploitative responding.

Two major accounts have been proposed for what causes people to switch from exploitation to exploration. One influential account that has emerged argues that *environmental uncertainty* is key in motivating exploration (Beesley, Nguyen, Pearson, & Le Pelley, 2015; Gold & Shadlen, 2007; Knox et al., 2012; Speekenbrink & Konstantinidis, 2015). That is, the less certain a participant is about the dynamics of their environment, the more likely they are to spend time exploring it (Mehlhorn et al., 2015). For example, if the quality of food at a restaurant is highly variable, you may explore many different meals before settling on a preferred one. By contrast, if the quality of meals is fairly consistent, you may quickly settle on a preferred meal. The key implication of this account is that exploration is an *intentional* attempt to reduce the amount of uncertainty in the environment (and thus aid informed decision making).

The other major account argues that in most cases, exploration can be explained by ratio choice rules (Sakai & Fukai, 2008). In their review, Gold and Shadlen (2007) suggested that most exploratory behavior might adequately be explained by a form of Herrnstein's (1961) Matching Law (See also Baum, 1974). The Matching Law states that the ratio of responding on each arm is equivalent to the ratio of reinforcement for each of those arms. That is, participants "match" how often they select each arm, based on the

perceived average reward for the selected arm compared to others. For example, in a two-armed bandit task, where arm A is reinforced 3 times as often as arm B, the Matching Law states that people will select arm A 3 times as often as arm B. Importantly, while participants still preferentially select the optimal arm (A), they will also switch to the other, sub-optimal arm (B) on 25% of trials. In this case, switching away from the optimal arm does not represent an intentional attempt to lessen uncertainty in the environment, but instead represents participants employing a (somewhat crude) ratio choice rule.

Baum (1974) provided an extension to Herrnstein's (1961) "Simple" Matching Law to account for a wider array of choice behavior. This "Generalized Matching Law" included two additional parameters: *bias* and s*ensitivity*, where *bias* reflects a tendency for selecting a given option over other available options (irrespective of the reinforcement rate for each option), and *sensitivity* determines how strictly a participant conforms to the choice ratio for their selections. The Generalized Matching Law has been shown to account for a wider variety of choice behavior than the Simple Matching Law (Baum, 1974; Schneider & Lickliter, 2010), and is the version applied in this paper.

It is important to note that, even when employing a ratio choice rule like the Matching Law, participants can still update their knowledge of the environment by picking sub-optimal responses (as determined by the choice rule). However, the crucial distinction is that exploratory choices occur on the basis of a ratio determined by the choice rule, and are not an intentional attempt to lessen uncertainty. For the purpose of the current paper, this type of behavior may be considered synonymous with the phenomenon known as *probability matching* (Sakai & Fukai, 2008; Shanks, Tunney, & McCarthy, 2002 – though strictly these two phenomena are slightly different, see Shanks et al., 2002).

The main difference between the uncertainty account and the ratio choice rule account of exploration is that, in the former, uncertainty is a catalyst for participants to explore (and thus lessen the total uncertainty in the task); while in the latter, exploration occurs as a product of some choice function. One key issue in attempting to differentiate these two accounts is that it is difficult to increase uncertainty without changing the reward value of the different arms. For example, in the commonly used "walking bandit task" (Daw et al. 2006; Laureiro-Martinez et al., 2015; Speekenbrink & Konstantinidis, 2015), uncertainty is implemented by stochastically walking (slowly changing) the mean reward for each arm every trial. While this does serve to make the value of each arm uncertain, it also necessarily causes changes to the probability of picking each arm as given by ratio choice rules. Therefore, it is hard to determine whether an attempt to reduce the uncertainty in the task, or a predetermined ratio choice rule, is responsible for motivating exploration under these circumstances.

While it is possible to use cognitive modeling techniques to examine whether uncertainty motivates exploration (e.g., Daw et al., 2006; Knox et al., 2012; Speekenbrink & Konstantinidis, 2015; Stevyers, Lee, & Wagenmakers, 2009), the conclusions of these methods have been mixed. A recent study by Beesley et al. (2015) argued that attention may be another viable metric for assessing the EE Trade-Off. Beesley et al. conducted a study in which participants were presented with two cues and were asked to make a choice between two responses. One cue was informative about what the optimal response was on that trial, while the other cue was task-irrelevant. Beesley et al. measured participants' attention by tracking eye-gaze on the two cues. They showed that when cues were perfect predictors of the optimal response, participants attended to the informative cue over the task-irrelevant cue. However, when cues were *imperfect* predictors of the optimal response (i.e., predicted the optimal response on only two-thirds of trials), participants increased their attention to both the informative and task-irrelevant cue, indicating greater exploration of the cues. Beesley et al. argued that these findings were synonymous with the EE trade-off. The implication, therefore, is that exploration can be exhibited in behavioral domains *outside* of participant choice, and that exploration cannot be solely explained by ratio choice rules (the predictions of which are restricted to the response domain alone).

One limitation to the Beesley et al. task was that the experimenters provided participants with feedback about which response was optimal on each trial. Therefore, while participants appeared to explore the information in the task by altering their attentional processing, they had no incentive to explore different responses to find the one that was most optimal (as they were told which response was optimal regardless of their choice). Thus, the current paper aims to assess whether uncertainty can induce exploratory behavior in both participants' attention and responses, and hence provide wider support for the idea that uncertainty drives exploration. This would imply that exploration itself is perhaps a more complex, intentional process, which would place it outside the scope of ratio choice rules alone. Furthermore, it would suggest that exploration can manifest itself across more than one aspect of behavior (choice and attention). We used a four-armed bandit task where we manipulated uncertainty both within and between subjects. We measured responding and gaze-time to the different arms during the task. In line with our hypotheses, we found that participants made fewer optimal responses and spent longer fixating on task elements when the task had an element of uncertainty, suggesting that uncertainty acts as a catalyst for exploration.

Method

This experiment aimed to examine the effect of uncertainty on attention and responding in a four-armed bandit task. Participants completed a variant of the bandit task, where on every trial they were presented four arms and asked to pick two. Two of the arms conferred 30 points (the *High Value* [HV] arms) and the remaining two arms conferred 15 points

(the *Low Value* [LV] arms). After making their choice, participants were rewarded with the cumulative score associated with each arm. For example, if the participant selected an HV arm worth 30 points and an LV arm worth 15 points, they received a reward of 45 points for that trial.

The experiment was conducted in two stages. Stage 1 of the task had a deterministic reward structure, in which each arm always yielded the same amount. Stage 1 was designed to be simple, such that participants could quickly learn the structure of the task and engage in what might be considered an exploitative pattern of behavior. In Stage 2 of the task, rewards were drawn stochastically from a uniform distribution for each combination of arms, with the mean reward value set at the same value as in the first stage. Stages 1 and 2 were coined the *Certain* and *Uncertain* stages respectively. In terms of participants' responding, we hypothesized that when rewards became uncertain, participants would make more exploratory, non-optimal responses, and this exploration would be greater for participants who experienced greater uncertainty (consistent with Knox et al., 2012; Speekenbrink & Konstantinidis, 2015). In terms of gaze-time, we hypothesized that when rewards became uncertain in Stage 2, participants would increase their gaze-time to all arms in the task. Furthermore, we hypothesized that the greater the level of uncertainty in those rewards in Stage 2, the more gaze time that would be allocated to the arms in the task, with more gaze time to HV arms over LV arms (consistent with Beesley et al., 2015).

Design

The design of the experiment is shown in Table 1. The key manipulation of the amount of uncertainty present in Stage 2 was manipulated between-subjects. Uncertainty was operationalized as the range of possible scores around the mean reward value that could be received following a trial (in Stage 2). For example, a reward distribution of ±3 (Low Uncertainty) meant that after the participant made their selections, they received the cumulative score of those arms (e.g., 45 points if they picked one HV arm and one LV arm), ±3 points (uniformly distributed across trials). Therefore, in this case, the participant could receive a score from 42 to 48. The Low Uncertainty condition had a reward distribution of ±3, and the High Uncertainty condition had a reward distribution of ±18. The crucial difference between these two conditions was that in the High Uncertainty condition, the two score distributions *overlapped*, such that participants could sometimes earn more points after a choice of an HV arm and an LV arm than after a choice of two HV arms. By comparison, for participants in the Low Uncertainty condition, the optimal response was always picking two HV arms. The dependent variables were proportion of HV arms picked, and gaze-time on HV and LV arms as a proportion of trial time.

Table 1: Design of Experiment 1

Uncertainty condition	HV Reward	LV Reward	Reward uncertainty (Stage 2)
Low	30	15	±3
High	30	15	±18

Participants

Sixty-five UNSW Sydney undergraduate students were recruited in exchange for course credit. The two highest scoring participants received a $20 prize.

Apparatus and Materials

Participants were tested individually in a quiet room. During the task, participants' eye-gaze was tracked using a 58.4cm widescreen Tobii eye-tracking monitor (TX-300). Participants were seated approximately 60cm from the monitor, and had their heads steadied by a chin rest. The eye-tracker was calibrated at the start of the task. The experiment was run in MATLAB using the Psychophysics Toolbox extension (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Participants mades all responses via a standard keyboard and mouse.

The four arms in the experiment were represented as four colored squares of 200 by 200 pixels (visual angle of approximately 5°). The four colors were always red, green, blue, and yellow (*Figure 1*). Color assignment to design elements (i.e., HV and LV arms) was counterbalanced between participants (24 permutations).

Procedure

At the start of the experiment, participants were instructed that they would be playing a simple guessing game, where the objective of the game was to maximize the number of points they received. On each trial, the four colored arms were presented in the four quadrants of the screen. The location of each arm was counterbalanced between trials, with a full counterbalance of positions taking 24 trials. Participants used the mouse to select two arms. Participants were allowed to deselect arms they had selected by clicking on the arm a second time. Once the participant had selected two arms, a small "Submit" (120 by 60 pixels) button appeared in the center of the screen. If the participant selected more than two arms, or deselected an arm, the button disappeared.

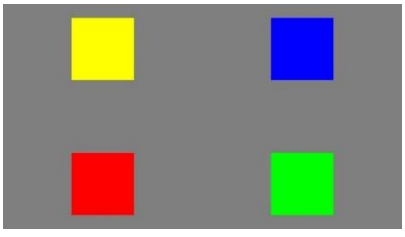

Figure 1: A sample screen from Experiment 1.

Once the submit button was clicked, the four arms and the cursor disappeared, and the participant was told how many points they had earned on that trial, as well as the total points accumulated so far. Points were calculated by aggregating the value of the two arms the participant had selected, with the addition of the reward uncertainty in Stage 2 (see Design). Participants then pressed the spacebar to start the next trial. The location of the cursor was reset to the center of the screen on each trial.

Stage 1 consisted of 96 trials and Stage 2 144 trials. The start of Stage 2 was not signaled to participants in any way. The only difference between Stage 1 and 2 was the addition of the variability (stochastic noise) for rewards (see Design and Table 1).

Results

Data were collapsed into blocks of 24 trials for analysis. If a participant had less than 50% of trials with valid eye-tracking data recorded, they were excluded from analysis (n = 10). In addition, participants who selected the HV arms less than 70% of the time in the final block of Stage 1 were inferred to have not learnt the associations adequately, and were also excluded (n = 7). For each exclusion, we ensured a complete counterbalancing of design elements by recruiting a new participant with the same counterbalancing conditions. Trials in which the participant took two standard deviations longer than their mean trial time were excluded from all analyses.

Response data are shown in *Figure 2* and were analyzed in three parts, Stage 1 (blocks 1 to 4), the between-stage transition period (blocks 4 and 5), and Stage 2 (blocks 5 to 10), using a repeated measures ANOVA with a within-subjects factor of block and a between-subjects factor of condition. Effect sizes are reported as generalized eta-squared, η^2_G (see Bakeman, 2005). In Stage 1, a significant effect of block was observed, $F(3, 138) = 98.84$, $p < .001$, $\eta^2_G = .447$, with participants in both conditions increasing selections of HV arms as they progressed through Stage 1.

During the transition from Stage 1 to Stage 2, a significant effect of block was observed, $F(1, 46) = 80.4$, $p < .001$, $\eta^2_G = .465$, with participants decreasing their selections of HV arms from Stage 1 to Stage 2. A significant effect of condition was also observed, $F(1, 46) = 18.06$, $p < .001$, $\eta^2_G = .165$, with participants less likely to choose the HV arms in the High Uncertainty group. Finally, there was a significant interaction between block and condition, $F(1, 46) = 22.56$, $p < .001$, $\eta^2_G = .197$, with the proportion of HV arm choices showing a greater decrease in the high uncertainty conditions than the low uncertainty condition during the transition period.

In Stage 2, a significant effect of block was observed, $F(5, 230) = 22.68$, $p < .001$, $\eta^2_G = .165$, with participants increasing their selections of HV arms over the course of Stage 2. A significant effect of condition, $F(1, 46) = 24.71$, $p < .001$, $\eta^2_G = .243$, and a significant interaction between condition and block, $F(5, 230) = 5.02$, $p < .001$, $\eta^2_G = .042$, were observed, with participants picking HV arms less frequently in the High Uncertainty condition, but also showing a greater increase in their selection of HV arms over the course of Stage 2, compared to participants in the Low Uncertainty condition.

Gaze-time data are shown in *Figure 3*. Gaze-time was calculated as the summed time of all fixations on the different arms in the task. A fixation was determined to have occurred if a participant's gaze did not deviate more than 75 pixels vertically or horizontally for at least 150ms. Total fixation time was calculated by extending this time until the participant's gaze exited the 75 pixel limit (in accordance with Beesley et al., 2015; Salvucci & Goldberg, 2000). Proportion of gaze-time was calculated as the total fixation on each arm divided by the total trial time. Again, these data were analysed using a repeated measures ANOVA, with a within-subjects factor of block, a within-subjects factor of arm value (high and low), and a between-subjects factor of condition. A significant effect of block was observed in Stage 1, $F(3, 138) = 18.85$, $p < .001$, $\eta^2_G = .062$, with participants decreasing their total gaze-time to arms throughout Stage 1. A significant effect of arm value was also observed, $F(1, 138) = 276.77$, $p < .001$, $\eta^2_G = 0.675$, along with a significant interaction between block and arm value, $F(3, 138) = 17.86$, $p < .001$, $\eta^2_G = .066$, with participants gazing more at HV arms than LV arms, and this difference increasing over the course of Stage 1.

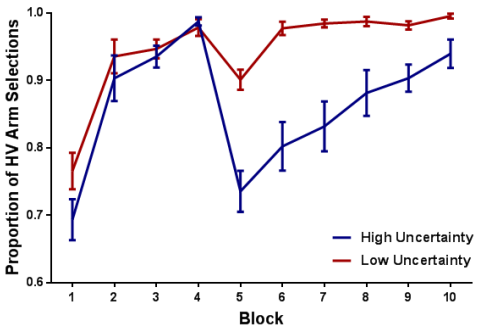

Figure 2: Proportion of HV arms selected in each block. Stage 1 occurred in Blocks 1 to 4, while Stage 2 occurred in Blocks 5 to 10. Error bars represent ±1 *SEM*.

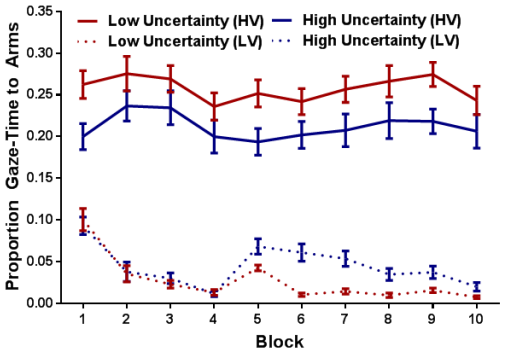

Figure 3: Proportion of trial time gazing at HV and LV arms in each block for each condition. Stage 1 occurred in Blocks 1 to 4, while Stage 2 occurred in Blocks 5 to 10. Error bars represent ±1 *SEM*.

In the transition from Stage 1 to Stage 2, there was a significant effect of block, $F(1, 46) = 15.80$, $p < .001$, $\eta^2_G = .039$, with participants increasing their total gaze-time at the onset of uncertainty. The significant effect of arm value was maintained, $F(1, 46) = 282.29$, $p < .001$, $\eta^2_G = .713$. There was no effect of condition observed on gaze-time, $F(1, 46) = 2.21$, $p = .144$, and there was no interaction between block and condition, $F < 1$. In Stage 2, the significant effect of arm value was maintained, $F(1, 46) = 362.42$, $p < .001$, $\eta^2_G = .721$, and no effect of condition was observed, $F < 1$. No effect interaction between block and condition was observed in Stage 2, $F(5, 230) = 1.19$, $p = .316$[1].

Discussion

In a reinforcement learning task, participants earned points for combinations of responses. In Stage 1, one combination of responses was optimal and participants readily learnt this relationship. In Stage 2, we introduced variation in the number of points received, while keeping the mean number of points per response constant. When moving from the certainty of Stage 1 to the uncertainty of Stage 2, participants in both conditions reduced their rate of optimal responding, and this reduction was greater for participants who experienced greater uncertainty. Following this change in behavior at the outset of Stage 2, participants in both conditions increased their rate of optimal responding over the course of Stage 2.

In the High Uncertainty condition, the choice behavior of participants is well predicted by the Matching Law. However, crucially in the Low Uncertainty condition the Matching Law fails to predict the drop in optimal responding at the onset of. If a participant in the Low Uncertainty condition were following the Matching Law, they should not show a decrease in optimal responding at the onset of uncertainty. This finding suggests that exploratory choice cannot be solely explained by the Matching Law, and provides support to the idea that uncertainty can drive exploration. The reason why participants in the Low Uncertainty condition chose to switch away from the optimal response at the onset of uncertainty is not immediately clear. However, one possible explanation is that when participants perceived that the nature of the task had changed (i.e., rewards were no longer confined to three set values), they felt compelled to explore the other previously discounted responses to ensure they had not changed in any significant way.

In terms of the attentional data, we found support for two of our three hypotheses. Unsurprisingly, participants began to pay more attention to HV arms over LV arms over the course of the experiment. This is compatible with a host of research from the associative learning literature (See Le Pelley et al., 2016, for a review), showing that participants are likely to direct their attention to the most valuable predictors in a task (also see Le Pelley et al., 2015). Furthermore, there is evidence that participants will attend more to arms they are intending to select prior to making their response (e.g., Manohar & Husain, 2013). As participants selected more HV arms, this likely contributed to participants preferentially attending to them over LV arms.

Crucially, we have shown evidence that an onset of uncertainty is associated with an increase in attention. Once rewards became uncertain at the onset of Stage 2, participants in all conditions increased their gaze-time to all arms in the task. Our data are in line with the findings of Beesley et al. (2015), and provide support to the idea that uncertainty can instigate exploratory behavior in both the choice responses and attentional bias. We argue that these data are beyond the scope of ratio choice rules, which do not provide a natural account of attentional changes under conditions of uncertainty and would not predict changes in response rate across the course of Stage 2. While the notion that uncertainty increases attentional processing of stimuli is not novel (Pearce & Hall, 1980), very little is known about attentional processing in multi-armed bandit tasks like the one used in the current experiment. The current findings suggest that pursuing this line of research may be important to gaining a more complete understanding of human decision-making.

However, we did not find evidence for gaze-time interacting with the level of uncertainty. If the uncertainty account of exploration is correct, we should have observed greater exploration under greater uncertainty. Instead, the amount of gaze-time participants paid to the arms was comparable under both levels of uncertainty. One possible reason for this is that moving from a completely certain environment to an environment with *any* level of uncertainty may cause attention to increase. Yu and Dayan (2005) showed that participants behave differently when uncertainty is expected (i.e., present for the entire task) compared to when uncertainty is unexpected (i.e., a period of uncertainty occurs suddenly, following a period of certainty). It may be the case that when unexpected uncertainty occurred, attention increased by a set amount in response (regardless of the degree of that uncertainty). Also, while gaze-time does appear to be affected by uncertainty, the effect-size in our study was much smaller in comparison to the effect of uncertainty on responding ($\eta^2_G = .039$ compared to $\eta^2_G = .465$). This may suggest that while changes in response rate and changes in overt attention are signals of exploration under uncertain conditions, that uncertainty affects these behavioral markers by distinct mechanisms. Alternatively, these data might suggest that gaze-time was less sensitive to uncertainty than was

[1] All data and analyses can be accessed on the Open Science Framework at osf.io/y6hqp.

participants' responding, which made it harder to detect any effect of the different levels of uncertainty and gaze-time.

In summary, we have shown that the introduction of uncertainty into a four-armed bandit task caused a general increase in attending, and a decrease in optimal responding. This provides support for the idea that environmental uncertainty causes an increase in exploratory behavior, and challenges the idea that exploration can be explained purely by ratio choice rules.

Acknowledgements

This research was supported by an Australian Government Research Training Program (RTP) Scholarship, and by Australian Research Council grant DP140103268, awarded to Tom Beesley, Mike Le Pelley, and Chris Mitchell. The authors thank Ben Newell and Fred Westbrook for their help.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379-384.

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology (Hove), 68*, 2175-2199.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching, *Journal of the Experimental Analysis of Behaviour, 22*, 231-242.

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 B: Biological Sciences, 362*, 933-942.

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 (Methodological)*, 148-177.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*, 267-272.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology, 2*, 398.

Laureiro-Martínez, D., Brusoni, S., Canessa, N., & Zollo, M. (2015). Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. *Strategic Management Journal, 36*, 319-338.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*, 1111-1140.

Le Pelley, M. E., Pearson, D., Griffiths, O., & Beesley, T. (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General, 144*, 158-171.

Manohar, S. G., & Husain, M. (2013). Attention as foraging for information and value, *Frontiers in Human Neuroscience*, 62-77.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*, 191-215.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532-552.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Sakai, Y., & Fukai, T. (2008). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation, 20*, 227-251.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*, 71–78.

Schneider, S, M., & Lickliter, R. (2010), Choice in quail neonates: The origins of the generalized matching law *Journal of the Experimental Analysis of Behaviour, 94*, 315-326.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233-250.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. Topics in Cognitive Science, 7 , 351–367.

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* (Vol. 1): MIT press Cambridge.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*, 681-692.