Uncertainty and exploration in a restless bandit problem

Maarten Speekenbrink

Experimental Psychology, University College London

Emmanouil Konstantinidis

Department of Social and Decision Sciences, Carnegie Mellon University

Abstract

Decision-making in noisy and changing environments requires a fine balance between exploiting knowledge about good courses of action and exploring the environment in order to improve upon this knowledge. We present an experiment on a restless bandit task in which participants made repeated choices between options for which the average rewards changed over time. Comparing a number of computational models of participants' behaviour in this task, we find evidence that a substantial number of them balanced exploration and exploitation by considering the probability that an option offers the maximum reward out of all the available options.

Keywords: Dynamic decision making; Exploration-exploitation trade-off; Restless multi-armed bandit task; Uncertainty; Volatility

Uncertainty and exploration in a restless bandit problem

Introduction

In many situations, the expected utility of an action is initially unknown and can only be learned from experience. In such situations, we can take actions in order to maximise the utility experienced ("exploiting" the environment), but also take actions which we think might not provide as good outcomes, but which help us to learn more about the outcomes associated with that action ("exploring" the environment). Performing well in these situations requires a fine balance between exploration and exploitation (Sutton & Barto, 1998). Multi-armed bandit tasks have proven a useful paradigm to study the exploration-exploitation trade-off, theoretically (e.g., Gittins, 1979; Whittle, 1988) as well as empirically (e.g., Acuna & Schrater, 2008; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Knox, Otto, Stone, & Love, 2012; Steyvers, Lee, & Wagenmakers, 2009; Yi, Steyvers, & Lee, 2009).

For standard bandit problems, in which the expected rewards of unchosen arms remain unchanged, the optimal decision strategy can be determined through dynamic programming (Berry & Fristedt, 1985) or by calculating a "Gittins index" (Gittins, 1979) for each arm of the bandit, reflecting the expected total future rewards associated with the arm at a given time. Acuna and Schrater (2008) showed that, allowing for computational constraints, human decisions follow this optimal strategy reasonably well. Although standard bandit tasks have generated useful results, in real-life situations, the expected rewards of unchosen options do often change. For instance, when choosing a restaurant, we should allow for the possibility that the quality of the food on offer changes over time. It then makes sense to sometimes revisit restaurants in which our experiences were less than optimal, to assess whether the chefs may have improved upon their skills. For what is now a "restless" bandit problem, optimal decision strategies have proven elusive (Papadimitriou & Tsitsiklis, 1999), although heuristic strategies have been proposed (Whittle, 1988).

Daw et al. (2006) investigated human decision making in a restless bandit task, and found that exploration appeared to be unrelated to the uncertainty regarding the average reward associated with each arm. Exploration was best described by a heuristic strategy in which arms are chosen probabilistically according to their relative expected rewards (the "softmax" decision rule; Sutton and Barto, 1998). The lack of an effect of uncertainty on explorative decisions is disappointing, considering that rationally, this should be a driving factor of exploration (Cohen, McClure, & Yu, 2007). Knox et al. (2012) used a restless two-armed bandit task with a simplified structure, which allowed them to derive the optimal decision strategy. In their task, the rewards of the two arms alternated in their superiority, "leapfrogging" over each other. The results showed that people appeared to act reflectively, updating their beliefs that the arms switched in superiority, but that they could not use these beliefs to plan further ahead in time than for the immediate decision. Moreover, in contrast to Daw et al., Knox et al. found evidence that exploration was driven by uncertainty regarding the associated rewards. As Knox et al. used a much more constrained task than Daw et al., it is unclear whether this difference in results is due to the task, or to the way in which the effect of uncertainty on decisions was formalized. In the present paper, we will try to reconcile these conflicting results, by considering an alternative way to incorporate uncertainty into explorative decisions than the heuristic strategy used by Daw et al.

To illustrate our model, consider a relatively simple task in which, on each trial t, the reward $R_j(t)$ associated with arm j is drawn from a Normal distribution, with a mean $\mu_j(t)$ that changes over trials according to a random walk:

$$\begin{aligned} R_j(t) &= \mu_j(t) + \epsilon_j(t) \qquad & \epsilon_j(t) &\sim N(0, \sigma_\epsilon) \\ \mu_j(t) &= \mu_j(t-1) + \zeta_j(t) \qquad & \zeta_j(t) &\sim N(0, \sigma_\zeta) \end{aligned} \tag{1}$$

Note that there are two sources of variability in this process: the observation variance σ^2_ϵ, reflecting the extent to which rewards vary around their mean, and the innovation variance σ^2_ζ, reflecting the volatility of the environment (time-dependent variation in the mean rewards associated with each arm). An ideal Bayesian learner with knowledge of the properties of this process would update her belief about the average rewards based on the observed rewards; $p(\mu_j(t)|R_{1:t}, C_{1:t})$, the posterior distribution of arm j's average reward, conditional upon the obtained rewards $R_{1:t} = (R_1, \ldots, R_t)$ and choices made $C_{1:t}$, can be computed by the Kalman filter (Kalman, 1960; Kalman & Bucy, 1961). This posterior distribution, together with the structural model, allows the agent to derive a prior distribution $p(\mu_j(t+1)|R_{1:t}, C_{1:t})$ for an arm's average reward on the next trial. In turn, these prior distributions can be used to derive prior predictive distributions for the actual rewards that can be obtained by playing each arm:

$$p(R_j(t+1)|R_{1:t}, C_{1:t}) = \int p(R_j(t+1)|\mu_j(t+1))p(\mu_j(t+1)|R_{1:t}, C_{1:t})\, \mathrm{d}\mu_j(t+1). \tag{2}$$

How should these beliefs regarding obtainable rewards be used to choose the next arm to play? A "greedy" agent would always choose the arm with the highest expected reward (according to the prior predictive distribution). While this is optimal on the last play, if there are more plays left, it is generally beneficial to sometimes choose a different arm in order to check that its average reward has not surpassed that of the currently favoured arm. In a restless bandit task, the longer an arm has not been played, the higher the (subjective) probability that it may now provide better rewards. If exploration is based on this probability, the probability of exploration increases with the time that an arm has not been played. The difference between this explorative strategy and a greedy one is illustrated in Figure 1. As shown there, the explorative strategy clearly outperforms the greedy strategy.

The probability that an arm provides the maximum reward naturally combines both the expected value (mean) and the associated uncertainty (variance) of the prior predictive distributions. While a strategy which bases decisions on this probability is myopic in the sense that it is solely based on the chance of obtaining the highest possible reward for the immediate decision, it nevertheless allows for a reasonable balance between exploitation and exploration. As the probability of maximum reward increases with uncertainty, the probability of exploring an arm increases the longer its outcome has not been observed. But the probability of maximum reward is also dependent on the expected reward, such that this increase is larger for arms that are closer to the currently favoured arm in expected reward. The strategy of choosing arms according to the probability that they provide the maximum reward was first proposed by Thompson (1933) in the context of a classical bandit task, and has become known as "Thompson sampling". While Daw et al. (2006) did not find evidence that exploration is related to uncertainty, they only considered a heuristic decision strategy with an exploration bonus which increased linearly with the standard deviation of the prior distribution. The probability of maximum reward strategy offers a more principled way in which to combine expectancy and uncertainty and it has been shown to outperform other heuristic strategies (Granmo & Berg, 2010; Gupta, Granmo, & Agrawala, 2011; Viappiani, 2013). In addition, strategies such as those with an exploration bonus require fine-tuning of their parameters (i.e.., the relative weight given to exploration vs exploitation). In the probability of maximum reward strategy, the exploration-exploitation trade-off follows naturally from assumptions about the process governing the rewards (e.g., Equation 1).

In the present paper, we investigate whether humans performing a restless multi-armed bandit task use this strategy to make their decisions. We compare the strategy to a number of heuristic decision strategies proposed for multi-armed bandit tasks, contrasting also Bayesian "model-based" learning and two popular "model-free" learning strategies. Bayesian learning relies upon a model of the environment (e.g., Equation 1) to update beliefs about the (average) rewards associated with each arm. Model-free learning does not require a model of the environment, focussing instead on learning the expected rewards associated with actions directly (see Sutton & Barto, 1998, for a more in-depth discussion of model-based vs model-free reinforcement learning).

Method

We investigated decision-making in a restless four-armed bandit task similar to that used by Daw et al. (2006). Four versions of the task were constructed in which (a) the average rewards of the arms changed either completely unpredictably or with a small trend, and (b) the volatility of the changes was either stable or there were periods of relatively high volatility. We expected people who notice an increase in volatility to make more exploratory decisions, due to the associated increase in uncertainty. People who notice the trends could be expected to make relatively less exploratory decisions, as the changes in average rewards are more predictable.

Participants

Eighty participants (41 female), aged between 18 and 56 ($M = 22$, $SD = 6.72$), took part in this study on a voluntary basis. Participants were randomly assigned to one of the four experimental conditions: stable volatility with trend (ST), stable volatility without trend (SN), variable volatility with trend (VT), and variable volatility without trend (VN).

Task

All participants completed a restless four-armed bandit task. On each of 200 trials, they were presented with four decks of cards (arms) and asked to draw a card from one of them. After choosing an arm, they were informed of the reward obtained from playing the arm. For each arm $j = 1, \ldots, 4$, the reward $R_j(t)$ on trial t was randomly drawn from a Normal distribution with a mean $\mu_j(t)$ which varied randomly and independently according to a random walk. More precisely, the rewards were generated according to the following schedule:

$$\begin{aligned} R_j(t) &= \mu_j(t) + \epsilon_j(t) & \epsilon_j(t) &\sim N(0, \sigma_\epsilon) \\ \mu_j(t) &= \lambda \mu_j(t-1) + \kappa_j + \zeta_j(t) & \zeta_j(t) &\sim N(0, \sigma_\zeta(t)) \end{aligned} \tag{3}$$

The averages of the arms were initialized as $\mu_j(1) = -60, -20, 20, 60$ for arm $j = 1, \ldots, 4$ respectively. A decay parameter $\lambda = .9836$ was used so that values remained closer to 0 than with a pure random walk. In the ST and VT conditions, the trend parameter had values $\kappa_j = 0.5, 0.5, -0.5, -0.5$ for arms 1 to 4 respectively. In the SN and VN conditions, the values were $\kappa_j = 0$ for all arms. The reward error variance was $\sigma^2_\epsilon = 16$ in all conditions. The innovation variance was $\sigma^2_\zeta(t) = 16$ on all trials in the stable volatility conditions (SN and ST). In the variable volatility conditions, the innovation variance was the same on half of the trials. On trials 51-100 and 151-200, the innovation variance was increased to $\sigma^2_\zeta(t) = 256$. The schedule in Equation 3 was used to generate four sets of reward sequences in each condition, matching the seed in the random number generated used over conditions. One example of the resulting rewards in the four conditions is provided in Figure 2. A similar schedule was used by Daw et al. (2006), but they did not include trends or changes in volatility and used only posive rewards.

Procedure

Participants completed the 200 trials of the task individually at their own pace in a single session. At the start of the task, participants were told that they would be presented with four decks of cards and that their task was to select on each trial a card from any deck they chose. The only goal of the game was to win as many points as possible. Participants were informed that some decks may be better than others, but that the amount they tend to give may vary, so that this can change. They were not informed of the total number of trials in the task.

After reading the instructions, participants started the experimental task. On each trial, they were presented with the four decks of cards and selected a card from one deck via a mouse click. The number of points won or lost was then displayed for 1.5 seconds, along with either a smiley or a frowning face for wins or losses respectively. Throughout the task, a counter displayed the total points obtained thus far.

Behavioural results

One participant (age = 21) in the SN condition was excluded from further analysis as she only chose one arm throughout the whole task. All other participants played each arm at least once.

Performance

Given the differences between the conditions in obtainable reward magnitudes (see e.g. Figure 2), total reward obtained is not an unambiguous measure of performance. We therefore chose to focus on whether, on a given trial, the arm with the maximum reward was chosen, which we will refer to as an advantageous choice. Average proportions of advantageous choices, by block (4 blocks of 50 trials each) and condition, are depicted in Figure 3. Choice behaviour was analysed with a generalized linear mixed-effects model, using a binomial distribution for the number of advantageous choices in each block. In addition to fixed effects for Block, Volatility, and Trend, subject-specific random intercepts were included. Note that this model is structurally similar to a repeated-measures ANOVA, but takes into account the non-normal distribution of the number of advantageous choices. This analysis showed a significant main effect of Block, $\chi^2(3) = 295.11$, $p < .001$. Averaging over conditions, the proportion of advantageous choices increased from block 1 to block 3, while there was a small decrease from block 3 to block 4. In addition, there was a significant Volatility by Block interaction, $\chi^2(3) = 147.83$, $p < .001$, as well as a significant Trend by Block interaction, $\chi^2(3) = 86.14$, $p < .001$. Post-hoc comparisons showed that in block 2, performance in the stable volatility conditions was significantly better than in the variable volatility conditions ($p = .019$), and performance in the no trend conditions was significantly better than in the trend conditions ($p = .004$). In block 3, the reverse was true, with performance better in the variable volatility conditions ($p < .001$) and in the trend conditions ($p = .029$). In the remaining blocks, there was no effect of Volatility or Trend. The effects of Volatility and Trend on performance are likely due to their effect on the discriminability between the arms in terms of their average rewards. For instance, while high volatility may hinder discrimination between the arms due to large trial-by-trial variation in average rewards, when volatility reduces again in block 3, the arms are actually more discriminable than in the stable volatility conditions, as high volatility has pushed the means further apart.

Switching

```
## Warning in RET$pfunction("adjusted", ...):  Completion with error >
abseps
## Warning in RET$pfunction("adjusted", ...):  Completion with error >
abseps
## Warning in RET$pfunction("adjusted", ...):  Completion with error >
abseps
## Warning in RET$pfunction("adjusted", ...):  Completion with error >
abseps
```

For an initial analysis of explorative behaviour, we looked at how often people switched between arms (how often they chose a different arm than on the immediately preceding trial). Average switching proportions, by block and condition, are depicted in Figure 3. Switching behaviour was analysed with a similar generalized linear mixed-effects model as for the advantageous choices. This analysis showed a significant main effect of Block, $\chi^2(3) = 633.88$, $p < .001$, as well as a Volatility by Block interaction, $\chi^2(3) = 26.44$, $p < .001$, a Trend by Block interaction, $\chi^2(3) = 44.1$, $p < .001$, and a three-way interaction between Volatility, Trend, and Block, $\chi^2(3) = 16.17$, $p = .001$. No other effects were significant. Post-hoc analysis did not show any significant differences between pairs of conditions within each block. Comparisons of consecutive blocks within each condition showed that in the SN condition, there was a significant decrease in switching from block 2 to 3. In the VN condition, switching decreased from block 1 to 2 and from block 2 and 3, while there was an increase from block 3 to 4. In the ST and VT condition there was a decrease in switching from block 1 to 2 and from block 2 to 3. For all these comparisons, $p < .001$. As for the number of advantageous choices, this analysis indicates that there were no general effects of volatility or trend on switching behaviour. However, these manipulations did affect how switching behaviour developed during the task.

Of particular interest is whether participants in the variable volatility conditions show increased exploration in the blocks with high volatility. Focussing on switching behaviour in block 3 and 4, we see an increase from block 3 to block 4 in the variable volatility conditions, $\chi^2(1) = 36.15$, $p < .001$, while there is no difference in the stable volatility conditions, $\chi^2(1) = 1.92$, $p = .17$.

Modelling exploration and exploitation

Switching between arms is only a rough measure of exploration, as one can switch to another arm because one believes that one is now optimal (exploitation) or to gain more information about it (exploration). Indeed, if one switches back to the favoured arm after an explorative choice, this is clearly not an explorative choice itself, although the preceding choice may have been. We therefore use computational modelling to gain more insight into explorative decisions. In all models considered here, $u(t)$, the utility of the reward $R(t)$ received on trial t, is assumed to be described through the value function of Prospect Theory (Tversky & Kahneman, 1992; see also Ahn, Busemeyer, Wagenmakers, & Stout, 2008):

$$u(t) = \begin{cases} R(t)^\alpha & \text{if } R(t) \geq 0 \\ -\lambda |R(t)|^\alpha & \text{if } R(t) < 0 \end{cases}$$

where the parameter $\alpha > 0$ determines the shape of the utility function: when $\alpha < 1$, the curve is concave for gains (risk aversion) and convex for losses (risk seeking). The parameter $\lambda \geq 0$ can account for loss aversion: when $\lambda > 1$, a loss of x points has a larger negative utility than a win of x points has a positive utility.

After receiving a reward on trial t, participants are assumed to update their expectancies $E_j(t + 1)$ regarding the utility they will receive when choosing arm j on trial $t + 1$. We consider three possible mechanisms through which these expectancies are updated: Bayesian updating, the delta rule, and a decay rule.

Bayesian updating. This model-based learning strategy assumes the utility of arms is determined by a Gaussian process as in Equation 1. Optimal Bayesian inference regarding mean utilities is implemented by the Kalman filter (Kalman, 1960; see also Daw et al., 2006):

$$E_j(t) = E_j(t-1) + \delta_j(t)K_j(t)[u(t) - E_j(t-1)] \tag{4}$$

where $\delta_j(t) = 1$ if arm j was chosen on trial t, and 0 otherwise. The "Kalman gain" term is computed as

$$K_j(t) = \frac{S_j(t-1) + \sigma^2_\zeta}{S_j(t-1) + \sigma^2_\zeta + \sigma^2_\epsilon}$$

where $S_j(t)$ is the variance of the posterior distribution of the mean utility, computed as

$$S_j(t) = [1 - \delta_j(t)K_j(t)][S_j(t-1) + \sigma^2_\zeta] \tag{5}$$

Prior means and variances were initialized to $E_j(0) = 0$ and $S_j(0) = 1000$. For simplicity, we did not consider a model which learns possible trends or the level of volatility.

Delta rule. A popular model-free alternative to Bayesian inference is the delta rule (Gluck & Bower, 1988; Yechiam & Busemeyer, 2005):

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

The main difference between this rule and Bayesian updating (Equation 4) is that the learning rate $0 \leq \eta \leq 1$ is fixed in the delta rule, while the "learning rate" $K_j(t)$ in Bayesian updating depends on the current level of uncertainty.

Decay rule. While the two previous learning rules assume only the expectancy of the currently chosen arm is updated, according to the (model-free) decay rule (Ahn et al., 2008; Erev & Roth, 1998), expectancies of unchosen arms decay towards 0:

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

where the decay parameter $0 \leq \eta \leq 1$.

Choice rules

Choice rules describe how the expectancies are used to make a choice $C(t)$ between the arms. We consider six choice rules, the "probability of maximum utility" rule described in the introduction, and three more heuristic rules popular in reinforcement learning.

ϵ-greedy. The ϵ-greedy choice rule (Sutton & Barto, 1998) exploits the arm with the maximum expectancy with probability 1 - ϵ, and with probability ϵ chooses randomly from the remaining arms:

$$P(C(t) = j) = \begin{cases} 1 - \epsilon & \text{if } E_j(t) > E_k(t), \quad \forall k \neq j \\ \epsilon/3 & \text{otherwise} \end{cases}$$

Softmax (SM). The softmax choice rule (Luce, 1959; Sutton & Barto, 1998) can vary gradually between a pure exploitation (maximisation) and pure exploration through an inverse temperature parameter $\theta(t)$:

$$P(C(t) = j) = \frac{\exp\{\theta(t)E_j(t)\}}{\sum_{k=1}^{4} \exp\{\theta(t)E_k(t)\}}$$

The temperature is constant in the fixed softmax (SM$_f$) models: $\theta(t) = \theta_0$, with $\theta_0 \geq 0$. In the dynamic softmax (SM$_d$) models (Busemeyer & Stout, 2002), the temperature can increase or decrease over trials according to the schedule $\theta(t) = [t/10]^{\theta_0}$. In this case, θ_0 can take values along the whole real line.

Softmax with exploration bonus (SMEB). The softmax with exploration bonus choice rule (Daw et al., 2006) increases exploration compared to the standard softmax rule by adding an "exploration bonus" term, $\beta_j(t)$:

$$P(C(t) = j) = \frac{\exp\{\theta_0 E_j(t) + \beta_j(t)\}}{\sum_{k=1}^{4} \exp\{\theta_0 E_k(t) + \beta_k(t)\}}$$

The exploration bonus increases with uncertainty. For the Kalman filter model, we use the standard deviation of the prior distribution of mean utility: $\beta_j(t) = \beta_0\sqrt{S_j(t) + \hat{\sigma}_\zeta^2}$, with $\beta_0 > 0$, $\hat{\sigma}_\zeta^2$ the innovation variance assumed by the learner, and $S_j(t)$ computed as in

Equation 5. As the Delta and Decay models do not provide measures of uncertainty, we used a simple heuristic according to which the uncertainty increases linearly with the number of trials since a particular arm was last observed: $\beta_j(t) = \beta_0[t - T_j]$, where T_j is the last trial before the current trial t in which arm j was chosen.

Probability of maximum utility (PMU). The probability that an arm provides a higher utility than any of the other arms can be computed from the prior predictive reward distributions (Equation 2) as the probability that all pairwise differences between the reward of an arm and the rewards of the other arms are greater than or equal to 0. For the current task and generative model in Equation 1, there are three such pairwise differences scores for each arm. These pairwise difference scores follow a multivariate Normal distribution. Hence, the probability that arm j is chosen on trial t is

$$\begin{aligned} P(C(t) = j) &= P(\forall k : u_j(t) \geq u_k(t)) \\ &= \int_0^\infty \Phi(\mathbf{M}_j(t), \mathbf{H}_j(t)) \end{aligned}$$

where Φ is the multivariate Normal density function with mean vector

$$\mathbf{M}_j(t) = \mathbf{A}_j \mathbf{E}(t)$$

and covariance matrix

$$\mathbf{H}_j(t) = \mathbf{A}_j \text{diag}(\mathbf{S}(t) + \hat{\sigma}_\epsilon^2) \mathbf{A}_j^T.$$

Here, $\mathbf{E}(t)$ is a vector with the prior expectancies on trial t, $\text{diag}(\mathbf{S}(t) + \hat{\sigma}_\epsilon^2)$ is a diagonal matrix with the variances of the prior predictive reward distribution for each arm, which equal $S_j(t) + \hat{\sigma}_\epsilon^2$, where $\hat{\sigma}_\epsilon^2$ is the observation variance assumed by the learner. The matrix $\mathbf{A}_j$ computes the pairwise differences between arm j and the other arms. E.g., for arm 1, this matrix is given as

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$$

Note that this choice rule requires prior predictive distributions for the rewards associated with each arm. As such distributions are not directly available from the model-free delta and decay learning rules, this choice rule only applies to Bayesian learning.

Model estimation and inference

For each individual participant, model parameters were estimated by maximum likelihood using the Nelder-Mead simplex algorithm implemented in the `optim` function in R (R Core Team, 2014). To evaluate the fit of the models, we computed the Akaike (AIC) and Schwartz (BIC) information criteria, reported as difference scores between a non-learning null model, which assumes a fixed probability of choosing each arm[1], and the model of interest (cf. Yechiam & Busemeyer, 2005). For these difference scores, negative values of Δ(AIC) and Δ(BIC) indicate that the model fitted worse than the null model, while increasing positive values indicate better fit. Finally, we computed Akaike and Schwarz weights, w(AIC) and w(BIC) (cf. Wagenmakers & Farrell, 2004). Schwarz (BIC) weights approximate the posterior probability of the models (assuming equal prior probability). Similarly, Akaike (AIC) weights can be interpreted as reflecting the probability that, given the observed data, a candidate model is the best model in the AIC sense (it minimizes the Kullback-Leibler discrepancy between the model and the true model) in the set of models under consideration.

Modelling results

Table 1 contains the fit measures for all models. Focussing first on the Δ(AIC) and Δ(BIC) scores, which assess how well each model performs compared to the (non-learning) null model, we see that on average the best fitting model is decay learning with a fixed

[1] The null model used was a simple multinomial model in which, on each trial, the arm choice is assumed to be an independent random draw from a multinomial distribution. The probabilities of this distribution were estimated for each participant and the null model has three parameters (the probability for three arms; the probability of the remaining arm is given from these).

softmax choice rule (Decay-SM$_f$). In general, the Decay rule always performs better (on average) than the other two learning rules. This is a common finding which is likely due to the ability of this rule to capture people's tendency to repeat previous choices (Ahn et al., 2008). While the delta learning rule performs on average a little better than Bayesian updating with the Kalman filter, the differences between these learning rules are less marked. Out of the different choice rules, the ϵ-greedy rule clearly performs the worst, while the fixed Softmax (SM$_f$) rule performs best for each learning rule. When we look at the number of participants who are best fit by each model, a different picture arises: Bayesian updating with the probability of maximum utility choice rule (Bayesian-PMU) fits more participants best than any of the other models. This is also the model with the highest average Akaike and Schwarz weights. As such, the evidence for this model is more marked than the evidence for the other models.

The discrepancy between the results for the Δ(AIC) and Δ(BIC) scores on the one hand, and the Akaike and Schwarz weights on the other, is due to the fact that when the Bayesian-PMU model fits best, it fits decidedly better than the other models, resulting in weights close to 1. For participants with a different best fitting model, the differences between the AIC and BIC values of the next best-fitting models are generally smaller, resulting in less extreme weights.

Median parameter estimates for the models are given in Table 2. We will not discuss these fully, but rather focus on the Bayes-PMU, Delta-SM$_d$ and Decay-SM$_f$ models, the version of each learning rule which fitted most participants best. For all three models, the α parameter of the utility function reflects risk-aversion for wins and risk seeking for losses. However, in the Bayes-PMU and Delta-SM$_d$ models, there is little evidence of risk-aversion (the median parameter estimates of λ are smaller than 1), while the Decay-SM$_f$ model does indicate loss-aversion. This difference is likely to due to the fact that the expectancy of arms with average losses will decay towards 0 when unchosen, making them relatively more favourable over time. A large value for λ will push the expectancy of losing arms further from 0, slowing this process somewhat. For the Bayes-PMU model, the median estimates indicate that participants generally acted as though σ^2_ζ, the innovation variance, was larger than σ^2_ϵ, the observation variance.The larger the assumed innovation variance is compared to the observation variance, the more the uncertainy increases for unchosen arms, resulting in relatively more exploration[2]. For the Delta-SM$_d$ model, the median value of θ_0 indicates an increase from $\theta = 0.79$ on trial 1 to $\theta = 1.35$ on trial 200. In the Decay-MS$_f$ model, θ is constant with a median value of $\theta = 0.38$.

While the parameter estimates provide some information regarding the level of exploration according to the different models, it can be difficult to assess their precise effect on exploration and to make cross-model comparisons. For instance, the effect of the softmax inverse temperature parameter ($\theta(t)$) on exploration depends on the relative expectancies $E_j(t)$, which in turn depend on the shape of the utility function (as determined by α and λ) and the learning rate (η). For a more direct assessment of explorative behaviour, we can compare the models in terms of their predicted probability that a participant chooses a non-maximising arm (i.e. an arm for which the expected reward is lower than the expected reward of at least one other arm). These probabilities, computed for each participant according to their best fitting parameters and then averaged, are depicted in Figure 4. While the models show a reasonable level of agreement in terms of the predicted level of exploration, there are some notable differences. In general, the Bayesian-PMU model predicts the highest level of exploration, followed by the Delta-SM$_d$ model, with the Decay-SM$_f$ model predicting the least amount of exploration. The differences tend to be somewhat larger in the stable volatility conditions. The level of predicted exploration seems itself more volatile in the variable volatility conditions as compared to the stable volatility conditions. In the blocks with high volatility, there is a

[2] Note that the median variance parameters differ substantially between the different versions of the Bayesian updating model. In the current models, in which the predicted choices depend mostly on the relative expectancies, the relative value of σ_ζ compared to σ_ϵ matters more than the absolute values of these parameters.

rapid initial rise in predicted exploration, followed by a relatively rapid decrease, and then rising again towards the end of the block. Interestingly, the models show relatively little disagreement in these blocks, and hence the dynamics of exploration may be more driven by the volatility in the rewards themselves, rather than the resulting increase in uncertainty. However, we should note that the Bayesian model was a relatively simple one which assumed the innovation variance was constant throughout the task. Although analytically intractible, a model which can detect and adapt to changes in the level of volatility, such as a switching Kalman filter model (Kim, 1994), may be a more realistic one for this task. For such models, a rapid increase in volatilty would directly result in a rapid increase in uncertainty about the rewards associated with the arms. There is already some evidence that people adjust their learning to changes in volatility (Behrens, Woolrich, Walton, & Rushworth, 2007), suggesting that participants in our experiment may have done the same. We intend to explore this possibility and the precise manner in which people can achieve such meta-learning in future research.

Conclusion

We found evidence that a substantial proportion of our participants explored to reduce uncertainty in their beliefs. This was evident from the finding that participants switched arms more in periods of increased volatility, and from computational modelling where a Bayesian model that balances exploration and exploitation according to the probability that arms will provide the maximum utility fitted the largest number of individual participants and, on average, had the highest posterior probability in a set of competing models. This supports previous findings by Knox et al. (2012) in a more constrained task, but contrasts with the findings of Daw et al. (2006) who found no influence of uncertainty in a task similar to the one used here. However, Daw et al. only considered a Bayesian learning model with a Softmax with exploration bonus (SMEB) choice rule to model the influence of uncertainty, for which we also found little evidence.

The effect of uncertainty on explorative decisions seems thus more nuanced (cf. Cohen et al., 2007), affecting choices through its effect on predicted distributions of the utility associated with the different options.

Acknowledgements

We thank James Carragher, Belle Cartwright, Melvin Wan and Tiffany Yung for their assistance in collecting the data.

References

Acuna, D., & Schrater, P. (2008). Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 200–300).

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

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

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

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*, 253–262.

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 B*, *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.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, *88*, 848–881.

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Granmo, O.-C., & Berg, S. (2010). Solving non-stationary bandit problems by random sampling from sibling kalman filters. In N. García-Pedrajas, F. Herrera, C. Fyfe, J. M. Benítez, & M. Ali (Eds.), *Trends in applied intelligent systems* (Vol. 6098, p. 199-208). Berlin: Springer.

Gupta, N., Granmo, O.-C., & Agrawala, A. (2011). Thompson sampling for dynamic multi-armed bandits. In *2011 10th international conference on machine learning and applications and workshops (ICMLA)* (Vol. 1, p. 484-489).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering*, *82*, 35–45.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering*, *83*, 95–108.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, *60*, 1–22.

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*, 1–12.

Luce, D. R. (1959). *Individual choice behavior.* New York: Wiley.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research*, *24*, 293–305.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

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: MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Viappiani, P. (2013). Thompson sampling for bayesian bandits with resets. In P. Perny, M. Pirlot, & A. TsoukiÃăs (Eds.), *Algorithmic decision theory* (Vol. 8176, p. 399-410). Springer Berlin Heidelberg.

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

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, *25*, 287–298.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*, 387–402.

Yi, M. S., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, *2*, 5.

Table 1

Modelling results. Values of $\Delta(\cdot)$ and $w(\cdot)$ are averages and the standard deviation is given in parentheses. Values of $n(\cdot)$ are the total number of participants best fit by the corresponding model.

Learning	Choice	Δ(AIC)	w(AIC)	n(AIC)	Δ(BIC)	w(BIC)	n(BIC)
Bayesian	ϵ-greedy	192.16 (96.71)	0 (0)	0	185.57 (96.71)	0 (0)	0
	SM$_f$	237.37 (90.07)	0.04 (0.09)	2	230.78 (90.07)	0.01 (0.03)	0
	SM$_d$	236.22 (90.09)	0.06 (0.17)	2	229.62 (90.09)	0.03 (0.13)	2
	SMEB	226.71 (94.8)	0.01 (0.02)	0	216.81 (94.8)	0 (0)	0
	PMU	223.43 (107.36)	0.25 (0.4)	21	220.13 (107.36)	0.26 (0.41)	21
Delta	ϵ-greedy	194.7 (94.97)	0 (0)	0	191.4 (94.97)	0 (0)	0
	SM$_f$	237.5 (91.7)	0.06 (0.13)	6	234.2 (91.7)	0.09 (0.2)	8
	SM$_d$	236.68 (91.4)	0.11 (0.24)	12	233.38 (91.4)	0.15 (0.3)	12
	SMEB	231.39 (93.4)	0.02 (0.05)	0	224.8 (93.4)	0.01 (0.01)	0
Decay	ϵ-greedy	206.11 (95.52)	0.01 (0.11)	1	202.81 (95.52)	0.01 (0.11)	1
	SM$_f$	244.55 (91.69)	0.19 (0.28)	18	241.25 (91.69)	0.22 (0.33)	18
	SM$_d$	240.59 (95.7)	0.2 (0.33)	17	237.29 (95.7)	0.22 (0.34)	17
	SMEB	240.5 (91.99)	0.05 (0.08)	0	233.9 (91.99)	0.01 (0.02)	0

Table 2

Median parameter estimates of the fitted models.

Choice	σ_ξ	σ_ϵ	η	ϵ	θ_0	β_0	α	λ
Bayesian updating								
ϵ-greedy	5.57	7.57		0.14			0.63	2.15
SM$_f$	112.72	5			0.5		0.57	0.18
SM$_d$	244.79	3.03			0.11		0.35	0
SMEB	2745.46	588.75			0.33	0.54	0.2	0
PMU	1.24	0.69					0.61	0.22
Delta rule								
ϵ-greedy			0.76	0.14			0.95	1.51
SM$_f$			0.89		0.5		0.55	0.17
SM$_d$			0.87		0.1		0.35	0
SMEB			1		0.48	0.54	0.15	0
Decay rule								
ϵ-greedy			0.52	0.13			1.28	2.4
SM$_f$			0.64		0.38		0.36	2.06
SM$_d$			0.52		0.11		0.08	1.01
SMEB			0.64		0.48	0.25	1.73	0

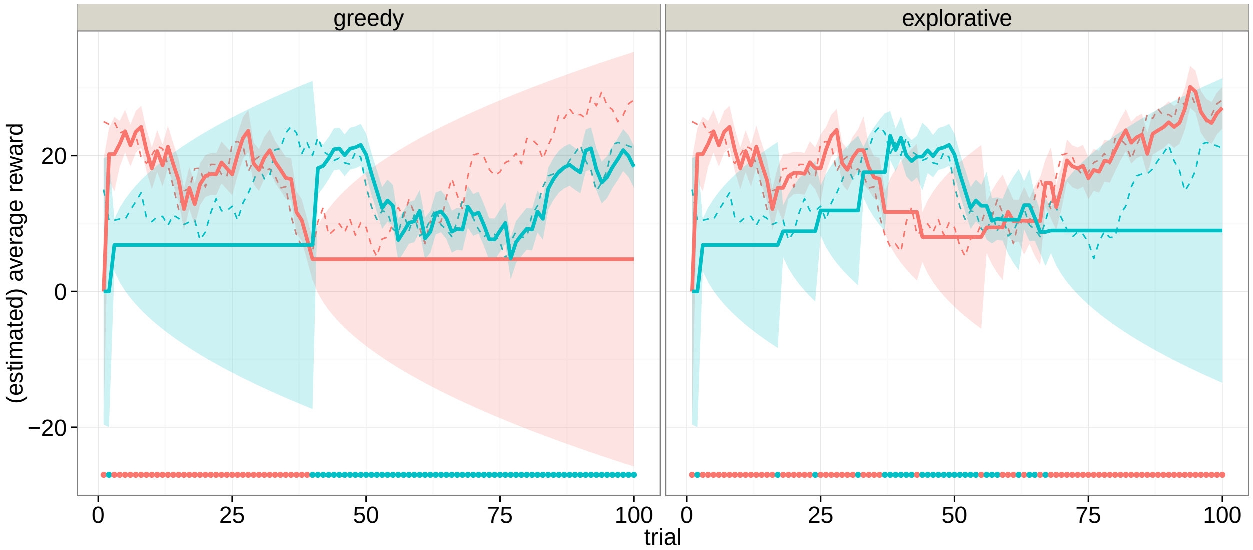

Figure 1. Learning and decision making in a restless two-armed bandit task. Two arms (blue and red) have changing average rewards (broken lines), generated according to Equation 1. On each trial, an agent chooses an arm (dots at the bottom of the graphs), observes the associated reward, and then updates her belief about the average reward for that arm. These posterior beliefs form the basis of the prior belief on the next trial (solid lines show the prior means and areas the 95% highest density intervals of the prior distribution). The "greedy" agent (left panel) always chooses the arm with the highest expected reward. After sampling once from both arms, she always chooses the one with the highest expected reward until the prior mean falls below the prior mean of the unchosen arm. A problem with this strategy is that it ignores the uncertainty in the prior distribution, which increases for unchosen arms due to the innovations $\zeta_j(t)$. After not choosing an arm for a prolonged period, the probability that the mean reward of this arm is higher than the mean reward of the chosen arm can become substantial. The "explorative" agent (right panel) bases her choices on this probability. On each trial she chooses an arm randomly according to the probability that this arm will provide the highest reward in the set of arms. This clearly gives better results than the greedy strategy and the agent mostly chooses the arm with the highest average reward.

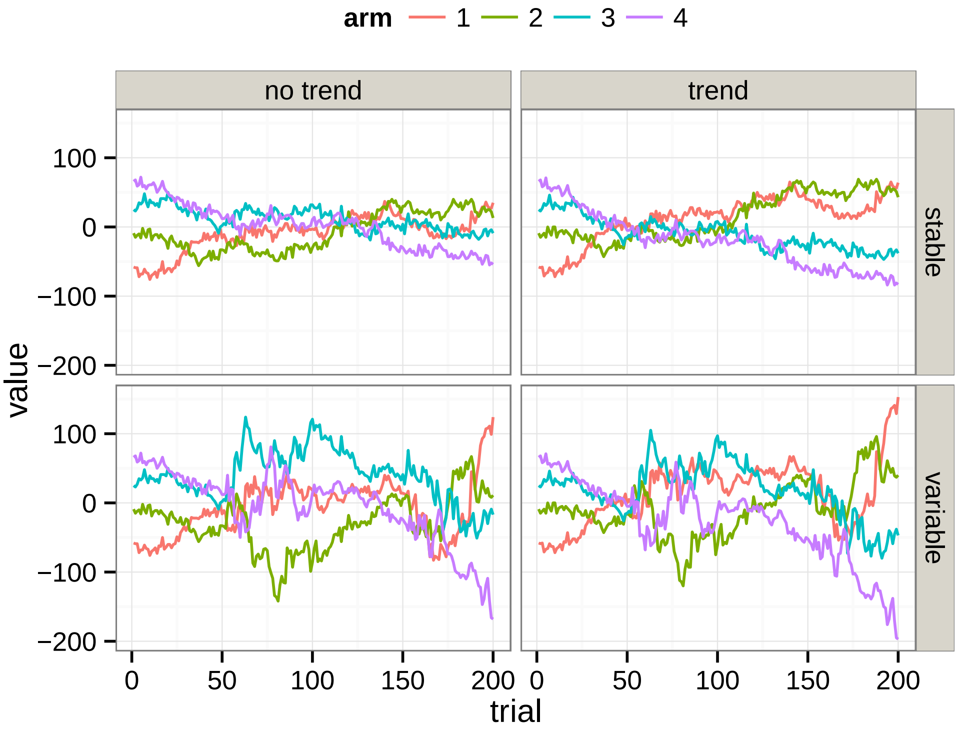

Figure 2. Example rewards in the four-armed restless bandit task.

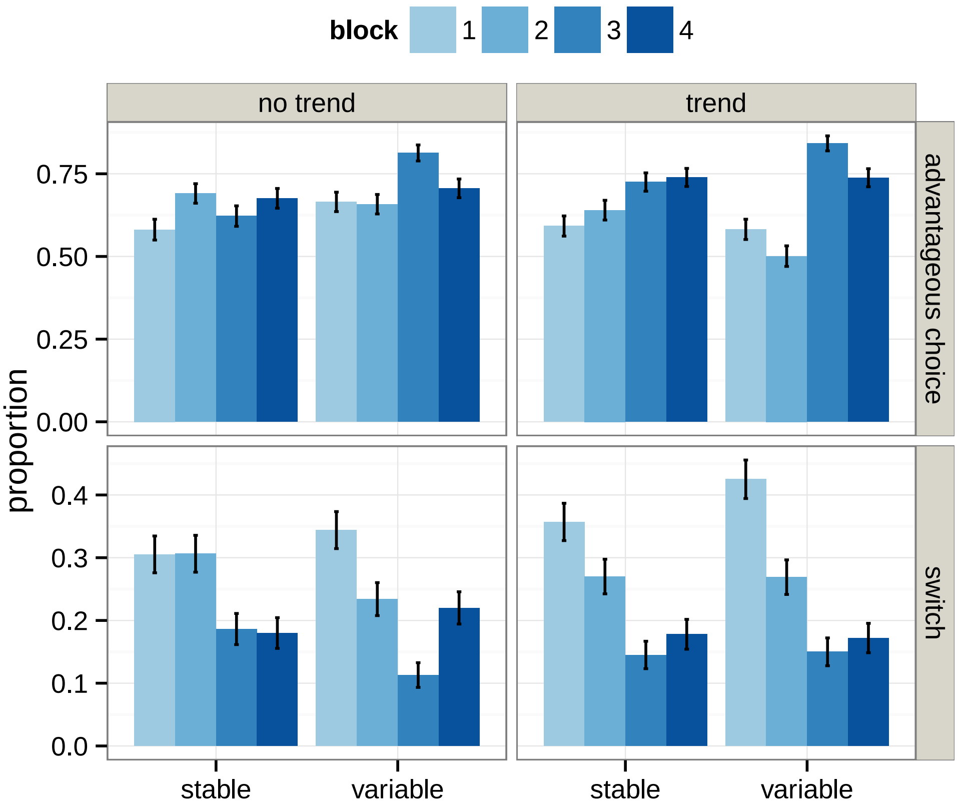

Figure 3. Proportion of advantageous choices and switches by block (50 trials each) and condition.

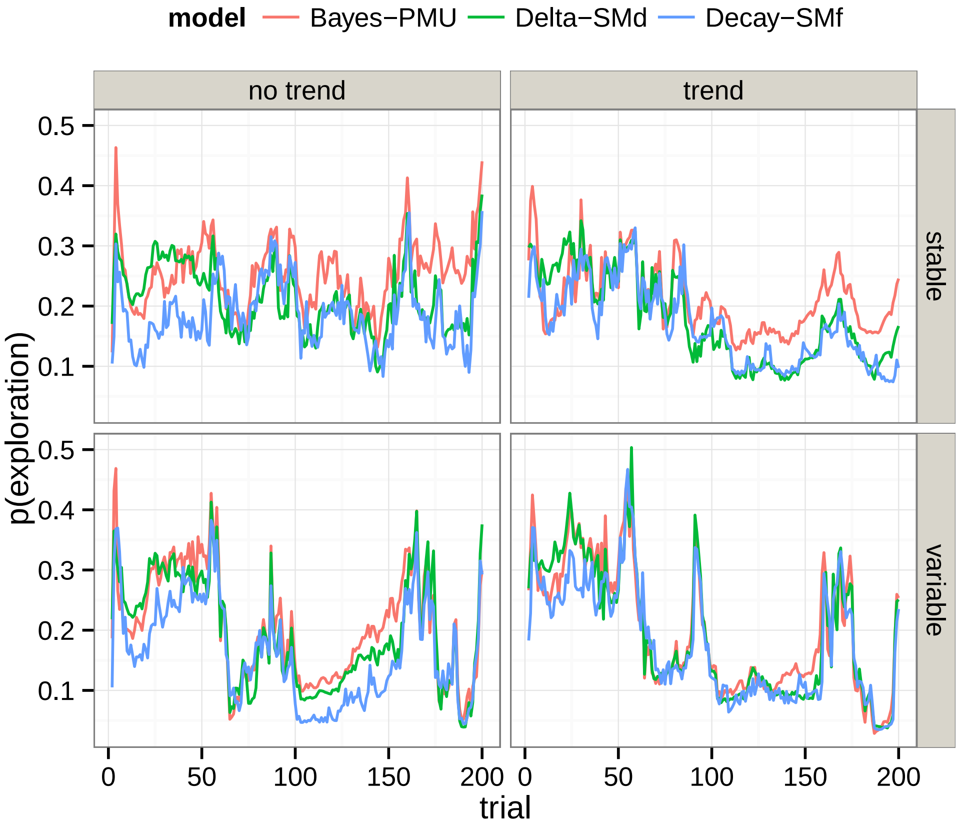

Figure 4. Average probability of an explorative choice by condition according to the Bayes-PMU, Delta-SM$_d$, and Decay-SM$_f$ models.