Dissociating different forms of random exploration

Magda Dubois (magda.dubois.18@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Johanna Habicht (johanna.habicht.15@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Jochen Michely (j.michely@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Rani Moran (r.moran@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Ray Dolan (r.dolan@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Tobias Hauser (t.hauser@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London, UK

Abstract

The arbitration between making the most out of current knowledge (exploitation) and gathering new knowledge (exploration) is central to decision making. It has been proposed that humans use two distinct strategies for exploration. Directed exploration which targets high information conveying options, and random exploration which assigns weights to options relative to their value estimates. Here we suggest that humans use a third strategy, 'tabula rasa' exploration, which in contrast to the traditional 'random' exploration ignores all prior knowledge about the world. We tested this hypothesis using a novel three-bandit task in which the expected values, the prior information and the time horizon is manipulated. Using computational modeling, we found evidence for tabula rasa exploration in addition to directed and random exploration.

Keywords: explore-exploit dilemma; model comparison

When seeking to optimise rewards over time in a finite decision space, a thinking agent will most certainly face the explore-exploit dilemma. She will need to decide whether to go for a known option with the highest expected reward (exploitation) or for lesser known options (exploration) to make sure to not miss out on possibly even higher rewards. Although humans solve this dilemma on a daily basis, the mechanisms involved are still not fully understood. Exploration-exploitation decisions can be studied using the multi-armed bandit problem in which a gambler has to choose how to play from slot machines in order to maximize her reward (Sutton & Barto, 1998). By varying the time horizon (i.e. the number of choices that could be made in the future) humans modulate their exploration behaviour using two distinct strategies: directed and random (Wilson, Geana, White, Ludvig, & Cohen, 2014). Directed exploration is the idea that humans are biased towards highly informative options. This is commonly implemented by adding an 'information' bonus to the expected reward of an option, for example using the Upper Confidence Bound (UCB) algorithm. Random exploration is based on the assumption that humans inject stochasticity in their decision (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), which influences the weight of the relative value estimate, e.g. via softmax (Sutton & Barto, 1998) or Thompson sampling (Thompson, 1933) . In addition, recent findings have suggested that a hybrid model combining a stochastic UCB model (UCB + softmax) with a Thompson sampling model is best accounted for human exploration behaviour (Gershman, 2018). Both those strategies take into account some knowledge of the world. Here we suggest that humans can also use a third type of exploration strategy ('tabula rasa') which is completely agnostic to the environment. This is a reflection of what is known in reinforcement learning as ε-greedy algorithm (Sutton & Barto, 1998). Tabula rasa exploration thus predicts a constant probability of exploration independent of the states by occasionally substituting the greedy action with a random action. To test this hypothesis we developed a novel task and used computational modelling to investigate these three forms of exploration.

Methods

Sixty healthy volunteers (30 female, 30 male; mean age = 23.30) were recruited. All participants provided written informed consent and the study was approved by the University College London research ethics committee.

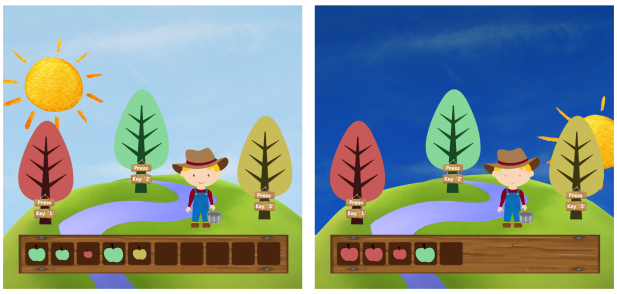

Figure 1: Maggie's Farm task. Left: Long horizon. In this trial red is tree D, green is tree A and yellow is tree B. Right: Short horizon. In this trial red is tree A, green is tree B and yellow is tree C.

Participants had to choose between trees that produced apples with different sizes in two different horizon condition (Figure 1). They were instructed to collect the biggest apples and that they would receive cash bonus according to their performance. To distinguish between different types of exploration (Wilson et al., 2014), we manipulated the horizon (i.e. number of apples to be picked: 1 in the short horizon, 6 in the long horizon) and within games the mean reward μ (i.e. apple size) and the information I (i.e. apples shown at the beginning of the trial) of each option. Trees were generated from 4 different generative groups:

$$\begin{aligned} &TreeA: \ \mu_A \sim N(5.5,\ 1.4), \quad I_A = 3 \\ &TreeB: \ \mu_B = \mu_A \pm 1 \ or \ 2, \quad I_B = 1 \\ &TreeC: \ \mu_C = (\mu_A \ or \ \mu_B) \pm 1 \ or \ 2, \quad I_C = 0 \\ &TreeD: \ \mu_D = min(\mu_A,\ \mu_B,\ \mu_C) - 1, \quad I_D = 1 \end{aligned} \tag{1}$$

For each type of tree x, the apples were sampled from $N(\mu_x, 0.8)$, bounded to $[2, 10]$, and rounded to the closest integer. On each trial, three trees from different groups were available to choose from. Tree A and Tree B are a reproduction of the 'Horizon task' (Wilson et al., 2014) and allow us to distinguish between random and directed exploration. In this task we added Tree C to examine how directed exploration applies in a complete naive scenario, and a very low-valued option, Tree D, to measure tabula-rasa exploration. There were 50 trials of each tree category combination for both short and long horizon, resulting in a total of 400 trials. We then developed a set of Bayesian generative models, where each model assumed that different characteristics accounted for participants' behavior. The binary coefficients w_1, w_2, w_3, w_4, w_5 are used to indicate which components were included in the different models. Similarly to previous studies (Gershman, 2018), the mean $Q_t(x)$ and variance $\sigma_t^2(x)$ of each tree x are tracked using Kalman filtering (Bishop, 2006) and γ is a horizon-dependent weighting factor on the uncertainty bonus. We computed the sum Λ_t of the value of each tree (the first term), the directed exploration component of UCB (the second term) and the random exploration component of Thompson sampling (the third term) for each tree x. Additionally, as gaining information about an unknown stimuli can be intrinsically rewarding (Dubey & Griffiths, 2017), we added a novelty bonus η (non zero for tree C only). Therefore, for each tree x we have:

$$\begin{aligned} &\Lambda_t(x) = w_1 Q_t(x) \\ &+ w_2 \gamma \sigma_t(x) + w_3 \frac{Q_t(x)}{\sqrt{\sum_x \sigma(x)_t^2}} + w_4 \eta \delta_{[x=C]} \end{aligned} \tag{2}$$

The choice policy was computed using a softmax with a horizon-dependent τ controlling the choice stochasticity (analogous to a choice temperature). To measure tabula-rasa exploration, we extended it with an ε-greedy component using a horizon dependent ε. The probability of choosing tree x is:

$$P(c_t = x) = \frac{\exp(\tau^{-1}\Lambda_t(x))}{\sum_i \exp(\tau^{-1}\Lambda_t(i))} \times (1 - w_5\varepsilon) + w_5 \frac{\varepsilon}{3} \tag{3}$$

We set $[w_4, w_5] = [0, 0]$ and compared Gershman's hybrid model with $[w_1, w_2, w_3] = [1, 1, 1]$ to the stochastic UCB model with $[w_1, w_2, w_3] = [1, 1, 0]$ and the Thompson sampling model with $[w_1, w_2, w_3] = [1, 0, 1]$. We also created 3 extended variants from each of these 3 models by setting $[w_4, w_5] = [1, 0]$, $[w_4, w_5] = [0, 1]$ and $[w_4, w_5] = [1, 1]$, for a total of 12 models. For model selection, we used the Bayesian information criterion (BIC).

Results

The winning model (i.e. the model with the lowest BIC) was found to be the model with $[w_1, w_2, w_3, w_4, w_5] = [1, 1, 0, 1, 1]$, that is a stochastic UCB model with both a novelty bonus and an ε-greedy parameter. Therefore, the sum of directed and random exploration components of each tree x is:

$$\Lambda_t(x) = Q_t(x) + \gamma\sigma_t(x) + \eta(x) \tag{4}$$

And the probability of choosing tree x :

$$P(c_t = x) = \frac{exp(\tau^{-1}\Lambda_t(x))}{\sum_i exp(\tau^{-1}\Lambda_t(i)} \times (1 - \varepsilon) + \frac{\varepsilon}{3} \tag{5}$$

As shown in previous studies (Wilson et al., 2014) , γ and τ reflect directed and random exploration respectively. Following our hypothesis, ε reflects tabula rasa exploration. The novelty of tree C is taken into account by the novelty bonus η. We then averaged the model parameters of each participant. We compared the parameters between horizons (Figure 2) and found strong evidence of γ being modulated by the horizon ($p < 0.001$) with marginally significant horizon effects on τ ($p = 0.054$) and ε ($p = 0.052$).

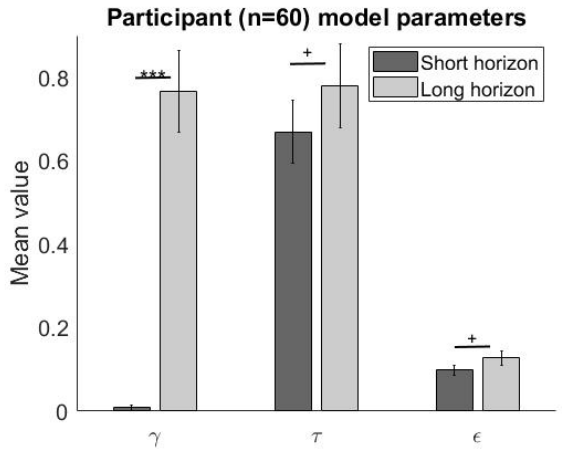

Figure 2: Model parameters reveal effects of decision horizon (*** is $p < 0.01$, + is $p = 0.05$). The parameters γ, τ and ε reflect directed, random and tabula rasa exploration respectively and show (marginally) significant horizon effects

Conclusion

In this study we examined different forms of exploration. Our results reproduce previous research dissociating directed and random exploration and they suggest that participants also engage in another, more agnostic, form of random exploration: tabula rasa exploration.

References

Bishop, C. M. (2006). *Machine Learning and Pattern Recogniton*.

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

Dubey, R., & Griffiths, T. L. (2017). A rational analysis of curiosity. Retrieved from http://arxiv.org/abs/1705.04351

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*.

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples Author. *Biometrika*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the exploreexploit dilemma. *Journal of Experimental Psychology: General*.