Exploration strategies in human decision making

Robert C. Wilson
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
rcw2@princeton.edu

Andra Geana
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
ageana@princeton.edu

John M. White
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
jmw1729@princeton.edu

Elliot A. Ludvig
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
eludvig@princeton.edu

Jonathan D. Cohen
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
jdc@princeton.edu

Abstract

The tradeoff between pursuing a known reward (exploitation) and sampling unknown, potentially better opportunities (exploration) is a fundamental challenge faced by all adaptive organisms. Theories formalize the value of exploration (gathering information) as an information bonus. However, this may be difficult to compute; a simpler alternative is to increase decision noise, driving random exploration. Relatively few studies have characterized human exploratory behavior, and most have failed to find an information bonus, suggesting it relies entirely on random exploration. However, these previous studies have either confounded reward and information or failed to account for baseline levels ambiguity aversion and decision noise. To overcome these limitations, we conducted a sequential choice task that independently manipulated reward, information, and number of choices. Contrary to previous work, we found that humans do show an information bonus when given the opportunity to explore. In addition we found adaptive changes in decision noise consistent with a type of random exploration that is subject to cognitive control.

Keywords: explore-exploit, decision noise, information seeking

Acknowledgements

This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. This research was funded by RCW was funded by a J. Insley Blair Pyne fund award to JDC. JDC, EAL and JMW were supported by NIH Grant #AG024361 to JDC.

1 Introduction

When you go to your favorite restaurant do you always order the same item on the menu, or do you sometimes try something new? Sticking with an old favorite ensures a good meal, but if you are willing to explore you might discover something better. This simple conundrum, deciding between something you know about or trying something new and unknown, is referred to as the exploration-exploitation dilemma [1, 2]. Whether deciding on a meal, a career, or a life partner, this is an important and recurrent problem at all levels of decision-making.

Theoretical accounts suggest two distinct strategies for resolving this dilemma. One is directed exploration [3], in which choices are biased towards ambiguous (and hence more informative) options with an 'information bonus'. The other strategy is random exploration [4], in which choices are biased by internal decision noise.

Directed strategies are often derived from theories of optimal decision-making that ensure the greatest amount of reward in the long run, whereas random strategies reflect simpler heuristics that may be less costly to implement in practice. For example, in our hypothetical restaurant scenario, directed exploration would involve weighing the relative costs and benefits of each meal against the expected gain in information from choosing the unknown option. By contrast, random exploration might involve something as simple as tossing a coin to decide.

Previous work looking for directed exploration has lead to mixed results, with some studies finding it [5, 6, 7] and others not [8, 9, 10]. We believe that these mixed findings are due to two complications that arise in explore-exploit experiments, namely ambiguity aversion and the confounding of information and reward. Ambiguity aversion can be loosely summarized as a 'fear of the unknown' and leads people to irrationally avoid ambiguous options even if choosing them would be beneficial [11]. In explore-exploit situations, this bias runs counter to any information bonus, making directed exploration harder to detect. The confound between reward and information is subtle and arises because participants only receive information about the options they choose. When these choices are made freely, participants, not surprisingly, choose the more rewarding options more often and thus become better informed about them. This correlation between information and reward makes identifying independent effects of information (such as the information bonus) more difficult.

Relatively little work has looked at random exploration. While many studies implicitly use decision noise in their choice models, often this is treated more as a 'fudge factor' to overcome limitations of the model rather than as actual neuronal noise in the participant's head. As with ambiguity aversion, without a measurement of baseline decision noise when the explore-exploit dilemma is absent, it is impossible to dissociate noise in the subject's head from misspecification of the model.

To address these limitations, we examined decision-making behavior in a simple task in which participants were given prior information about each of two options and then allowed to make a series of choices between them. We experimentally manipulated the amount of information that participants were given about each option (i.e., the ambiguity of each), as well as the number of choices they would be allowed to make between those options (i.e. the game horizon). By controlling the information subjects received we removed the confound between reward and information. By varying the horizon we could dissociate baseline levels of decision noise and ambiguity aversion from exploration induced changes. We quantified our analysis by fitting a formal model of the decision-making process to participants behavior. This approach allowed us to determine not only the overall amount of exploratory behavior, but also to dissociate the influence of the two potential strategies discussed above that might be driving such behavior: information seeking (directed exploration) and decision noise (random exploration). We found evidence for both types of exploratory behavior that was adaptively modulated by the opportunity to explore: participants exhibited greater directed and random exploration in games with longer horizons, when exploration yielded information that could be used to achieve greater reward overall.

2 Methods

2.1 Behavioral task

Participants played 150 games of a sequential two-armed bandit gambling task (see figure 1A). In each game they made repeated decisions between two options, A and B. Each option paid out between 0 and 100 points sampled from a Gaussian distribution with a fixed standard deviation of 8 points. The means of the underlying Gaussian were different for the two options, remained stable within a game (sequence of decisions), but changed with each new game. Participants were instructed to maximize the points earned over the entire task, and points were converted to money at the end of the session.

The first four trials of each game were forced-choice trials, in which only one of the options was available for the participant to choose. We used these forced-choice trials to manipulate the relative ambiguity of the two options, by providing the participant with different amounts of information about each before their first free choice. The four forced-choice

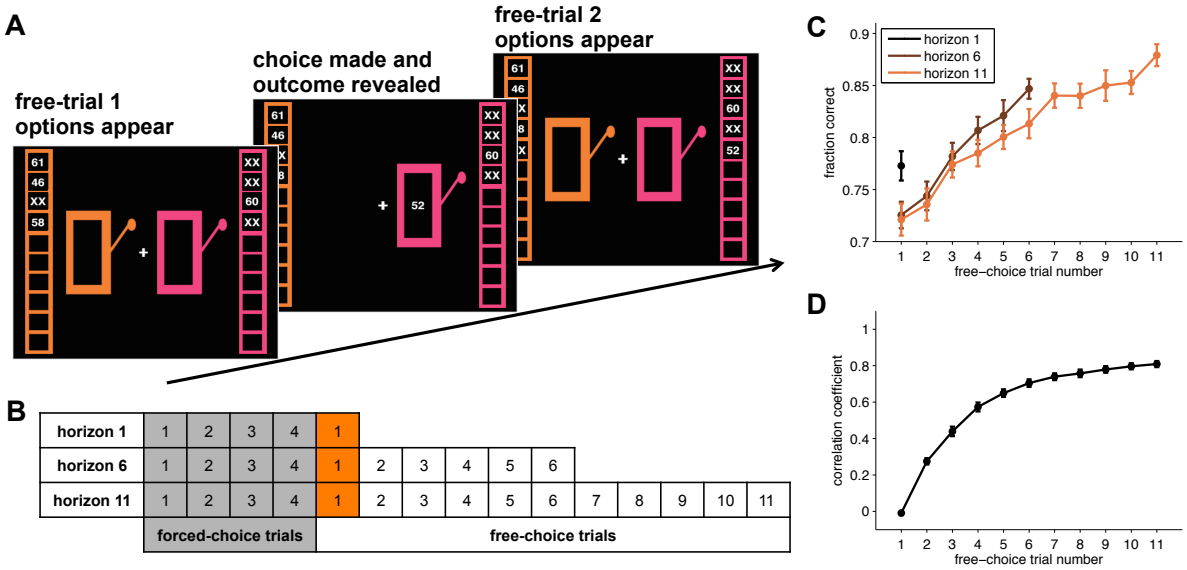

Figure 1: Task design. (A) Example screen shots showing the first free-choice trial in a horizon 6 game, after the choice is made and the start of the second free-choice trial. (B) Schematic showing the different trial types in the three horizon conditions. Each game begins with four forced-choice trials before a sequence of 1, 6 or 11 free-choice trials. Only the first free choice trial is the same across conditions (orange) and is the focus of subsequent analysis. (C) Learning curves showing the fraction of times the correct option (i.e., the option with the higher generative mean) is chosen as a function of free choice trial number for the different horizon conditions. This clearly shows that participants perform at above chance levels and improve as the game progresses. (D) Correlation between the difference in observed means of each option and the difference in the number of times each option has been played as a function of free choice trial number. There is no significant correlation only on the very first trial and a significant correlation thereafter as participants get more information about the more rewarding options that they have selected.

trials set up two ambiguity conditions: unequal ambiguity (or [1 3]) in which one option was forced to be played once and the other three times, and equal ambiguity (or [2 2]) in which each option was forced to be played twice.

Crucially, this manipulation of ambiguity ensured that participants were exposed to a specified amount of information about each option regardless of how rewarding it was. Furthermore, the relative amount of information provided about each option was manipulated independently of the relative difference in their means. Thus on the first free choice (the fifth trial in each game), the difference in the number of times each option had been sampled (and hence its ambiguity) was decorrelated from the difference in mean payout of that option (mean correlation coefficient r(31)= -0.0083, p = 0.56). After the forced-choice trials, participants made either 1, 6 or 11 free choices (figure 1B). At the beginning of each game, the number of upcoming free-choice trials (i.e., the game horizon) was indicated by the length of reward history bars on the sides of the screen (figure 1A), that contained an empty space in which the outcome of each subsequent trial would appear.

2.2 Behavioral model

To avoid the confound between reward and information, our analysis focusses only on the first trial. To model participants choices on this trial, we assumed that they compute a value, Q_a, for each option, a, given as the sum of the expected reward, R_a, information I_a and spatial location s_a; i.e.,

$$Q_a = R_a + AI_a + bs_a + \epsilon_a \tag{1}$$

where A denotes the information bonus, b the spatial bias and ϵ the noise which we assume to be sampled from a logistic distribution with variance $\sigma_{decision}$. R_a was the mean of the examples seen on the forced trials. I_a was defined such that the information bonus was given by the indifference point of the choice curves. This implies that $I_a = 0$ in the equal condition, $I_a = +1/2$ if $a = A$ in the unequal condition and $I_a = -1/2$ if $a \neq A$ in the unequal condition. s_a was defined such that $s_a = +1$ when option A was physically on the right side of the screen and $s_a = -1$ when A was on the left.

Parameters A, $\sigma_{decision}$ and B were estimated using a maximum *a posteriori* fit for each participant in each condition. For the priors, on the information bonus, A, we used a Gaussian distribution with mean 0 and standard deviation 20, for the noise, $\sigma_{decision}$, an exponential distribution with mean 20 and for the spatial bias, B, a uniform distribution between -50 and 50.

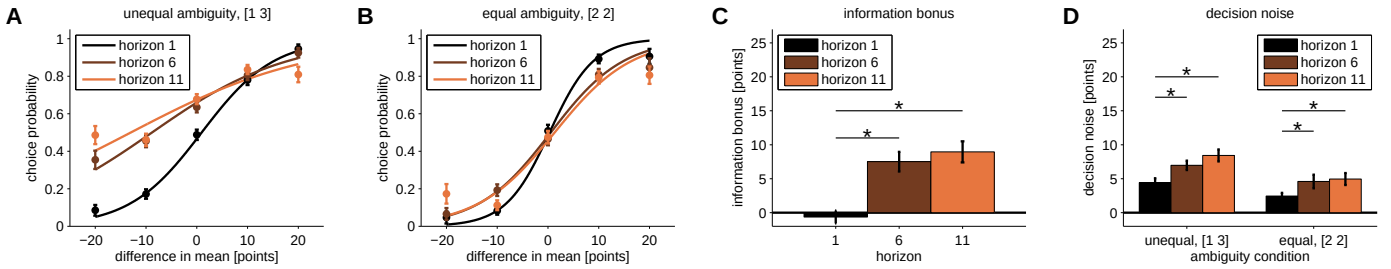

Figure 2: Experimental results. (A) Choice curves for the unequal ambiguity condition showing the fraction of times the more ambiguous option is chosen on the first free-choice trial as a function of the difference in mean between the ambiguous option and the better known option. As the horizon increases the ambiguous option is chosen more often, indicating an information bonus, and the slope of the curves decrease indicating a change in decision noise. (B) Choice curves for the equal ambiguity condition. As the horizon increases there is no change in indifference point, as both options have equal ambiguity, but the slope of the curves still decreases consistent with an increase in decision noise with horizon. (C) The information bonus extracted from the model fits to participant behavior showing an increase in information bonus with horizon. Error bars are s.e.m across participants a *** indicates significant difference (paired t-test) at $p < 10^{-5}$ (horizon 6 vs horizon 1, $t(32) = 5.6$, $p < 10^{-5}$ and horizon 11 vs horizon 1, $t(32) = 5.8$, $p < 10^{-5}$). (D) Participants' decision noise extracted from the model fits showing an increase in decision noise with horizon in both ambiguity conditions. * denotes a significant difference (paired t-test) at $p < 0.05$ (horizon 6 vs horizon 1, $t(32) = 2.6$, $p < 0.05$), ** indicates $p < 0.01$ (horizon 11 vs horizon 1, $t(32) = 3.3$, $p < 0.01$). (E, F) Ambiguity bonus and decision noise for the optimal model. Note that for the optimal model, the decision noise is zero in all conditions.

3 Results

3.1 Model-free analysis

Performance was above chance in all horizon conditions (figure 1C) and improved throughout the game for the longer horizon conditions, indicating that participants understood the task and continued to learn during the free-choice trials. Also, as expected, after the first free-choice trial significant correlations appeared between mean reward and ambiguity (figure 1D). As noted above, this was not the case for the first free-choice trial, thus confirming our ability to examine these factors independently on that trial. Therefore, unless otherwise noted, all of the subsequent analyses presented below focus on the first free-choice trial.

We computed the probability of choosing option A on the first free-choice trial as a function of the difference in sample means observed on the forced plays. By convention, we defined option A to be the more ambiguous one in the unequal ambiguity condition (i.e., the option played only once in the [1 3] forced-choice trials). For the equal ambiguity condition, option A was defined randomly for each trial. The resulting choice curves are plotted in figure 2A and B for the unequal and equal conditions respectively.

In all conditions the probability of choosing option A on the first free choice increased as a function of the difference in mean between them. Furthermore, for that choice, increasing the horizon in the unequal condition increased the probability of choosing the ambiguous option. For example, in horizon 11, even when the mean of the ambiguous option was 8 points lower than the alternative, it was still chosen 50% of the time. This change in the indifference point – the point at which participants were equally likely to choose either option – is consistent with directed exploration driven by an information bonus; that is, for the first free-choice trial in the long horizon conditions, participants behaved as though the ambiguous option had more value. In the equal ambiguity condition, because the same amount of information was available for both options, there was no such information bonus.

In addition to the shift in the indifference point, there was also a change in the slope of the choice curves with horizon (figure 2A and B). Curves at long horizons became increasingly shallow in both ambiguity conditions. This change in slope is consistent with random exploration induced by an increase in decision noise, that is, participants choices became more random and hence less correlated with the difference in means as the horizon increased.

3.2 Model-based analysis

We quantified these observations by fitting the participants choices with the model described in the Methods. The results of these fits are shown in figure 2C. We found a systematic increase in both the information bonus (ANOVA main effect of horizon, $F(2, 32) = 26.9$, $p < 10^{-8}$) and decision noise (ANOVA main effect of horizon, $F(2, 32) = 19.4$, $p < 10^{-6}$) as

the horizon increased. As expected, the fitted values for the spatial bias were not significantly different from zero for any horizon and ambiguity condition.

To test whether the change in information bonus with horizon was consistent with theories of optimal exploration, we computed the optimal information bonus for each horizon condition in the task. In the interests of space we omit the details, but note that these bear a close qualitative resemblance to the estimates of the information bonus for the human participants (figure 2C). Quantitatively, however, the correspondence is weaker and participants appear to exhibit an information bonus that is about twice the optimal value.

4 Discussion

In this paper we presented a simple task designed to dissociate two strategies of exploration: directed and random. Our results suggest that humans use both strategies in our task. Both of these results are surprising.

Our finding that humans use directed exploration with an information bonus runs counter to many previous studies. We believe that this is because these earlier studies fail to take into account baseline ambiguity aversion and subtly confound reward and information. In our experiment, baseline ambiguity aversion is measured in the horizon 1 condition while reward and information are deconfounded by the forced-choice trials. Qualitatively, the presence of directed exploration is in line with theories of optimal decision making [3]. Quantitatively, however, humans tend to use an information bonus that is *larger* than the optimal value. This suggests that while people use directed exploration, they lack the ability to do it perfectly. This is in line with the fact that, even for a simple task such as ours, finding the optimal bonus is computationally demanding [3].

Our findings also show exploration related changes in decision noise in addition to an information bonus. This suggests that humans use decision noise as a separate (simpler, and therefore presumably less costly) means of driving exploration. Such a noise-driven strategy is used to great effect in machine learning [2]) where its simplicity allows it to be applied to situations in which computing the optimal information bonus is intractable. Thus, random exploration driven by decision noise may represent a reasonable adjunct to the theoretically optimal, but costly computations required to quantify the information bonus in many situations. These results also provide empirical support for the Adaptive Gain Theory of exploratory decision making [12]. This biological theory of exploration proposes that such exploration induced changes in decision noise are modulated by activity of the locus coeruleus. Neuroscientific studies have begun to examine these mechanisms [12], but these remain to be fully characterized. Indeed, this area remains one in need of further adaptive exploration.

References

[1] L.P. Kaelbling, M.L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 1996.

[2] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

[3] J.C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gans, editor, *Progress in statistics*, pages 241–266. Amsterdam, The Netherlands: North-Holland, 1974.

[4] D. Luce. *Individual Choice Behavior.* Wiley, NY., 1959.

[5] R.J. Meyer and Y. Shi. Choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, 41(5):817–834, 1995.

[6] J. Banks, M. Olson, and D. Porter. An experimental analysis of the bandit problem. *Economic Theory*, 10(1):55–77, 1997.

[7] Michael J Frank, Bradley B Doll, Jen Oas-Terpstra, and Francisco Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, 12(8):1062–1068, 2009.

[8] Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–9, Jun 2006.

[9] M. Steyvers, M.D. Lee, and E.J. Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53:168–179, 2009.

[10] Elise Payzan-LeNestour and Peter Bossaerts. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, 7(1):e1001048, 2011.

[11] Daniel Ellsberg. Risk, ambiguity and the savage axioms. *The Quaterly Journal of Economics*, 75(4):643–669, 1961.

[12] Gary Aston-Jones and Jonathan D Cohen. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci*, 28:403–50, 2005.