Journal of Experimental Psychology: General
2014, Vol. 143, No. 6, 2074–2081

0096-3445/14/$12.00 http://dx.doi.org/10.1037/a0038199

BRIEF REPORT

Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma

Robert C. Wilson, Andra Geana, and John M. White
Princeton University

Elliot A. Ludvig
Princeton University and University of Warwick

Jonathan D. Cohen
Princeton University

All adaptive organisms face the fundamental tradeoff between pursuing a known reward (exploitation) and sampling lesser-known options in search of something better (exploration). Theory suggests at least two strategies for solving this dilemma: a directed strategy in which choices are explicitly biased toward information seeking, and a random strategy in which decision noise leads to exploration by chance. In this work we investigated the extent to which humans use these two strategies. In our "Horizon task," participants made explore–exploit decisions in two contexts that differed in the number of choices that they would make in the future (the time horizon). Participants were allowed to make either a single choice in each game (horizon 1), or 6 sequential choices (horizon 6), giving them more opportunity to explore. By modeling the behavior in these two conditions, we were able to measure exploration-related changes in decision making and quantify the contributions of the two strategies to behavior. We found that participants were more information seeking and had higher decision noise with the longer horizon, suggesting that humans use both strategies to solve the exploration–exploitation dilemma. We thus conclude that both information seeking and choice variability can be controlled and put to use in the service of exploration.

Keywords: explore-exploit, decision making, information bonus, decision noise, reinforcement learning

Supplemental materials: http://dx.doi.org/10.1037/a0038199.supp

When you go to your favorite restaurant, do you always order the same thing, or do you 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 and something you do not, is commonly referred to as the exploration–exploitation dilemma (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). Whether deciding on a meal, a vacation destination, or a life partner, this problem is an important one to solve.

Theoretical accounts suggest two distinct strategies for resolving this dilemma. One is directed exploration (Gittins & Jones, 1974; Gittins, 1979; Auer, Cesa-Bianchi, & Fischer, 2002), in which the sampling of informative options is encouraged by an "information bonus." The other strategy is random exploration (Thompson, 1933; Watkins, 1989; Bridle, 1990), in which "decision noise" encourages exploration by chance.

Directed strategies derive from theories of optimal decision making that ensure the greatest amount of reward in the long run. Random strategies on the other hand, reflect simpler heuristics that are not generally optimal, but can perform "well enough" at a fraction of the computational cost. Thus in our hypothetical restaurant scenario, a directed explorer would weigh the benefits of a known meal against the information available for sampling something new, whereas a purely random explorer would simply toss a coin to decide.

Despite these differences in implementation, both strategies are driven by the same goal: increasing reward in the long run. Both

This article was published Online First October 27, 2014.

Robert C. Wilson, Princeton Neuroscience Institute, Princeton University; Andra Geana, Princeton Neuroscience Institute and Department of Psychology, Princeton University; and John M. White, Department of Psychology, Princeton University; Elliot A. Ludvig, Department of Psychology, Princeton University and Department of Psychology, University of Warwick; Jonathan D. Cohen, Princeton Neuroscience Institute and Department of Psychology, Princeton University.

The data and analysis scripts used in this study are available upon request and will be made freely available online after publication. We thank Peter Foster, Laura DeSouza, and Abigail Novick for help running participants and Paul Krueger, Michael Todd, David Barrack, and Matthew Nassar for helpful discussions. This project/publication 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. Robert C. Wilson was funded by a J. Insley Blair Pyne Fund Award to Jonathan D. Cohen.

Correspondence concerning this article should be addressed to Robert C. Wilson, Princeton Neuroscience Institute, Princeton University, Washington Road, Princeton, NJ 08544. E-mail: rcw2@princeton.edu

strategies try to achieve this goal by encouraging the exploration of informative options. Thus a key question in understanding how humans solve the explore–exploit dilemma is to understand whether and how they use these two strategies.

Previous work looking for directed exploration in humans has led to mixed results. Some authors have found evidence for this strategy (Meyer & Shi, 1995; Banks, Olson, & Porter, 1997; Frank, Doll, Oas-Terpstra, & Moreno 2009; Steyvers, Lee, & Wagenmakers, 2009; Lee, Zhang, Munro, & Steyvers, 2011; Payzan-LeNestour & Bossaerts, 2012; Zhang & Yu, 2013), and others have not (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Payzan-LeNestour & Bossaerts, 2011). We believe that this discrepancy occurs for two reasons: The first is a subtle confound between reward and information that makes directed exploration difficult to identify, and the second is an interaction with ambiguity preference.

The reward–information confound arises because most exploration tasks involve sequential decisions with partial feedback (Hertwig & Erev, 2009). This means that participants only receive information about the options they select, and thus the information available for a current decision depends crucially on choices made in the past. Because participants make choices to maximize their rewards, these past choices tend to favor options that have already proven to be rewarding, which is not surprising. As a result of this sampling bias, the more immediately rewarding options become better known to the participants. Thus the information available for playing a particular option becomes confounded with its reward, making directed exploration (favoring information over immediate reward) difficult to detect.

This informativeness–ambiguity interaction—the fact that more informative options are also more ambiguous—means that directed exploration may be masked in participants who are ambiguity-averse (Ellsberg, 1961; Camerer & Weber, 1992), or incorrectly identified when participants are ambiguity-seeking (Kahn & Sarin, 1988; Bier & Connell, 1994). Thus, to measure directed exploration, it is important to take account of baseline attitudes to ambiguity.

Random exploration in human behavior has been studied in less detail than directed exploration. Although a certain amount of choice variability is often assumed in many cognitive models (Luce, 1959), including many models of exploratory behavior (Aston-Jones & Cohen, 2005; Daw et al., 2006; Frank et al., 2009; Steyvers et al., 2009; Payzan-LeNestour & Bossaerts, 2011; Lee et al., 2011; Payzan-LeNestour & Bossaerts, 2012; Zhang & Yu, 2013), relatively little experimental work has explicitly tied this variability to exploration. One notable exception is (Lee et al., 2011), who used a model in which random exploration decreased over time, although this model did not provide the best fit to their data. Outside of the field of decision making, Wu et al. (2014) recently showed that humans use and adapt motor noise to aid motor learning. Likewise, there is a growing body of evidence that motor noise is used by song birds as a means of exploration in song learning and that this noise is generated by specific neural structures (Doya & Sejnowski, 1995; Ölveczky, Andalman, & Fee, 2005; Kao, Doupe, & Brainard, 2005; Tumer & Brainard, 2007).

In this study we designed a task to remove the problems associated with measuring directed exploration and to explicitly evaluate the contribution of choice variability to random exploration. In this *Horizon task*, participants played a series of games in which they made choices between two options that paid out probabilistic rewards. To remove the reward–information confound, we carefully controlled the amount of information subjects had about each option *before* they made their choices, thus "decorrelating" information and reward. To account for ambiguity preference and to assess the role of choice variability in exploration, we varied the number of trials in each game, the horizon. This horizon manipulation took advantage of the fact that, whereas ambiguity preference and intrinsic sources of choice variability should be fixed, endogenous factors, the amount of exploration should increase when subjects have more trials in which to explore. Thus, we were able to identify directed and random exploration as changes in information seeking and decision noise with horizon. Using this approach we found that humans do indeed use both strategies of exploration, and that they adapt these strategies based on the opportunity to explore.

Method

Participants

Thirty-one participants (20 women; mean age 19.7 years; range 18–24 years) were recruited from the Princeton student population. Participants were not paid for their performance or time, but received course credit for taking part in the experiment (a follow up experiment in which participants were paid based on their performance is reported in the Supplementary Material). All participants gave informed consent, and the study was approved by the Princeton University Institutional Review Board.

Horizon Task

Participants played 320 games (in four blocks of 80 games) of our Horizon task (see Figure 1A). Each game lasted either five or 10 trials and the two game lengths were interleaved and counterbalanced such that there were 160 games of each length.

In each game, participants made repeated decisions between two options. Each option paid out between 1 and 100 points that was sampled (rounded to the nearest integer) from a Gaussian distribution with a fixed standard deviation of 8 points. The generative means of the underlying Gaussians were different for the two options and remained stable within a game. In each game, the mean of one option was set to either 40 or 60 points and the mean of the other was set relative to the mean of the first, such that the difference between the two was sampled from 4, 8, 12, 20, and 30. Both the identity and the difference in means were counterbalanced over the entire experiment.

Participants were instructed in the task with the use of a set of illustrated onscreen instructions. These explicitly conveyed that the means of the two options were constant over a game and that the variability in the options was constant over the entire experiment. Participants were told to maximize the points they earned and that one option was always better on average. The full text of the instructions is provided in the Supplementary Material.

Choice and outcome history in each game remained onscreen inside each of the slot machines (Figure 1A). After a particular option was played, the reward on that trial was added to the slot machine, whereas the corresponding space for the unplayed option was filled with an "XX."

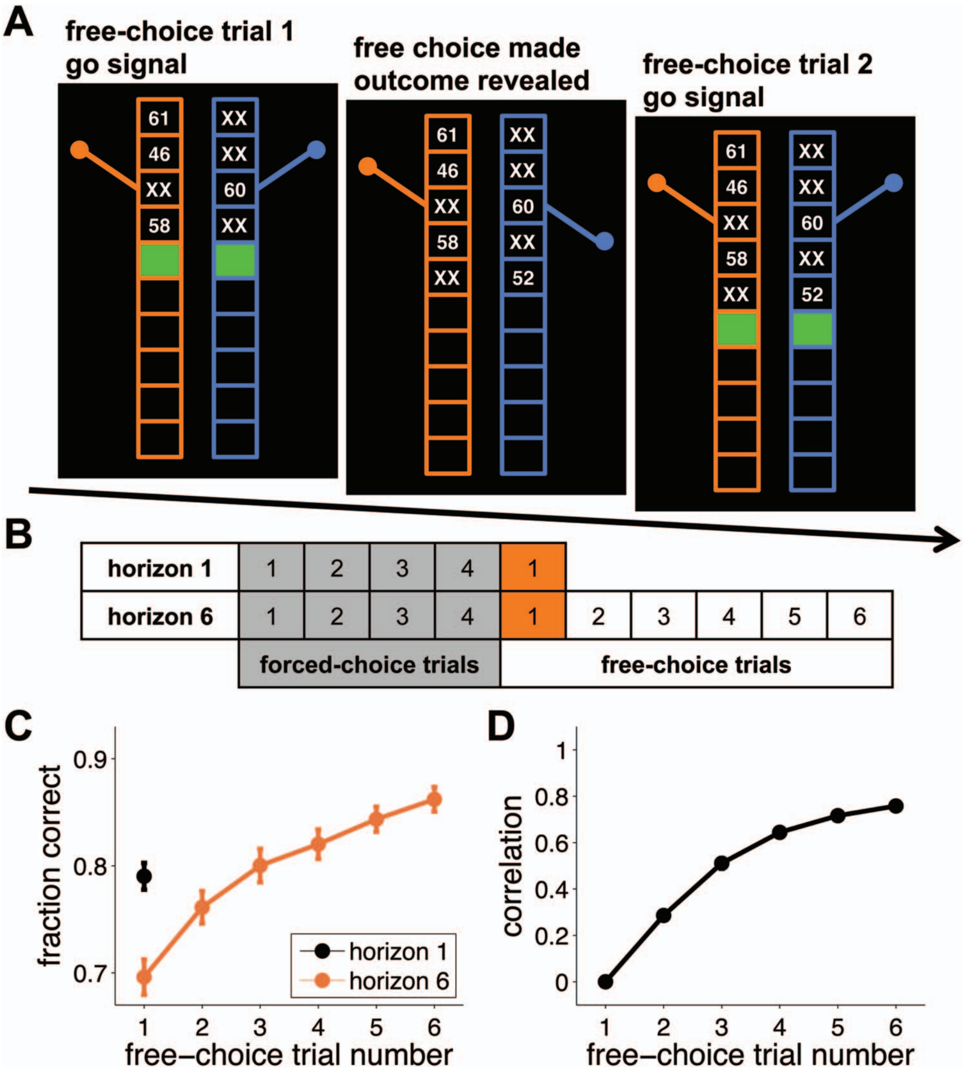

Figure 1. Task design (A). Example screen shots from a horizon 6 game showing the first free-choice trial, the result of the choice (after it is made), and the start of the second free-choice trial (B). Schematic showing the different trial types in the two horizon conditions. Each game began with four forced-choice trials before one or a sequence of six free-choice trials. In all conditions, the first free-choice trial (orange) was the main focus of subsequent analyses (C). Learning curves showing the fraction of times the correct option (i.e., the option with the higher generative mean) was chosen as a function of free-choice trial number for the different horizon conditions. This demonstrates that participants performed at above-chance levels and improved as the game progressed (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. Only the very first trial showed an absence of correlation; thereafter, there was a strong correlation as participants received more information about the more rewarding options that they selected. See the online article for the color version of this figure.

The first four trials of each game were forced-choice trials, in which only one of the options (cued by a green square inside the next available space) was available for participants to choose. We used these forced-choice trials to manipulate the information participants had about the two options from experience (Hertwig, Barron, Weber, & Erev, 2004) before their first free choice, while maintaining their active engagement in the task. The four forced-choice trials set up two information conditions: “unequal information” (or [1 3]), in which one option was forced to be played once and the other three times, and “equal information” (or [2 2]), in which each option was forced to be played twice.

Crucially, this manipulation 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 independent 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 the difference in available information) had no effect on the difference in mean payout of that option (repeated-measures ANOVA, $F(2, 89) = 0.09$, $p = .91$) thus removing the reward–information confound on this trial.

After the forced-choice trials, participants made either one or six free choices (Figure 1B). At the beginning of each game, the number of upcoming free-choice trials (i.e., the horizon) was indicated by the length of the slot machines (Figure 1A), which contained an empty space awaiting the outcome from each of the subsequent trials.

Finally, to test whether participants were performing the task, we assessed their performance over the entire experiment. Any participant whose performance was not significantly different from chance at a threshold of $p = .01$ was deemed not to have performed the task and was excluded from the analysis. One participant was excluded in this way, leaving 30 participants for the main analysis.

Model Fitting

We fit the behavior on the first free-choice trial using a simple logistic model. This model computes a value, Q_a, for each option, a, and makes probabilistic choices based on these values. In particular, Q_a is the weighted sum of the expected reward R_a, information I_a, and spatial location s_a,

$$Q_a = R_a + \alpha I_a + B s_a,$$

where α denotes the information bonus and B the spatial bias.

If we assume the values for each option are perturbed by logistic noise with variance σ_d and the model chooses the option with highest perturbed value, then the probability of choosing option a over option b is

$$p_a = \frac{1}{1 + \exp\left(\dfrac{R_b - R_a + \alpha(I_b - I_a) + B(s_b - s_a)}{\sigma_d}\right)}$$

The expected rewards R_a and R_b were set as the observed mean of the outcomes of the example plays for options a and b respectively, which assumes that participants have a linear utility function and weigh each outcome equally in the decision. Relaxing these assumptions did not change the main results (see Supplementary Material).

The information I_a was defined such that when option b was more informative than a in the unequal condition, $I_b - I_a = +1$, and $I_b - I_a = -1$, when b is less informative than a. In the equal condition, $I_b - I_a = 0$. This choice of I_a allows us to interpret the information bonus as the indifference point of the choice curves in Figure 2A. Similarly, the location variable s_a was defined such that $s_b - s_a = +1$ when b is on the right and a is on the left $s_b - s_a = -1$ when b is on the left, and a is on the right.

By fitting this choice function to participants' data we were able to estimate the information bonus α the bias B, and the magnitude of decision noise σ_d, separately for each participant in each information and horizon condition.

Results

Performance was above chance (50%) in all horizon conditions (Figure 1C) and improved throughout each game for the longer horizon condition, indicating that participants understood and were engaged in 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 information (Figure 1D). This result demonstrates how rapidly the confound between reward and information arises in a free-choice setting. Because of these strong correlations on later trials, our initial analysis focuses solely on the first free-choice trial when reward and information are independent.

Directed and Random Exploration on the First Free Choice

We computed the probability of choosing one of the options, Option A, on the first free-choice trial as a function of the difference in means of the samples observed on the forced plays. By convention, we defined Option A differently in the two information conditions. In the unequal condition, it was the more informative option (i.e., the option played only once in the [1 3] forced-choice trials). In the equal condition, because both options were equally informative, Option A was the option on the right hand side of the screen. The resulting empirical choice curves along with the average fits from the model 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 the two options. Furthermore, for that choice, increasing the horizon from 1 to 6 increased the probability of choosing the more informative option in the unequal condition. For example, in the Horizon 6 condition, even when the mean of the more informative option was 8 points lower than the alternative (−8 on the *x* axis), 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 indicative of directed exploration driven by an information bonus. That is, on the first free-choice trial in the Horizon 6 condition, participants behaved as though the more informative option had greater value.

In addition to the shift in the indifference point, there was also a change in the slope of the choice curves with the horizon (see Figure 2A and B). Curves in Horizon 1 were steeper than those in Horizon 6 for both information conditions. This change in slope is consistent with random exploration induced by an increase in decision noise. That is, participants' choices on the first free-choice trial became more random, and hence less correlated with the difference in means as the horizon increased.

The model fit confirmed these informal observations, as shown in Figure 2C–E. Consistent with the choice curves, there was a highly significant increase in information bonus between horizons 1 and 6, $t(29) = 5.05$, $p < 10^{-4}$. Likewise a repeated measures ANOVA found a significant increase in decision noise with horizon, $F(1, 119) = 65.97$, $p < 10^{-8}$ in addition to a small main effect of information condition, $F(1, 119) = 5.06$, $p < .05$ but no interaction between the horizon and information conditions, $F(1, 119) = 0$, $p = .96$. Furthermore, the effect of the horizon held for almost all subjects (see Figure 2F–H), with 25 out of 30 showing an increase in information bonus in the long-horizon condition, and 28 showing a similar increase in decision noise (observed in both of the information conditions).

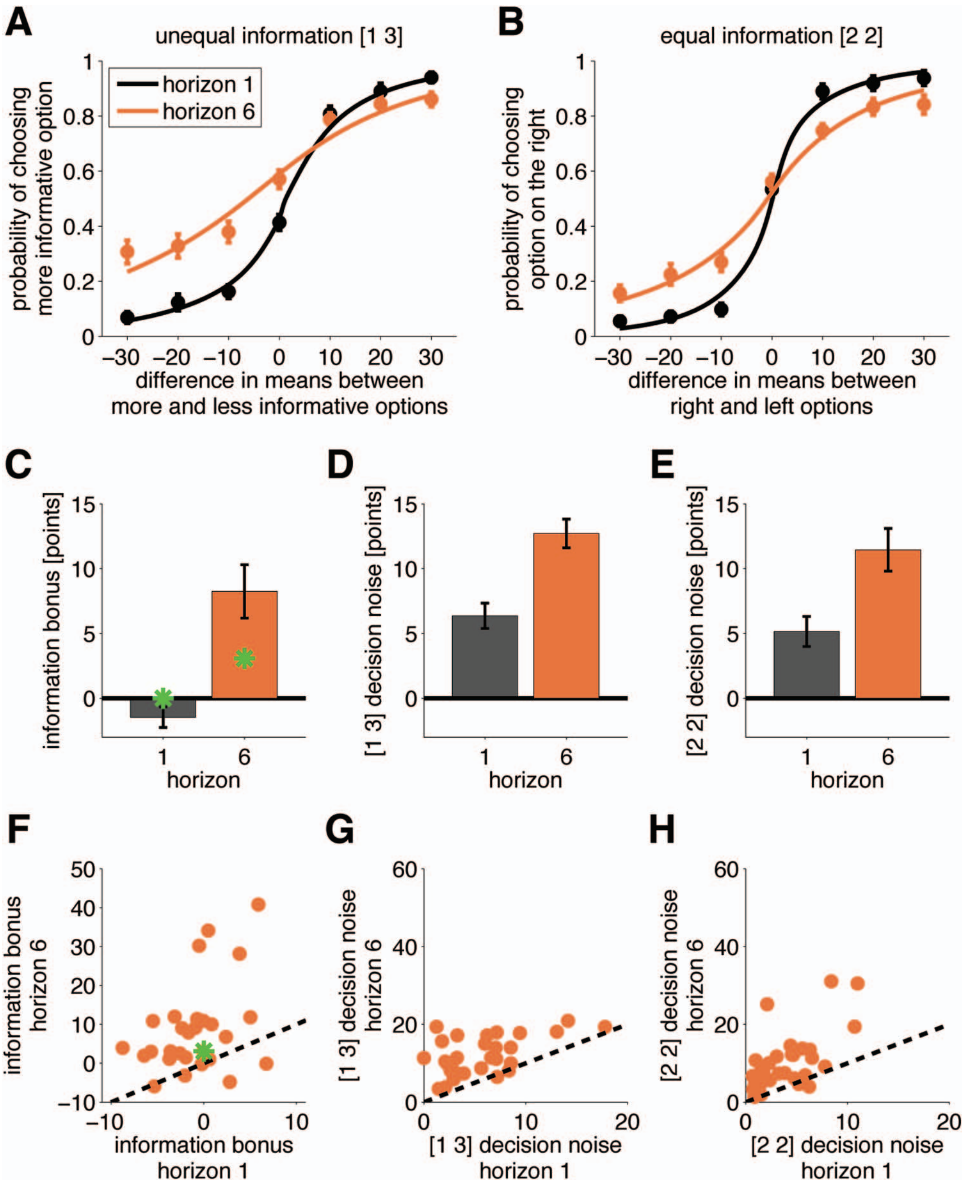

Figure 2. Behavior on the first free-choice trial. Choice curves for the unequal (A) and equal (B) information conditions. Filled circles = experimental data; solid lines = average over participants of model-derived choice curves (A). These curves show the fraction of times the more informative option was chosen on the first free-choice trial as a function of the difference in mean between the more informative option and the less informative option. As the horizon increased, the more informative option was chosen more often, indicating an information bonus. In addition, the slope of the curves decreased, indicating a change in decision noise (B). In the equal condition, as the horizon increased there was no change in indifference point because both options were equally informative. The slope of the curves, however, decreased again, consistent with an increase in decision noise with horizon. Mean parameter fits for the information bonus (C) and decision noise in the [1 3] condition (D) and decision noise in the [2 2] condition (E) showing an increase in information bonus and decision noise between horizons 1 and 6. Error bars are s.e.m. across participants. Scatter plots comparing parameter fits for individual subjects in horizon 1 and horizon 6. The dashed lines denote equality. This shows that the increase in information bonus (F) and decision noise (G and H) with horizon holds for almost all of the subjects. See the online article for the color version of this figure.

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 by running the same fitting procedure on simulated-choice data from the optimal model (see Supplementary Material). The optimal information bonus, plotted as the green stars in Figures 2C and 2F, bears a qualitative resemblance to the estimates of the information bonus for the human participants, although quantitatively, participants appear to exhibit a greater information bonus than the optimal value. It is worth noting that, in this task, optimal performance is associated with zero decision noise, which is clearly quite different from the behavior shown by humans (Figure 2D, E).

Random Exploration Decreases Over the Course of Horizon 6 Games

The change in decision noise between horizons 1 and 6 suggests that it reflects random exploration, rather than intrinsic variability in performance, because there should be less exploration with fewer opportunities to learn. Accordingly, this adaptive process should also occur over the course of the horizon 6 games: On early trials, decision noise should be high, and on later trials, decision noise should be low.

To test this hypothesis, we looked at choice behavior on the first, third and fifth free-choice trials, when both options had been chosen an equal number of times, that is, in the [2 2], [3 3], and [4 4] information conditions. By focusing on these equal information conditions, we removed any complications due to changes in directed exploration, as well as the deleterious effects of the reward–information confound (because there was no information difference to be confounded with reward).

The results in Figure 3 show a clear change in decision noise over the course of the game, with the slopes of the choice curves increasing (Figure 3A) and the fit decision-noise parameter decreasing (Figure 3B). A repeated-measures ANOVA on decision noise corroborated this observation, with a strong main effect of trial number, $F(1, 119) = 35.38$, $p < 10^{-10}$. The decrease in noise between the [2 2] and other conditions was large, [2 2] vs. [3 3], $t(29) = 5.73$, $p < 10^{-5}$; [2 2] vs. [4 4], $t(29) = 6.67$, $p < 10^{-6}$ and held for almost all subjects (see Figure 3C and D), whereas the decrease in noise between the [3 3] and [4 4] conditions was smaller, though still statistically significant, $t(29) = 2.22$, $p < .05$.

Discussion

In this study, we investigated the extent to which humans use directed and random exploration to solve the exploration–exploitation dilemma. The results indicate that humans use both strategies, with both information bonus and decision noise increasing between horizons 1 and 6.

For directed exploration, we showed that, when the horizon is longer, humans exhibit an information bonus that effectively increases the value of the more informative option, making sampling that option more likely. From a theoretical perspective this result

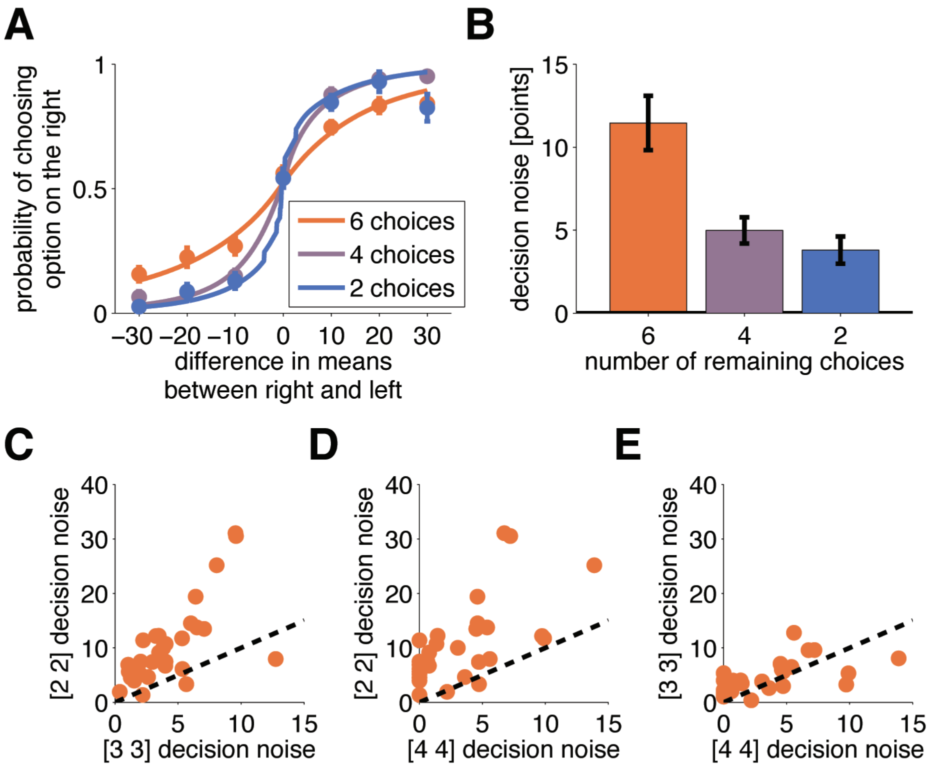

Figure 3. Change in decision noise across the horizon 6 game (A). Choice curves for the equal information conditions on the first [2 2], third [3 3] and fifth [4 4] free-choice trials in horizon 6 games. This shows an increase in slope on the later trials relative to the first trial consistent with a decrease in random exploration (B). Participants' mean decision noise extracted from the model fits showing a decrease in decision noise in the equal conditions over the horizon 6 games (C, D, E). Comparison between the individual participants' decision noise in the three information conditions shows that the decrease in decision noise between the first and later trials holds for most subjects (C) and (D), whereas there is no difference between the decision noise in the [3 3] and [4 4] conditions (E). See the online article for the color version of this figure.

is not surprising. It is well-known that information has real value for long horizons (Gittins & Jones, 1974; Gittins, 1979), and a carefully calibrated information bonus insures optimal or near-optimal exploration in many settings (Bubeck & Cesa-Bianchi, 2012).

Experimentally, however, previous results on directed exploration have been mixed, with some studies finding evidence for this strategy (Meyer & Shi, 1995; Banks et al.,1997; Frank et al., 2009; Steyvers et al., 2009; Lee et al., 2011; Payzan-LeNestour & Bossaerts, 2012; Zhang & Yu, 2013) and others failing to do so (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011). We believe that one reason for these mixed results is the subtle confound between reward and information that arises in sequential choice tasks and makes directed exploration both hard to observe and difficult to confirm. In the horizon task, we removed this confound on the first free-choice trial by manipulating reward and information *before* subjects made a free choice. This allowed us to unambiguously identify directed exploration on that trial. On later trials, however, the correlation between information and reward grew quickly making the behavior on later trials more difficult to interpret.

Although our results demonstrate the qualitative existence of directed exploration, they are limited in what they can tell us about its quantitative properties. In particular, because we had only two horizon and information conditions, it is impossible to measure how directed exploration changes parametrically with horizon and information. In a parallel experiment (Figure S6) with three horizon conditions (1, 6, and 11), we found no differences between horizon 6 and 11, consistent with a categorical effect of horizon (i.e., no information bonus for horizon 1, fixed bonus for horizon > 1). In contrast, however, analysis of later trials in the horizon 6 games (Figure S3) suggests a parametric decrease in information bonus over the game, although interpretation of this result is complicated by the reward–information confound discussed above, and because horizon decreases as information increases on these later trials.

For random exploration, our findings showed both an increase in decision noise between horizons 1 and 6 and a decrease in decision noise over the course of the horizon 6 games. Taken together, these findings provide compelling evidence that choice variability is adaptively modulated in the service of exploration.

Such noise-driven exploration has a long history in statistics (Thompson, 1933) and machine learning (Watkins, 1989; Bridle, 1990; Sutton & Barto, 1998) in terms of its simplicity, which allows the strategy to be applied to situations in which the optimal information bonus is hard to compute. 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 and may, furthermore, even meliorate any losses when the information bonus is wrong. In this light, the use of random exploration by our participants may reflect an effort (either conscious or unconscious) to compensate for their incorrect setting of the information bonus. More generally, investigating this potential *interaction* between the two exploration strategies is an important direction for future research.

An important question concerns the nature of the decision noise: Is it truly random noise, perhaps driven by internal neural variability (Faisal, Selen & Wolpert, 2008), or is it a deterministic process driven by external factors (Osborne, Lisberger & Bialek, 2005), or suboptimal inference by the brain (Beck et al., 2012)? All of these factors would be interpreted as noise by our model but their sources are very different. Because arbitrary deterministic strategies—such as going left on every fifth trial—are impossible to exclude, we used model fitting to rule out some of the better known cognitive factors (such as order effects, choice kernel, variance, trend bias, peak bias, and a nonlinear utility) as alternate explanations for the noise (see Supplementary Material). Future experiments with more carefully controlled stimuli—as has been used to great effect to measure the components of perceptual decision noise (e.g., Osborne et al., 2005)—will be needed to address this question more fully.

A related question asks, "How is the decision noise adapted?" Are existing sources of neural variability amplified, is noise explicitly injected into the system, or are the changes due to the adaptation of a deterministic process?" One possibility has been suggested by the adaptive gain theory of exploratory decision making (Aston-Jones & Cohen, 2005; McClure, Gilzenrat & Cohen, 2006), which proposes that decision noise is controlled by tonic levels of norepinephrine.

More generally, the question of the psychology and neuroscience of explore–exploit decisions is one that remains in need of further adaptive exploration.

References

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403–450. http://dx.doi.org/10.1146/annurev.neuro.28.061604.135709

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47,* 235–256. http://dx.doi.org/10.1023/A:1013689704352

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory, 10,* 55–77. http://dx.doi.org/10.1007/s001990050146

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron, 74,* 30–39. http://dx.doi.org/10.1016/j.neuron.2012.03.016

Bier, V. M., & Connell, B. L. (1994). Ambiguity seeking in multi-attribute decisions: Effects of optimism and message framing. *Journal of Behavioral Decision Making, 7,* 169–182. http://dx.doi.org/10.1002/bdm.3960070303

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2, pp. 211–217). Cambridge, MA: MIT Press.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning, 5,* 1–130.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty, 5,* 325–370. http://dx.doi.org/10.1007/BF00122575

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. http://dx.doi.org/10.1038/nature04766

Doya, K., & Sejnowski, T. J. (1995). A novel reinforcement model of birdsong vocalization learning. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 101–108). Cambridge, MA: MIT Press.

Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *The Quarterly Journal of Economics, 75,* 643. http://dx.doi.org/10.2307/1884324

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience, 9,* 292–303. http://dx.doi.org/10.1038/nrn2258

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12,* 1062–1068. http://dx.doi.org/10.1038/nn.2342

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

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gans (Ed.), *Progress in statistics* (pp. 241–266). Amsterdam, the Netherlands: North–Holland.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004

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

Kahn, B. E., & Sarin, R. K. (1988). Modeling ambiguity in decisions under uncertainty. *The Journal of Consumer Research, 15,* 265–272. http://dx.doi.org/10.1086/209163

Kao, M. H., Doupe, A. J., & Brainard, M. S. (2005). Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature, 433,* 638–643. http://dx.doi.org/10.1038/nature03127

Lee, M. D., Zhang, S., Munro, M. N., & Steyvers, M. (2011). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research, 12,* 164–174. http://dx.doi.org/10.1016/j.cogsys.2010.07.007

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

McClure, S. M., Gilzenrat, M. S., & Cohen, J. D. (2006). An exploration–exploitation model based on norepinepherine and dopamine activity. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 867–874). Cambridge, MA: MIT Press.

Meyer, R., & Shi, Y. (1995). Choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science, 41,* 817–834. http://dx.doi.org/10.1287/mnsc.41.5.817

Ölveczky, B. P., Andalman, A. S., & Fee, M. S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biology, 3,* e153. http://dx.doi.org/10.1371/journal.pbio.0030153

Osborne, L. C., Lisberger, S. G., & Bialek, W. (2005). A sensory source for motor variation. *Nature, 437,* 412–416. http://dx.doi.org/10.1038/nature03961

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7,* e1001048. http://dx.doi.org/10.1371/journal.pcbi.1001048

Payzan-LeNestour, E., & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: Estimation uncertainty and "unexpected uncertainty" both modulate exploration. *Frontiers in Neuroscience, 6,* 150. http://dx.doi.org/10.3389/fnins.2012.00150

Steyvers, M., Lee, M., & Wagenmakers, E. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53,* 168–179. http://dx.doi.org/10.1016/j.jmp.2008.11.002

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

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25,* 285–294. http://dx.doi.org/10.1093/biomet/25.3-4.285

Tumer, E. C., & Brainard, M. S. (2007). Performance variability enables adaptive plasticity of 'crystallized' adult birdsong. *Nature, 450,* 1240–1244. http://dx.doi.org/10.1038/nature06390

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University, Cambridge, England.

Wu, H. G., Miyamoto, Y. R., Gonzalez Castro, L. N., Ölveczky, B. P., & Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nature Neuroscience, 17,* 312–321. http://dx.doi.org/10.1038/nn.3616

Zhang, S., & Yu, A. J. (2013). Forgetful Bayes and myopic planning: Human learning and decision making in a bandit setting. *Advances in Neural Information Processing Systems, 26,* 2607–2615.

Received November 15, 2013
Revision received August 12, 2014
Accepted August 18, 2014 ■