The nature of decision noise in random exploration

Siyu Wang[1] and Robert C. Wilson[1,2]

[1]Department of Psychology, University of Arizona, Tucson AZ USA

[2]Cognitive Science Program, University of Arizona, Tucson AZ USA

August 5, 2020

Abstract

Human decision making is inherently variable. While this variability is often seen as a sign of suboptimality behavior, recent work suggests that variability can actually be adaptive. An example arises when we must choose between exploring unknown options or exploiting options we know well. A little randomness in these 'explore-exploit' decisions is remarkably effective as it encourages us to explore options we might otherwise ignore. Recent work suggests that people may actually use such 'random exploration' in practice, increasing their behavioral variability when it is more valuable to explore. A key question is whether the variability in random exploration is actually random. That is, is random exploration driven by stochastic processes in the brain or by some unobserved deterministic process that we have failed to account for when measuring behavioral variability? By designing an explore-exploit task in which, unbeknownst to them, participants are presented with the exact same choice twice, we provide a partial answer to this question. In particular, we find evidence that at least part (IT SOUNDS HARD TO ME TO ARGUE THAT THIS NUMBER IS NOT TASK SPECIFIC. - BOB SAYS: CAN WE PUT A NUMBER ON THIS E.G. 25%?) of the variability in 'random' exploration can be accounted for by deterministic processing of the stimulus. This still leaves open the possibility that much of random exploration is truly 'random,' but narrows the window of opportunity for stochastic processes to explain this behavior.

Introduction

Imagine trying to decide where to go to dinner on a date, you can go to your favorite restaurant, the one you both really enjoy and always go to, or you can try a new restaurant that you know nothing about. Such decisions, in which we must choose between a well-known 'exploit' option and a lesser known 'explore' option, are known as explore-exploit decisions. From a theoretical perspective, making optimal explore-exploit choices, i.e. choices that maximize long-term reward, is computationally intractable in most cases (Basu et al., 2018, Gittins and Jones, 1974). In part because of this difficulty, there is considerable interest in how humans and animals solve the explore-exploit dilemma in practice (Auer et al., 2002, Banks et al., 1997, Bridle, 1990, Daw et al., 2006, Frank et al., 2009, Gittins, 1979, Krebs et al., 1978, Lee et al., 2011, Meyer and Shi, 1995, Payzan-LeNestour and Bossaerts, 2011, Payzan-Lenestour and Bossaerts, 2012, Steyvers et al., 2009, Thompson, 1933, Watkins, 1989, Wilson et al., 2014, Zhang and Yu, 2013).

One particularly effective strategy for solving the explore-exploit dilemma is choice randomization (Bridle, 1990, Thompson, 1933, Watkins, 1989). In this strategy, the decision process between exploration and exploitation is corrupted by 'decision noise,' meaning that high value 'exploit' options are not always chosen and exploratory choices are sometimes made by chance. In theory, such 'random exploration,' is surprisingly effective and, if implemented correctly, can come close to optimal performance (Agrawal and Goyal, 2011, Bridle, 1990, Chapelle and Li, 2011, Thompson, 1933).

It has recently been shown that humans appear to use random exploration and can increase such decision noise when it is more beneficial to explore (Gershman, 2018, Wilson et al., 2014). In one of these tasks, known as the Horizon Task, the key manipulation is the horizon condition, i.e. the number of decisions remaining for the participant to make. Increasing the horizon makes exploration more valuable as there is more time to use the information gained by exploration to maximize future rewards. For example, if you are leaving town tomorrow (short horizon), you will probably exploit the restaurant you know and love, but if you are in town for a while (long horizon), you would be more likely to explore the new restaurant. Using such a horizon manipulation we found that people's behavior is more variable in long horizons than short horizons, suggesting that they use adaptive decision noise to solve the explore-exploit dilemma (Wilson et al., 2014).

One limitation of this previous work, is that it is difficult to tell whether what is measured as decision noise is truly random. That is, whether behavioral variability is due to intrinsic stochastic processes in the brain or whether it is due to deterministic processes that we simply have not observed. For example, in the restaurant example, my usual preference for one restaurant or another may be overruled if I see an ex romantic partner going into one of them. Avoiding an ex is a deterministic process, but if we fail to take ex's presence into account as scientists modeling the decision, then over a series of such decisions where the ex is present or not, we would mistakenly attribute the ensuing 'variability' in choice to randomness.

In this paper, we investigate the extent to which the apparent randomness in random exploration can be explained by such deterministic processing of the stimulus. In particular, we modify the Horizon Task to have people face the exact same explore-exploit choice twice. If the decision is a purely deterministic function of the stimulus, then people's choices should be identical both times. That is, their choices should be consistent, since the stimulus is the same both times. Conversely, the more their decision is driven by other processes, including both stochastic and unobserved deterministic processes, the less consistent their behavior should be. By analyzing behavior on this task in both a model-free and model-based manner, we show that at least some of the 'randomness' in random exploration must come from deterministic processing of the stimulus. This does not prove that random exploration is entirely deterministic or stochastic, but provides a lower bound on how much deterministic processes contribute to random exploration. This in turn sheds light on how random exploration may be implemented by amplifying the effects of task-irrelevant stimuli on choice.

Methods

Participants

80 participants (ages 18-25, 37 male, 43 female) from the University of Arizona undergraduate subject pool participated in the experiment. 14 were excluded on the basis of performance, using the same exclusion criterion as in (Wilson et al., 2014). In this exclusion criteria, we measured the accuracy of each participant's choices by calculating the percentage of times that a participant chose the bandit with the higher underlying mean payouts in the last choice of a long horizon game, intuitively people should figure out which bandit has a higher mean payout by the last trial and should have an accuracy measure significantly above 50%, specifically, we computed the likelihood that the measured accuracy can be achieved driven by making a completely random choice between the two options and excluded participants with a likelihood smaller than 99.999%, in other words, participants who didn't show an accuracy significant above chance with $p > 0.001$ were excluded in the analysis. This left 66 for the main analysis. Note that

including the 14 badly performing subjects did not change the main results (Supplementary Figures 1 - 3)

Task

The task was a modified version of the Horizon Task (Wilson et al., 2014) (Figure 1). In this task, participants play a set of games in which they make choices between two slot machines (one-armed bandits) that pay out rewards from different Gaussian distributions. In each game they made multiple decisions between two options. Each option paid out a random reward between 1 and 100 points sampled from a Gaussian distribution. The means of the underlying Gaussians were different for the two bandit options, remained the same within a game, but changed with each new game. One of the bandits always had a higher mean than the other. Participants were instructed to maximize the points earned over the entire task. To maximize their rewards in each game, participants need to exploit the slot machine with the highest mean, but they cannot identify this best option without exploring both options first.

The number of games participants played depended on how well they performed, which acted as the primary incentive for performing the task. Thus, the better participants performed, the sooner they got to leave the experiment. On average, participants played 151.6 games (minimum = 90 games, maximum = 192 games) and the whole task lasted between 12.34 and 32.74 minutes (mean 23.31 minutes). Participants played an average of 67.31 repeated pairs of games (minimum = 25 repeated pairs, maximum = 100 repeated pairs).

As in the original paper (Wilson et al., 2014), the distributions of payoffs tied to bandits were independent between games and drawn from a Gaussian distribution with variable means and fixed standard deviation of 8 points. Differences between the mean payouts of the two slot machines were set to either 4, 8, 12 or 20. One of the means was always equal to either 40 or 60 and the second was set accordingly. Participants were informed that in every game one of the bandits always has a higher mean reward than the other. The order of games was randomized. Mean sizes and order of presentation were counterbalanced.

Each game consisted of 5 or 10 choices. Every game started with a fixation cross, then a bar of boxes appeared indicating the horizon for that game. For the first 4 trials - the instructed trials, we highlight the box on one of the bandits to instruct the participant to choose that option. On these trials, they have to press the corresponding key to reveal the outcome. From the fifth trial, boxes on both bandits will be highlighted and they are free to make their own decision. There was no time limit for decisions. During free choices participants could press either the left arrow key or right arrow key to indicate their choice

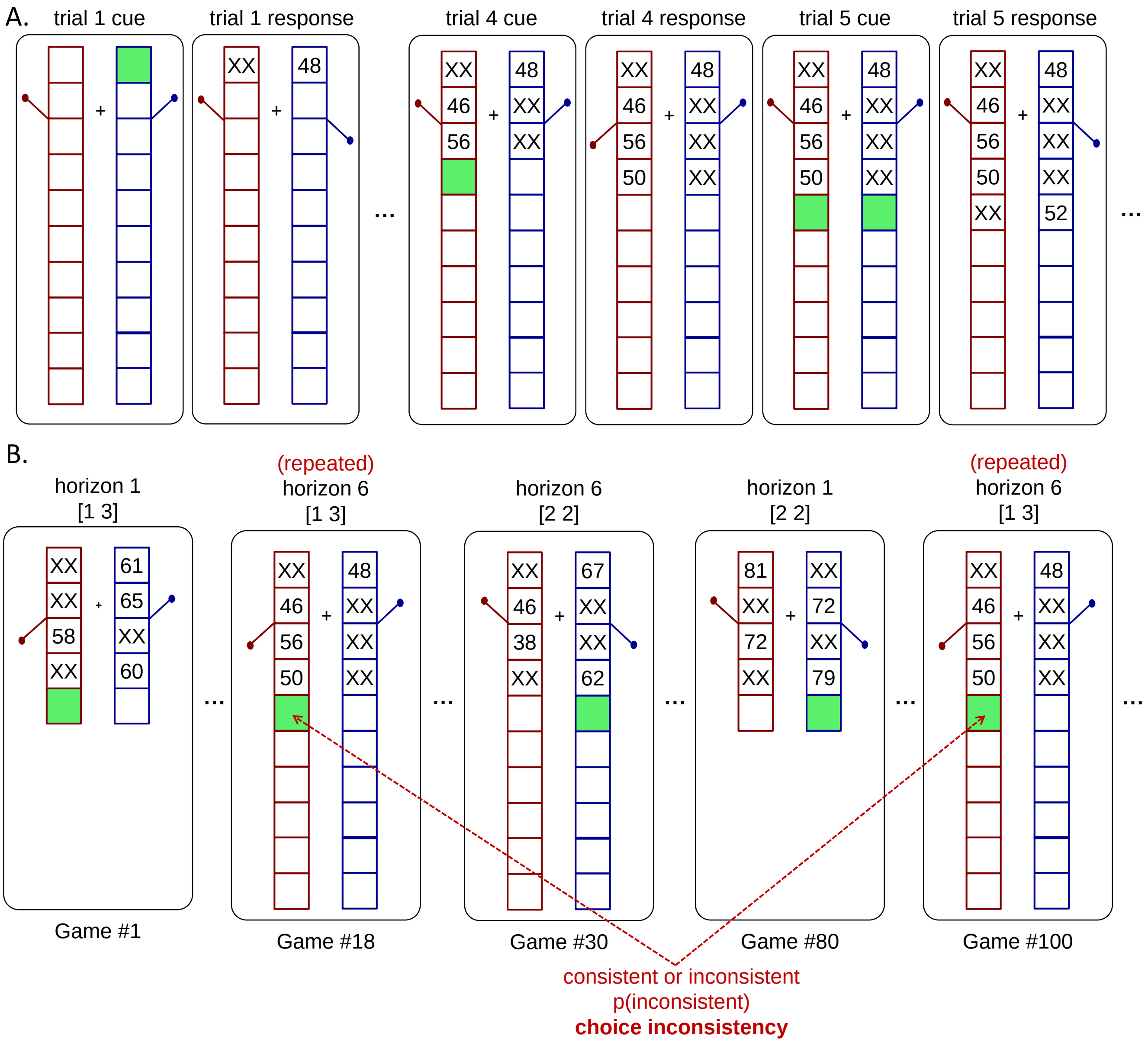


Figure 1: Schematic of the experiment. (A) Dynamics of an example horizon 6 game. Here the first four trials are forced trials in which participants are instructed which option to play. After the forced trials, participants are free to choose between the two options for the remainder of the game. (B) Different possible states of the game after the first free choice over the course of the experiment. Overall participants play 160 such games, with varying horizon (1 vs 6), uncertainty condition ([1 3] vs [2 2]) and observed rewards. In addition, all games are repeated (as Game 18 and 100 are here) such that participants will be faced with the exact same pattern of forced trials and exact same outcomes from those forced trials twice within each experiment. These repeated games allow us to compute the relative contribution of deterministic and random noise by analyzing the extent to which choices are *consistent* across the repeated games.

of left or right bandit. The score feedback was presented for 300ms. The task was programmed using Psychtoolbox in MATLAB (Brainard, 1997, Pelli, 1997).

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 bandit before their first free choice. The four forced-choice trials set up two uncertainty conditions: unequal uncertainty(or [1 3]) in which one option was forced to be played once and the other three times, and equal uncertainty(or [2 2]) in which each option was forced to be played twice. After the forced-choice trials, participants made either 1 or 6 free choices (two horizon conditions), Figure 1.

Model-based analysis

We modeled behavior on the first free choice of the Horizon Task using a version of the logistic choice model in (Wilson et al., 2014) that was modified to differentiate between components of the noise that are deterministically driven by the stimulus ('deterministic noise') and components of the noise that are not deterministically driven by the stimulus ('random noise'). Because the stimuli are identical in the repeated games, by definition, deterministic noise remains the same in repeated games, whereas random noise can change.

Hierarchical Bayesian Model

To model participants' choices on this first free-choice trial, we assume that they make decisions by computing the difference in value ΔQ between the right and left options, choosing right when $\Delta Q > 0$ and left otherwise. Specifically, we write

$$\Delta Q = \Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{1}$$

where, the experimentally controlled variables are $\Delta R = R_{right} - R_{left}$, the difference between the mean of the rewards shown on the forced trials, and ΔI, the difference information available for playing the two options on the first free-choice trial. For simplicity, and because information is manipulated categorically in the Horizon Task, we define ΔI to be +1, -1 or 0, +1 if one reward is drawn from the right option and three are drawn from the left in the [1 3] condition, -1 if one from the left and three from the right, and in [2 2] condition, ΔI is 0. n_{det} and n_{ran} are deterministic noise and random noise respectively.

The other variables are: the spatial bias, b, which determines the extent to which participants prefer the option on the right; the information bonus A, which controls the level of directed exploration; n_{det} denotes the deterministic noise, which is identical on the repeat versions of each game; and n_{ran} denotes random noise, which is uncorrelated between repeat plays and changes every game.

Each subject's behavior in each horizon condition is described by 4 free parameters: the information bonus A, the spatial bias, b, the standard deviation of the deterministic noise, σ_{det}, and the standard deviation of the random noise, σ_{ran} (Table 1, Figure 2). Each of the free parameters is fit to the behavior of each subject using a hierarchical Bayesian approach (Allenby et al., 2005). In this approach to model fitting, each parameter for each subject is assumed to be sampled from a group-level prior distribution whose parameters, the so-called 'hyperparameters', are estimated using a Markov Chain Monte Carlo (MCMC) sampling procedure. The hyper-parameters themselves are assumed to be sampled from 'hyperprior' distributions whose parameters are defined such that these hyperpriors are broad.

The particular priors and hyperpriors for each parameter are shown in Table 1. For example, we assume that the information bonus, A^{is}, for each horizon condition i and for each participant s, is sampled from a Gaussian prior with mean μ_i^A and standard deviation σ_i^A. These prior parameters are sampled in turn from their respective hyperpriors: μ_i^A, from a Gaussian distribution with mean 0 and standard deviation 10, and σ_i^A from an Exponential distribution with parameters 0.1.

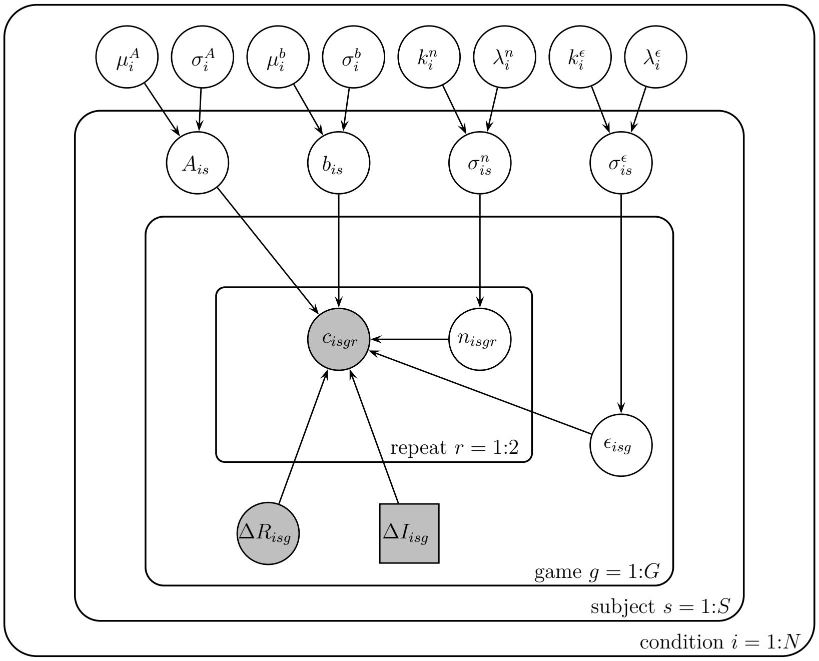
Parameter	Prior	Hyperparameters	Hyperpriors
information bonus, A_{is}	$A_{is} \sim$ Gaussian(μ_i^A, σ_i^A)	$\theta_i^A = (\mu_i^A, \sigma_i^A)$	$\mu_i^A \sim$ Gaussian(0, 100) $\sigma_i^A \sim$ Exponential(0.01)
spatial bias, b_{is}	$b_{is} \sim$ Gaussian(μ_i^b, σ_i^b)	$\theta_i^b = (\mu_i^b, \sigma_i^b)$	$\mu_i^b \sim$ Gaussian(0, 100) $\sigma_i^b \sim$ Exponential(0.01)
deviation of deterministic noise, σ_{isg}^{det}	$\sigma_i \sim$ Gamma(k_i^{det}, λ_i^{det})	$\theta_i^{det} = (k_i^{det}, \lambda_i^{det})$	$k_i^{det} \sim$ Exponential(0.01) $\lambda_i^{det} \sim$ Exponential(10)
deviation of random noise, σ_{isgr}^{ran}	$\sigma_{is} \sim$ Gamma(k_i^{ran}, λ_i^{ran})	$\theta_i^{ran} = (k_i^{ran}, \lambda_i^{ran})$	$k_i^{ran} \sim$ Exponential(0.01) $\lambda_i^{ran} \sim$ Exponential(10)

Table 1: Model parameters, priors, hyperparameters and hyperpriors.

Model fitting using MCMC

The model was fit to the data using Markov Chain Monte Carlo approach implemented in the JAGS package (Depaoli et al., 2016) via the MATJAGS interface (psiexp.ss.uci.edu/research/programs_data/jags). This package approximates the posterior distribution over model parameters by generating samples from this posterior distribution given the observed behavioral data.

In particular we used 4 independent Markov chains to generate 16000 samples from the posterior distribution over parameters (4000 samples per chain). Each chain had a burn in period of 2000 samples, which were discarded to reduce the effects of initial conditions, and posterior samples were acquired at a thin rate of 1. Convergence of the Markov chains was confirmed *post hoc* by eye.



Priors

$\mu_i^A \sim \text{Gaussian}(0, 100)$, $\sigma_i^A \sim \text{Exponential}(0.01)$

$\mu_i^b \sim \text{Gaussian}(0, 100)$, $\sigma_i^b \sim \text{Exponential}(0.01)$

$k_i^n \sim \text{Exponential}(0.01)$, $\lambda_i^n \sim \text{Exponential}(10)$

$k_i^\epsilon \sim \text{Exponential}(0.01)$, $\lambda_i^\epsilon \sim \text{Exponential}(10)$

Subject specific parameters

$A_{is} \sim \text{Gaussian}(\mu_i^A, \sigma_i^A)$

$B_{is} \sim \text{Gaussian}(\mu_i^B, \sigma_i^B)$

$\sigma_{is}^n \sim \text{Gamma}(k_i^n, \lambda_i^n)$

$\sigma_{is}^\epsilon \sim \text{Gamma}(k_i^\epsilon, \lambda_i^\epsilon)$

Stimulus driven noise for repeated game

$\epsilon_{isg} \sim \text{Logistic}(0, \sigma_{is}^\epsilon)$

Free noise for each game

$n_{isgr} \sim \text{Logistic}(0, \sigma_{is}^n)$

Observed choices

$\Delta Q_{isgr} \leftarrow \Delta R_{isg} + A_{is}\Delta I_{isg} + B_{is} + n_{isgr} + \epsilon_{isg}$

$c_{isgr} \sim \text{Bernoulli}\,(Q_{isgr} > 0)$

Figure 2: Schematic of the hierarchical Bayesian model using notation of Lee and Wagenmakers (2014b)

Parameter recovery

To be sure that our fit parameter values were meaningful, we tested the ability of our model fitting procedure to recovery parameters from simulated data. In particular, we simulated choices with the fitted parameters from the Hierarchical Bayesian analysis, and the re-fit the simulated choices to see whether we can recover the parameters.

Results of this parameter recovery procedure are shown in Supplementary Figure S4. As is clear from this figure, parameter recovery is good in terms of correlation between the true vs fitted parameters for all parameters apart from the bias in short horizon condition, which is likely due to this parameter being so close to zero (Supplementary Figure S4 Panel C). The recovery for the noise parameters, σ_{det} and σ_{ran}, is slightly better for horizon 1 than horizon 6. This is because it requires more trials to recover bigger noises, so with the same number of choices it is harder to recover overall bigger noises in horizon 6. In addition we see better recovery for random noise than deterministic noise because we effectively have half as many trials for deterministic noise since we are only generating one sample of deterministic noise for each repeated game pair.

One limitation of our model is that we observed a systematic underestimation of deterministic noise (despite the strong correlation). This is further illustrated in Supplementary Figure S5. In thee simulation with 0 random noise and full deterministic noise, our model perfectly recovered both random and deterministic noise, however in the simulation with full random noise and 0 deterministic noise, our model recovered only XXX $\%$ of deterministic noise, and the remaining shows up as random noise instead (See Supplementary Figure S5). Thus, our model provides a lower bound on deterministic noise and an upper bound on random noise.

Overall, we are able to recover both deterministic and random noises using our model to a satisfactory extent.

Data and code

Behavioral data as well as Matlab code to recreate the main figures from this paper will be made available on the Dataverse website upon publication.

Results

The Repeated-Games Horizon Task

We used a modified version of the ‘Horizon Task’ (Wilson et al., 2014) to show the influence of stimulus-driven ‘deterministic noise’ vs non-stimulus driven ‘random noise’ on people’s decisions (Figure 1). In this task, participants make repeated choices between two slot machines, or ‘one-armed bandits,’ that pay out probabilistic rewards. Because they are initially unsure as to the mean payoff of each bandit, this task requires that participants carefully balance exploration of the lesser known bandit with exploitation of the better known bandit to maximize their overall rewards.

Crucially, before people make their first choice in the Horizon Task, they are given information about the mean payoff from each bandit in the form of four example plays distributed either unequally between bandits (i.e. 1 play of one bandit, 3 plays of the other, the [1 3] condition) or equally (2 plays each, the [2 2] condition). These example plays allow us to manipulate exactly what people know about each option before they make their first choice.

Relative to the original Horizon Task, the key modification here is to give people 'repeated games,' in which they see exact same set of example plays twice in two separate games (separated by several minutes in time so as to avoid detection). By repeating the instructed plays for each game twice, we can set up a situation where (unbeknownst to the participants) they are faced with the exact same explore-exploit choice, with the exact same stimuli twice. Thus, if their behavior is a deterministic function of the stimuli, they will behave identically on the two games, that is their behavior on the two versions of each game will be consistent. Conversely, if their behavior is not driven by a deterministic function of the stimulus, then their choices on the repeated games should be inconsistent.

On average participants played 67.31 repeated games each allowing us to quantify the extent to which their behavior was a deterministic function of the stimulus or not.

Both behavioral variability and information seeking increase with horizon

Before discussing the results for repeated games, we first confirm that the basic behavior in this task is consistent with our previously reported results (Wilson et al., 2014). As in our previous work, we find evidence for two types of exploration in the Horizon Task. Random exploration, which is the main focus of this paper, where exploration is driven by noise, and directed exploration, where exploration is driven by information.

Random exploration is quantified in a model-free way as the probability of choosing the low mean option, $p(\text{low mean})$ in the equal, or [2 2], condition. This value increases with horizon, consistent with the idea that behavior is more random in horizon 6 (t(65) = 6.60, $p < 0.001$ for [1 3], t(65) = 8.14, $p < 0.001$ for [2 2]). Directed exploration, is measured as the probability of choosing the more informative option $p($high info in the unequal, or [1 3], condition. Again this measure increases with horizon, showing that people are more information seeking in horizon 6 (t(65) = 7.00, $p < 0.001$).

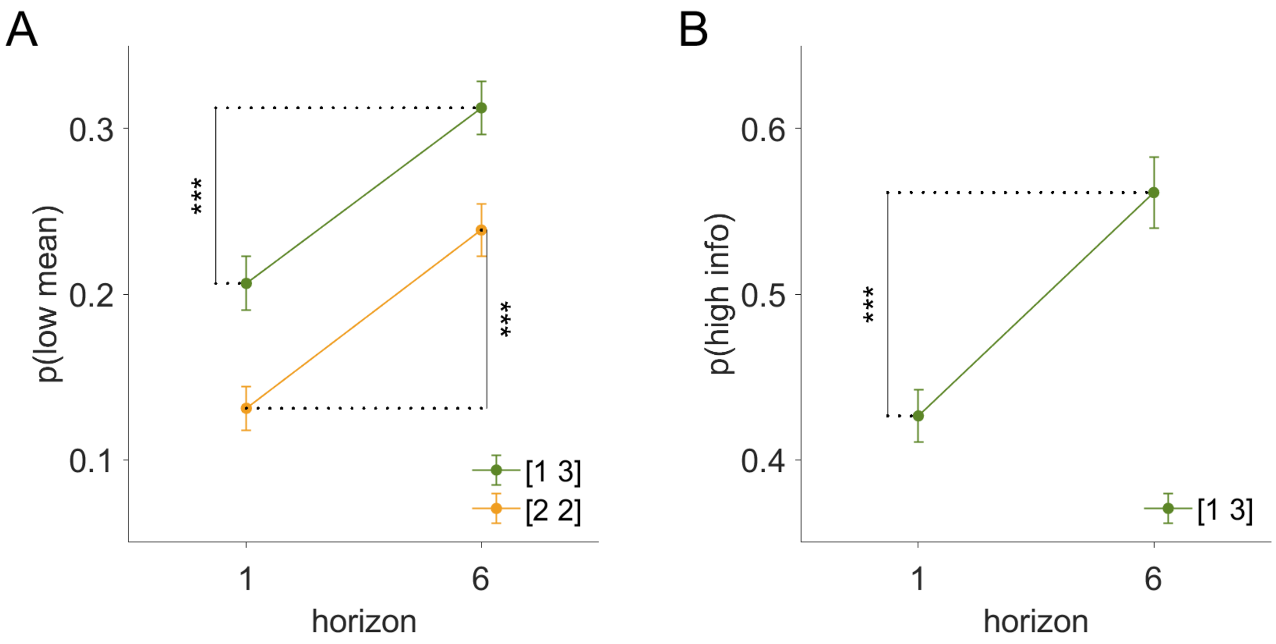


Figure 3: Replication of previous findings. Both $p(\text{low mean})$ (A) and $p(\text{high info})$ (B) increase with horizon suggesting that people use both random and directed exploration in this task.

Model-free analysis shows that random exploration may involve both random and deterministic noise

Next we asked whether participants' choices were consistent or inconsistent in the two repetitions of each game. The idea behind this measure is that purely deterministic noise should lead to consistent choices as the deterministic stimulus is identical both times. Conversely, if choice is not a deterministic function of the stimulus, participants' choices should be independent, and hence more inconsistent across the repetitions of the game.

To quantify choice inconsistency we computed the frequency with which participants made different responses for pairs of repeated games (Figure 4). Using this measure we found that participants made inconsistent choices in both the unequal ([1 3]) and equal ([2 2]) information conditions, suggesting that not all of the noise was stimulus driven (t-test vs zero revealed that inconsistency was greater than zero for all horizon and uncertainty conditions. For [1 3] condition, t(65) = 12.92, $p < 0.001$ for horizon 1, t(65) = 16.76, $p < 0.001$ for horizon 6; For [2 2] condition, t(65) = 9.67, $p < 0.001$ for horizon 1, t(65) = 17.74, $p < 0.001$ for horizon 6). In addition, we found that choice inconsistency was higher in horizon 6 than in horizon 1 for both [1 3] and [2 2] condition (F(1, 196) = 61.19, p < 0.001), suggesting that at least some of the horizon dependent noise is not a deterministic function of the stimulus.

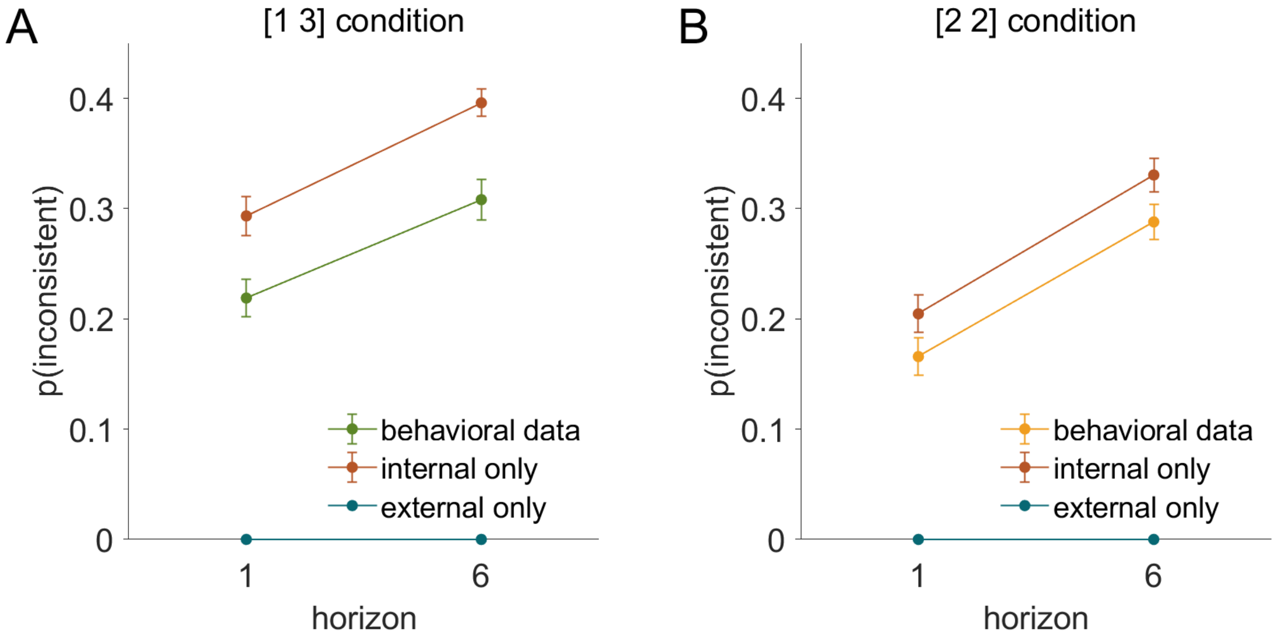


Figure 4: Model-free analysis suggests that both deterministic and random noise contribute to the choice variability in random exploration. For both the [1 3] (A) and [2 2] (B) condition, people show greater choice inconsistency in horizon 6 than horizon 1. However, the extent to which their choices are inconsistent lies between what is predicted by purely deterministic and random noise, suggesting that both noise sources influence the decision.

To gain more quantitative insight into these results, we computed theoretical values for the choice inconsistency for the purely deterministic and purely random noise cases. For purely deterministic noise this computation is simple because people should make the exact same decisions each time in repeated games, meaning that $p(\text{inconsistent}) = 0$ in this case. For purely random noise, the two games should be treated independently, allowing us to compute the choice inconsistency in terms of the probability of choosing the low mean option, $p(\text{low mean})$, as

$$\begin{aligned} p(\text{consistent}) &= p(\text{low mean})^2 + p(\text{high mean})^2 \\ &= p(\text{low mean})^2 + (1 - p(\text{low mean}))^2 \\ \text{hence,} \quad p(\text{inconsistent}) &= 1 - p(\text{consistent}) = 2p(\text{low mean})(1 - p(\text{low mean})) \end{aligned}$$

As shown in Figure 4, people's behavior falls in between the pure deterministic noise prediction and the pure random noise prediction Specifically, behavior is different from pure random noise prediction in the both the [1 3] condition (t(65) = 5.60, p $<$ 0.001 for horizon 1, t(65) = 5.62 p $<$ 0.001 for horizon 6) and [2 2] condition (t(65) = 3.86, p $<$ 0.001 for horizon 1, t(65) = 3.42, p $<$ 0.001 for horizon 6). Likewise, behavior is different from pure deterministic noise prediction in both the [1 3] condition (t(65) = 12.92, p $<$ 0.001 for horizon 1, t(65) = 16.76, p $<$ 0.001 for horizon 6) and the [2 2] condition (t(65) = 9.67, p $<$ 0.001 for horizon 1, t(65) = 17.74, p $<$ 0.001 for horizon 6). This suggests that at least some of the 'noise' is a deterministic function of the stimulus, although from this analysis it is not clear whether this deterministic noise increases with horizon or not.

Model-based analysis shows deterministic noise changes with horizon

To more precisely quantify the contribution of stimulus-driven deterministic noise, we turned to model fitting. We modeled behavior on the first free choice of the Horizon Task using a version of the logistic choice model in (Wilson et al., 2014) that was modified to differentiate deterministic noise from and random noise. In particular, we assume that in repeated games, the value of stimulus-driven deterministic noise is frozen whereas random noise is drawn independently both times.

Overview of model

As with our model-free analysis, the model-based analysis focuses only on the first free-choice trial since that is the only free choice when we have control over the experience participants have about two bandits. To model participants' choices on this first free-choice trial, we assume that they make decisions by computing the difference in value ΔQ between the right and left options, choosing right when $\Delta Q > 0$ and left otherwise. Specifically, we write

$$\Delta Q = \Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{2}$$

where, the experimentally controlled variables are $\Delta R = R_{right} - R_{left}$, the difference between the mean of rewards shown on the forced trials, and ΔI, the difference information available for playing the two options on the first free-choice trial. For simplicity, and because information is manipulated categorically in the Horizon Task, we define ΔI to be +1 if one reward is drawn from the right option and three are drawn from the left in the [1 3] condition, -1 if one from the left and three from the right, and in [2 2] condition, ΔI is 0. n_{det} and n_{ran} are deterministic noise and random noise respectively which are assumed to come from logistic distributions with mean 0.

The subject-and-condition-specific parameters are: the spatial bias, b, which determines the extent to which participants prefer the option on the right; the information bonus A, which controls the level of directed exploration; n_{det} denotes the deterministic noise, which is identical on the repeat versions of each game; and n_{ran} denotes random noise, which is uncorrelated between repeat plays and changes every game.

For each pair of repeated games, the set of forced-choice trials are exactly the same, so the deterministic noise, n_{det}, should be the same while the random noise, n_{ran} may be different. This is exactly how we distinguish deterministic noise from random noise. In symbolic terms, for repeated games i and j, $n^i_{det} = n^j_{det}$ and $n^i_{ran} \neq n^j_{ran}$.

Model fitting

We used hierarchical Bayesian analysis to fit the parameters of the model (see Figure 2 for an graphical representation of the model in the style of Lee and Wagenmakers (2014a)). In particular, we fit values of the information bonus A, spatial bias B, variance of random noise σ^2_{ran}, and variance of deterministic noise, σ^2_{det} for each participant in each horizon. Model fitting was performed using the MATJAGS and JAGS software (Depaoli et al., 2016, Steyvers, 2011) with full details given in the Methods.

Model fitting results

Posterior distributions over the group-level means of the deterministic and random noise variance are shown in Figure 5. Consistent with our model-free results, we see that both random and deterministic noise variances are non-zero and that random noise is about 2-3 times larger than the deterministic noise. In addition, we find that both random and deterministic noise increase with horizon. This increase was larger for random noise (M = 4.55, 100% of samples showed an increase in random noise with horizon) than deterministic noise (M = 1.78, 98.12% of samples showed an increase in deterministic noise with horizon). But intriguingly the relative increase in both types of noise was the same (Figure 6). That is when we compute the relative increase in deterministic noise with horizon, $\sigma_{det,horizon6}/\sigma_{det,horizon1}$, it is almost identical to the relative increase in random noise with horizon $\sigma_{ran,horizon6}/\sigma_{ran,horizon1}$.

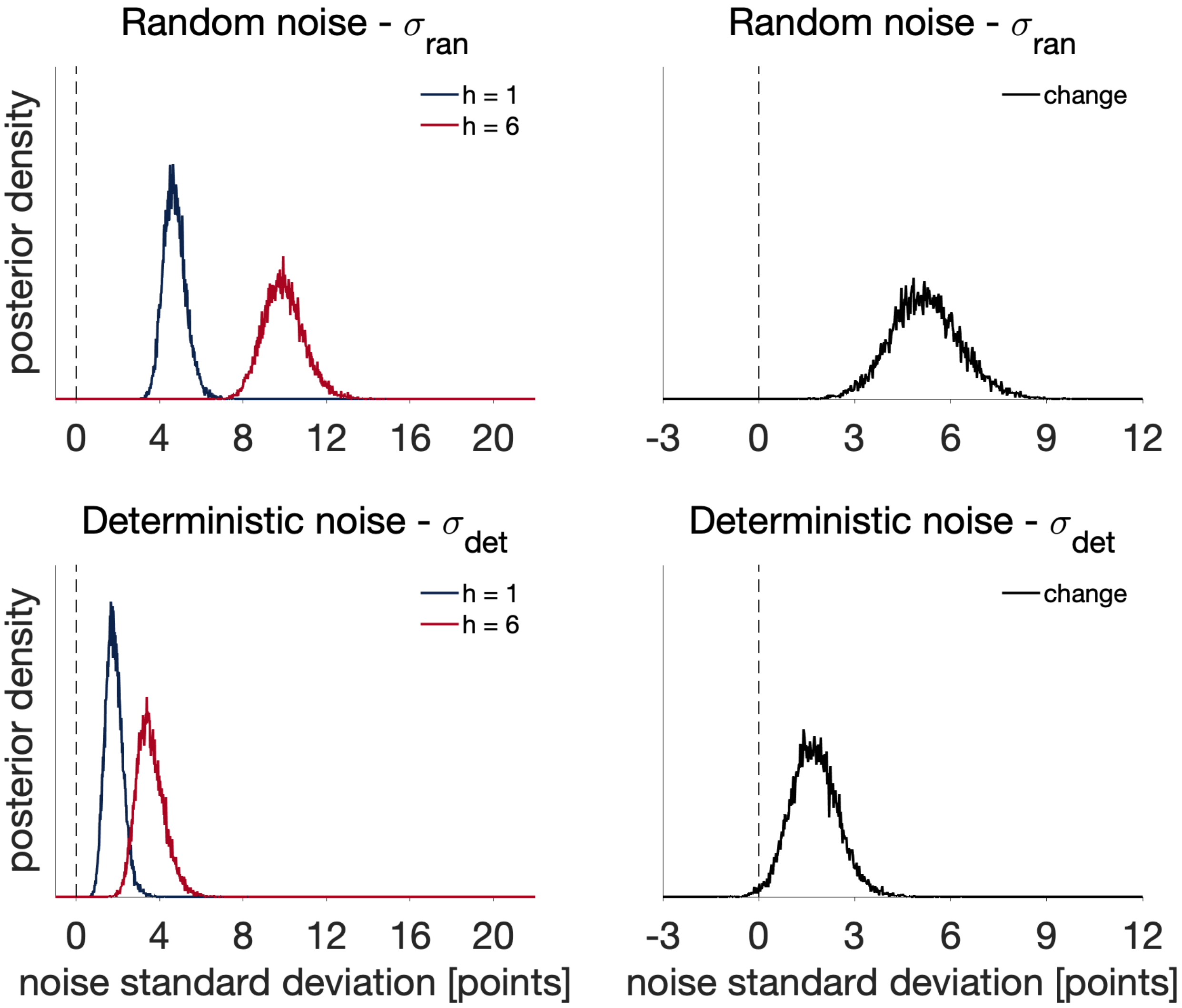


Figure 5: Model based analysis showing the posterior distributions over the group-level mean of the standard deviations of random and deterministic noise. Both random (A, B) and deterministic (C,D) noises are nonzero (A, C) and change with horizon (B, D). However, random noise has both a greater magnitude overall (A, C) and a greater change with horizon (B, D) than deterministic noise.

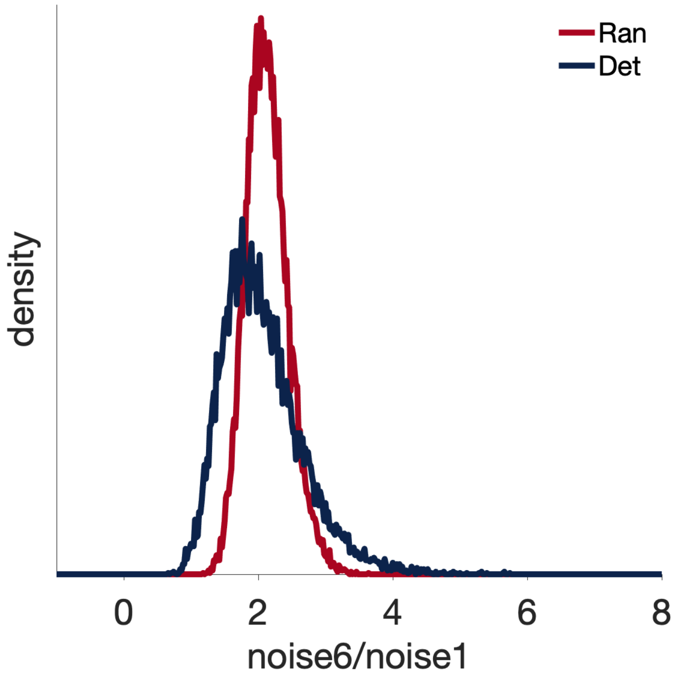


Figure 6: Model based analysis showing the posterior distributions over the ratio of the group-level mean of the standard deviations of random and deterministic noise between horizon 6 and horizon 1 respectivelly. The ratio in the standard deviations of noise between horizon 6 and horizon 1 is similar for random and deterministic noise.

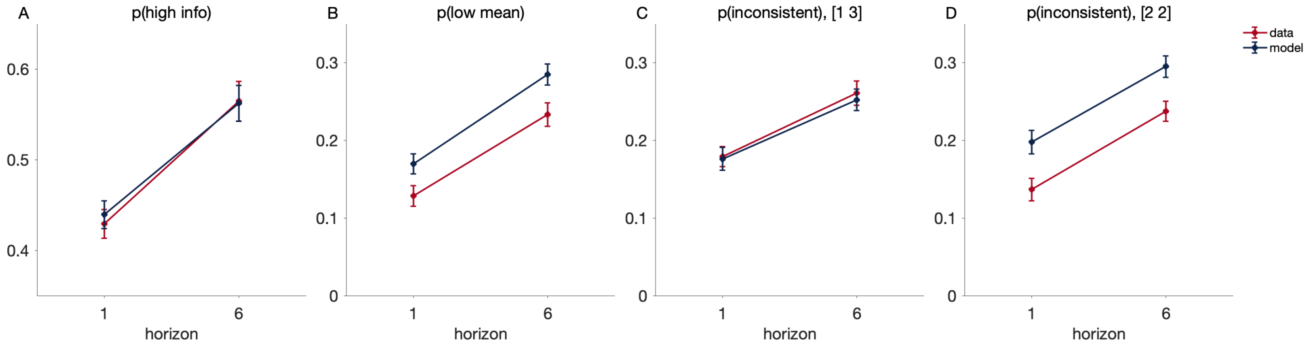


Figure 7: Our model accounts for all qualitative patterns of the data, namely, p(high info) and p(low mean) increase as a function of horizon, p(inconsistent) increases as a function of horizon for both [1 3] and [2 2] conditions and lies between the pure random and pure deterministic noise prediction.

Posterior predictive checks

In addition to fitting the model to behavior, it is also important to check whether the model captures the qualitative patterns of the data (Wilson and Collins, 2019) — specifically how p(high info), p(low mean) and p(inconsistent) change with horizon.

To perform this 'posterior predictive check,' we created a set of simulated data by taking the subject-level parameters from the hierarchical Bayesian fits and having the model play the same sequence of games as seen by the subjects. We then applied the same model-free analysis as described in the previous sections to this simulated data set and compared the model's behavior to that of participants. As shown in Figure 7, the model can account for all qualitative patterns in the data — the increase in p(high info), p(low mean), and p(inconsistent) with horizon, and that p(inconsistent) is in between pure random and pure deterministic noise. The quantitative agreement is almost perfect for p(high info), but the model seems to systematically underestimate p(low mean) and p(inconsistent), although the discrepancy is small (underestimating p(low mean) by 0.047 or 27.30%, and p(inconsistent) by 0.027 or 14.27%).

To check whether all aspects of the model were necessary to reproduce the qualitative pattern of findings, we also built and fit five additional versions of the model. These models varied whether deterministic and random noise are present or not and whether either types of noise is dependent on horizon. As shown in Supplementary Figure S5, only one of these models, where random noise is horizon dependent but deterministic noise is not, can capture the full qualitative pattern of responding (Note model F does not capture that p(inconsistent) is between pure random and deterministic?). However, the quantitative fit to the data is not as good (Supplementary Figure S5).

Discussion

In this paper, we investigated whether random exploration is really random. That is, to what extent is it driven deterministically by aspects of the stimulus we have previously ignored when measuring 'decision noise'. Using a version of the Horizon Task with repeated games, we found evidence that at least some of the noise in random exploration could be explained by such 'deterministic noise.' In particular, we found that deterministic noise accounted for XXX% of the overall variability in people's behavior and increased with horizon - a hallmark of an exploratory process. This suggests that at least some of the apparent randomness in random exploration is not random at all.

So where does this leave randomness in random exploration? Well, the remaining XXX % of the noise could be random or it could be deterministic because we can't control everything. In particular, while we controlled many aspects of the stimulus across repeated games (e.g. the outcomes and the order of the forced trials), we could not perfectly control *all* stimuli the participant received, which would vary, for example, based on exactly what they were looking at or whether they were scratching their nose. Thus conceptually, our estimate of deterministic noise is a lower bound. Conversely, our estimate of random noise is an upper bound as these 'missing' sources of deterministic noise would be interpreted as random noise in our model. In addition, our estimation method also systematically underestimate deterministic noise (part of the deterministic noise will show up as random noise in model fitting, see Supplementary figure SX), so methodologically our model also provides a lower bound of deterministic noise and an upper bound on random noise. Although there is still a considerable window for truly stochastic processes in the brain to be driving random exploration, our results suggest that at least some of the randomness is driven by a deterministic process instead.

Regardless of whether the remaining XXX$\%$ is deterministic or random, the fact that thee horizon change in the two noises are proportional to each other suggests a possible mechanism for random exploration, Specifically, a reduction in the strength with which reward drives the choice. In particular, in our analysis we excluded a scalar on $DeltaR$, if we include that then we get

$$\Delta Q = \beta \Delta R + A \Delta I + b + n_{det} + n_{ran} \tag{3}$$

A decrease in $beta$ here would be equivalent to an increase in variance of both deterministic and random noise. Such reduced coding of reward would be consistent with Ebitz et al. (2017) (NEED CHECK).

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem, 2011.

Greg Allenby, Peter Rossi, and Robert McCulloch. Hierarchical bayes models: A practitioners guide. 01 2005.

G. Aston-Jones and J. D. Cohen. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28:403–450, 2005.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Machine Learning. 47(235), 2002. URL `https://doi.org/10.1023/A:1013689704352`.

J. Banks, M. Olson, and D. Porter. An experimental analysis of the bandit problem. *Economic Theory*, 10:55, 1997.

Debabrota Basu, Pierre Senellart, and Stéphane Bressan. Belman: Bayesian bandits on the belief–reward manifold, 2018.

D. H. Brainard. The Psychophysics Toolbox. *Spat Vis*, 10(4):433–436, 1997.

M. S. Brainard and A. J. Doupe. What songbirds teach us about learning. *Nature*, 417(6886):351–358, May 2002.

J.S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. *Advances in Neural Information Processing Systems*, 2:211–217, 1990.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf`.

N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, Jun 2006.

Sarah Depaoli, James P. Clifton, and Patrice R. Cobb. Just another gibbs sampler (jags): Flexible software for mcmc implementation. *Journal of Educational and Behavioral Statistics*, 41 (6):628–649, 2016. doi: 10.3102/1076998616664876. URL `https://doi.org/10.3102/1076998616664876`.

J. Drugowitsch, V. Wyart, A. D. Devauchelle, and E. Koechlin. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6):1398–1411, Dec 2016.

B. Ebitz, T. Moore, and T. Buschman. Bottom-up salience drives choice during exploration. *Cosyne*, 2017.

M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.*, 12(8):1062–1068, Aug 2009.

Samuel J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 2018. ISSN 18737838. doi: 10.1016/j.cognition.2017.12.014.

J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *J. R. Statist. Soc. B*, 41(2):148–177, 1979.

J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, 1974.

M. Jepma, R. G. Verdonschot, H. van Steenbergen, S. A. Rombouts, and S. Nieuwenhuis. Neural mechanisms underlying the induction and relief of perceptual curiosity. *Front Behav Neurosci*, 6:5, 2012.

M. H. Kao, A. J. Doupe, and M. S. Brainard. Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature*, 433(7026):638–643, Feb 2005.

Waitsang Keung, Todd A Hagen, and Robert C Wilson. Regulation of evidence accumulation by pupil-linked arousal processes. *bioRxiv*, 2018. doi: 10.1101/309526. URL `https://www.biorxiv.org/content/early/2018/04/28/309526`.

J.R. Krebs, A. Kacelnik, and P. Taylor. Test of optimal sampling by foraging great tits. *Nature*, 275:27–31, 1978. doi: doi:10.1038/275027a0.

M.D. Lee, S. Zhang, M.N. Munro, and M. Steyvers. Psychological models of human and optimal performance on bandit problem. *Cognitive Systems Research*, 12:164–174, 2011.

Michael D. Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014a. doi: 10.1017/CBO9781139087759.

Michael D. Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014b. doi: 10.1017/CBO9781139087759.

R. Meyer and Y. Shi. Choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, 41:817, 1995.

S. Nieuwenhuis, D. J. Heslenfeld, N. J. von Geusau, R. B. Mars, C. B. Holroyd, and N. Yeung. Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage*, 25(4):1302–1309, May 2005.

E. Payzan-LeNestour and P. Bossaerts. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput. Biol.*, 7(1):e1001048, Jan 2011.

E. Payzan-Lenestour and P. Bossaerts. Do not Bet on the Unknown Versus Try to Find Out More: Estimation Uncertainty and ”Unexpected Uncertainty” Both Modulate Exploration. *Front Neurosci*, 6:150, 2012.

D. G. Pelli. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis*, 10(4):437–442, 1997.

M. Steyvers. matjags. An interface for MATLAB to JAGS version 1.3. 2011. URL `http://psiexp.ss.uci.edu/research/programs_data/jags/`.

M. Steyvers, M. Lee, and E. Wagenmakers. A Bayesian analysis of human decisionmaking on bandit problems. *Journal of Mathematical Psychology*, 53:168, 2009.

D. G. R. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, and A. Y. Karpova. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, 159 (1):21–32, Sep 2014.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL `http://www.jstor.org/stable/2332286`.

Christopher M. Warren, Robert C. Wilson, Nic J. van der Wee, Eric J. Giltay, Martijn S. van Noorden, Jonathan D. Cohen, and Sander Nieuwenhuis. The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, 12(4):1–17, 04 2017. doi: 10.1371/journal.pone.0176034. URL `https://doi.org/10.1371/journal.pone.0176034`.

C. J. C. H. Watkins. Learning from delayed rewards. *Ph.D thesis, Cambridge University*, 1989.

R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen*, 143(6):2074–2081, Dec 2014.

Robert C. Wilson and Anne G.E. Collins. Ten simple rules for the computational modeling of behavioral data. *eLife*, 2019. ISSN 2050084X. doi: 10.7554/eLife.49547.

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