What is 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 3, 2018

Abstract

[150 words for Nature Human Behavior, currently 162] Human decision making appears to contains components that are inherently random. While this behavioral variability has traditionally been seen as noise, recent work suggests that random choices may actually be adaptive. An example arises when deciding between exploring unknown options or exploiting options we know well. Theory shows that a little randomness in explore-exploit decisions is remarkably effective, leading to greater rewards overall. Meanwhile, experiments show that people use such 'random exploration' in practice, increasing their behavioral variability when it is more valuable to explore. Despite this progress, the nature of adaptive decision noise is unknown and it is unclear whether it is generated internally, from stochastic processes in the brain, or externally, from stochastic stimuli in the world? Here we show that, while both types of noise are present in explore-exploit decisions, random exploration is dominated by internal noise. This suggests that random exploration depends on adaptive noise processes in the brain which are subject to (perhaps unconscious) cognitive control.

Introduction

Imagine trying to decide where to go to dinner with a friend. You can go to your favorite restaurant that you both really enjoy and always go to, or you can try the new restaurant that just opened and about which you know nothing. 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 incredibly hard (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) ADDED TWO NEW PAPERS ON THOMPSON SAMPLING, NOT SURE WHETHER THEY ARE OKAY.

Recently we have shown that humans appear to actually use random exploration and actively adapt their decision noise to solve simple explore-exploit problems (Wilson et al., 2014). The key manipulation in the task is the horizon condition, i.e. the number of decisions remaining 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 place. Using such a horizon manipulation we found that people have greater decision noise in the long versus the short horizon condition.

However, a key limitation of this work was that the source of the decision noise used for exploration is unknown. Of particular interest is whether the adaptive decision noise that is linked to exploration arises externally, in the input from the world, or is generated internally, within the brain. In the restaurant case, an example of external noise would be if you happen to spot an old friend walking in to one of the restaurants. Conversely, internal noise would arise from stochastic neural processes tossing a metaphorical coin in your head. Previous work makes a strong case for both types of noise being relevant to behavior. For instance, external, stimulus-driven noise is thought to be a much greater source of choice variability in perceptual decisions than internal noise (Brunton et al., 2013). Conversely internal, neural noise is thought to drive exploratory singing behavior in song birds (Kao et al., 2005) and the generation and control of this internal noise has been linked to specific neural structures (Kao et al., 2005).

In this paper, we investigate which source of noise, internal vs external, drives random exploration in humans in a simple explore-exploit task adapted from our previous work (Wilson et al., 2014). To distinguish between the two types of noise, we had people make the exact same explore-exploit decision twice. If decision noise is purely externally driven, then people's choices should be identical both times, that is their choices should be consistent, since the stimulus is the same both times. Meanwhile, if noise is internally driven their choices should be inconsistent, since the internal noise is different both times. By analyzing behavior on this task in both a model-free and model-based manner, we show that, while both types of noise are present in explore-exploit decisions, the contribution of internal noise to random exploration far exceeds that contributed by the stimulus.

Results

The Repeated-Games Horizon Task

We used a modified version of the 'Horizon Task' (Wilson et al., 2014) to show the influence of internal vs external 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) or equally (2 plays each). These example plays allow us to manipulate exactly what people know about each option before they make their first choice. Thus, by giving people the exact same example plays twice in two separate games (separated by several minutes in time so as to avoid detection), the example plays allow us to probe how participants respond to the exact same explore-exploit choice twice.

These 'repeated games' are the key manipulation in this paper and allow us to distinguish between external and internal sources of noise. Specifically, if noise is externally driven, then choices on repeated games should be consistent. Conversely if noise is internally driven, then choices on repeated games should be inconsistent.

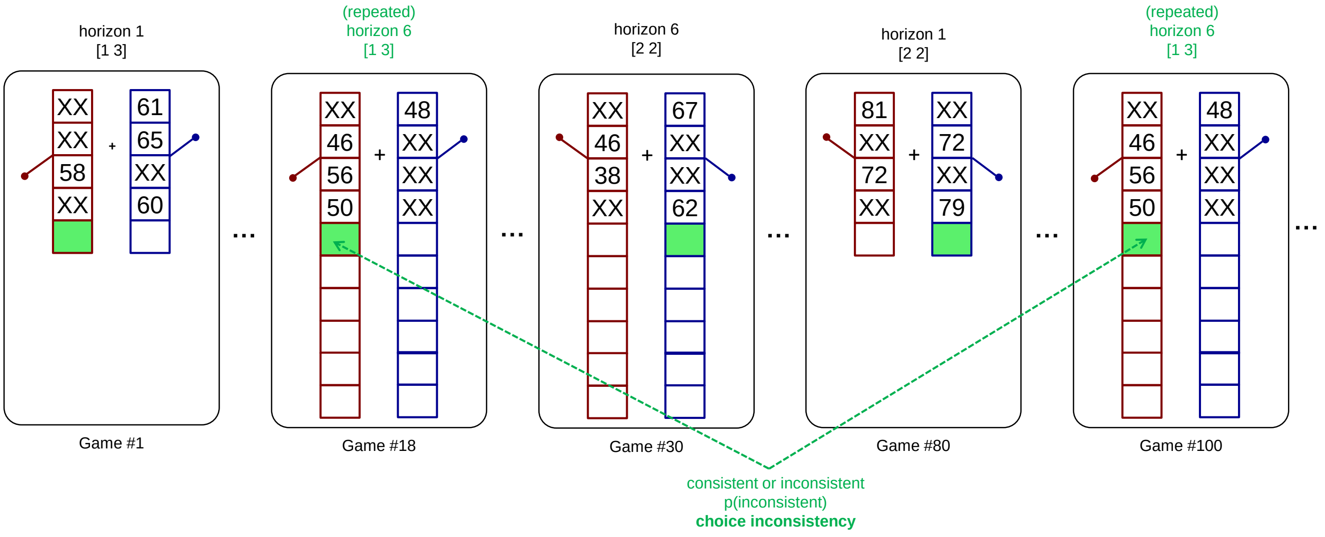


Figure 1: A key additional manipulation here is repeated games. Each pair of repeated games with identical example trials will appear twice during the experiment. We setup the repeated games such that they are at least 5 games apart from each other. A model-free measure of choice inconsistency which reflex the underlying decision noise is defined as the proportion of inconsistent choices for repeated games.

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(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(59) = 6.17, p < 0.001 for [1 3], t(59) = 7.26, 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(59)=6.88, p < 0.001).

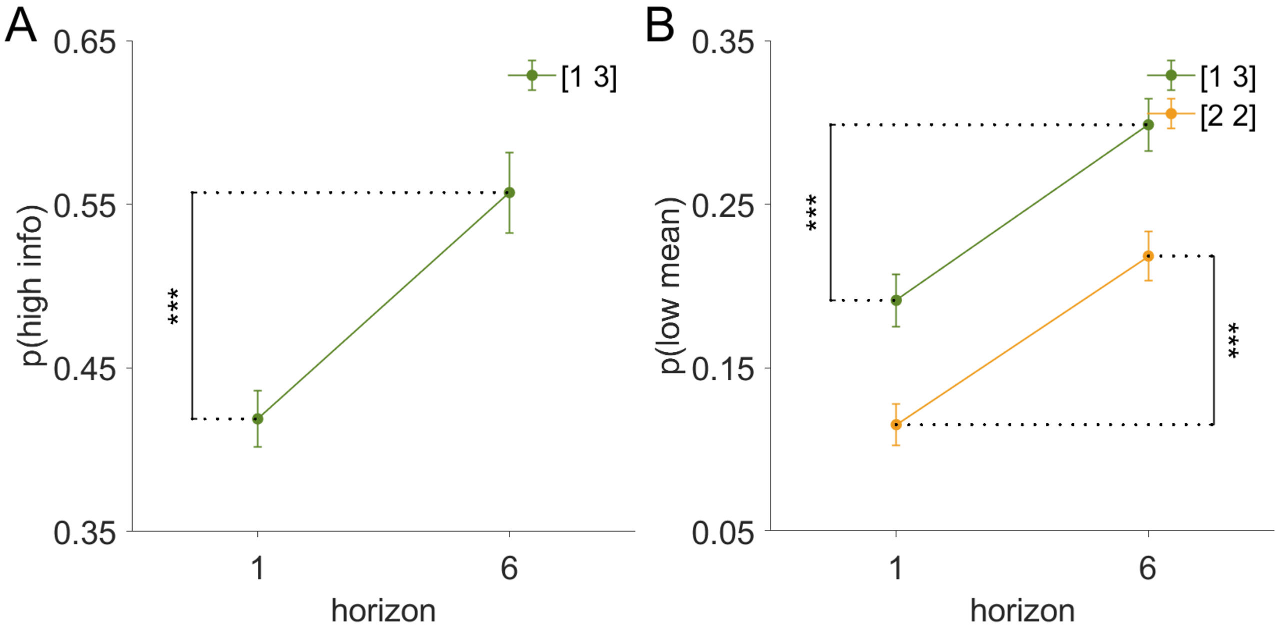


Figure 2: Both directed and random exploration increase with horizon. Choice inconsistency also increases with horizon for both [1 3] and [2 2] conditions.

Model-free analysis shows that random exploration may involve both internal and external 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 external noise should lead to consistent choices as the external stimulus is identical both times. Conversely, purely internal noise should lead to independent choices, and hence more inconsistent choices both times.

To quantify choice inconsistency we computed the frequency with which participants made different responses for pairs of repeated games (Figure 3). Using this measure we found that participants made inconsistent choices in both the unequal ([1 3]) and equal ([2 2]) information conditions (STATS), suggesting that not all of the noise was stimulus driven. 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, 236) = 37.98, p < 0.001), suggesting that at least some of the horizon dependent noise is internal.

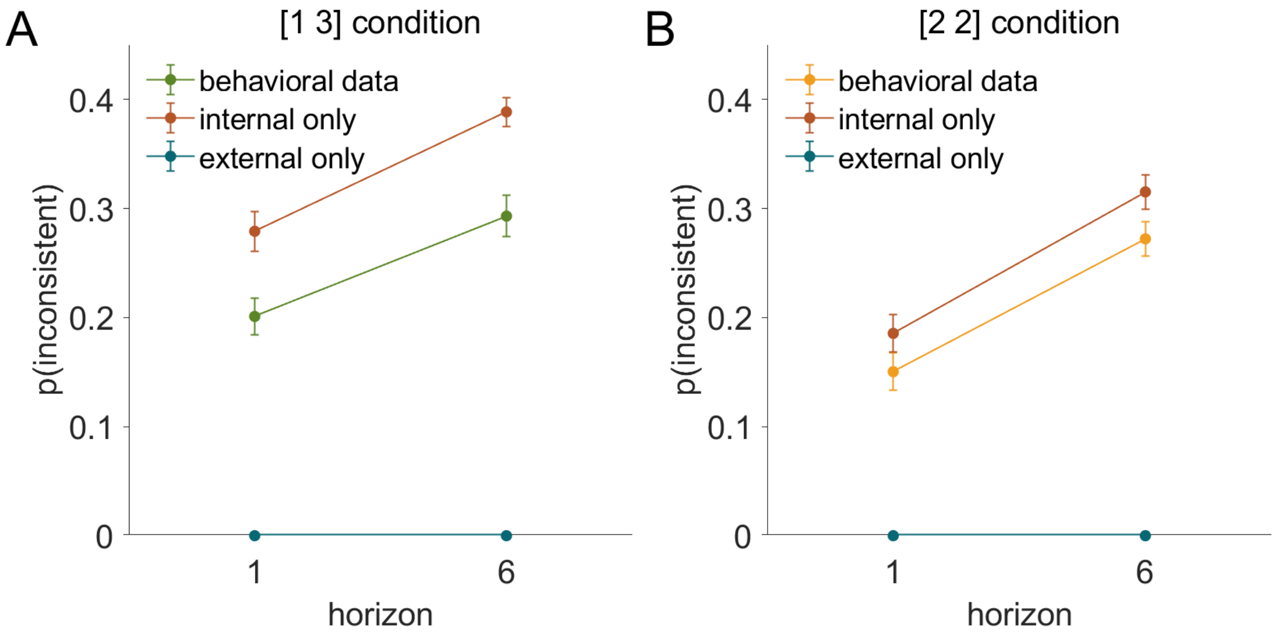


Figure 3: Both external and internal noise contribute to the choice variability in random exploration. For both [1 3] and [2 2] condition, there is a significant difference between people's behavior and predicted choice inconsistency assuming that only external noise exists where people should behave identically in repeated games. Also, there is a significant difference between people's behavior and predicted choice inconsistency assuming that only internal noise exists where people treat repeated games independently.

To gain more quantitative insight into these results, we computed theoretical values for the choice inconsistency for the purely external and purely internal noise cases. For purely external 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 internal 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

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

As shown in Figure 3, people's behavior falls in between the pure external noise prediction and the pure internal noise prediction (Behavior is different from pure internal noise prediction in both [1 3] condition, t(59) = 5.10, p < 0.001 for horizon 1, t(59) = 5.65, p < 0.001 for horizon 6; and [2 2] condition, t(59) = 3.49, p < 0.001 for horizon 1, t(59) = 3.31, p < 0.001 for horizon 6. Behavior is different from pure external noise prediction in both [1 3] condition, t(59) = 11.78, p < 0.001 for horizon 1, t(59) = 15.41, p < 0.001 for horizon 6; and [2 2] condition, t(59) = 8.76, p < 0.001 for horizon 1, t(59) = 17.02, p < 0.001 for horizon 6.), suggesting that both external and internal noise are present in driving this choice inconsistency. Since choice inconsistency only reflects internal noise, Figure 3 suggests that internal noise increases with horizon.

Model-based analysis shows that random exploration is dominated by internal noise

To more precisely quantify internal and external 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 internal and external noises. In particular, we assume that in repeated games, external noise remains the same whereas internal noise can change.

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 information bias between the 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

$$\Delta Q = \Delta R + A\Delta I + b + n_{ext} + n_{int} \tag{1}$$

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, -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_{ext} and n_{int} are external noise and internal noise respectively.

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_{ext} denotes the external, external noise, which is identical on the repeat versions of each game; and n_{int} denotes internal 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 external noise, n_{ext}, should be the same while the internal noise, n_{int} may be different. This is exactly how we distinguish external noise from internal noise. In symbolic terms, for repeated games i and j, $n^i_{ext} = n^j_{ext}$ and $n^i_{int} \neq n^j_{int}$.

Model fitting

We used hierarchical Bayesian analysis to fit the parameters of the model (see Figure 9 for an graphical representation of the model in the style of Lee and Wagenmakers (2014)). In particular, we fit values of the information bonus A, spatial bias B, variance of internal noise σ^2_{int}, and variance of external noise, σ^2_{ext} 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 external and internal noise variance are shown in Figure 4. Consistent with our model-free results, we see that both internal and external noise variances are non-zero and that internal noise is about 2-3 times larger than the external noise. In addition, we find that internal noise increases dramatically with horizon (M = 4.55, p < 0.001) whereas there is only a slight increase of external noise (M = 1.78, p = 0.009). Taken together, these results suggest that random exploration is dominated by internal noise.

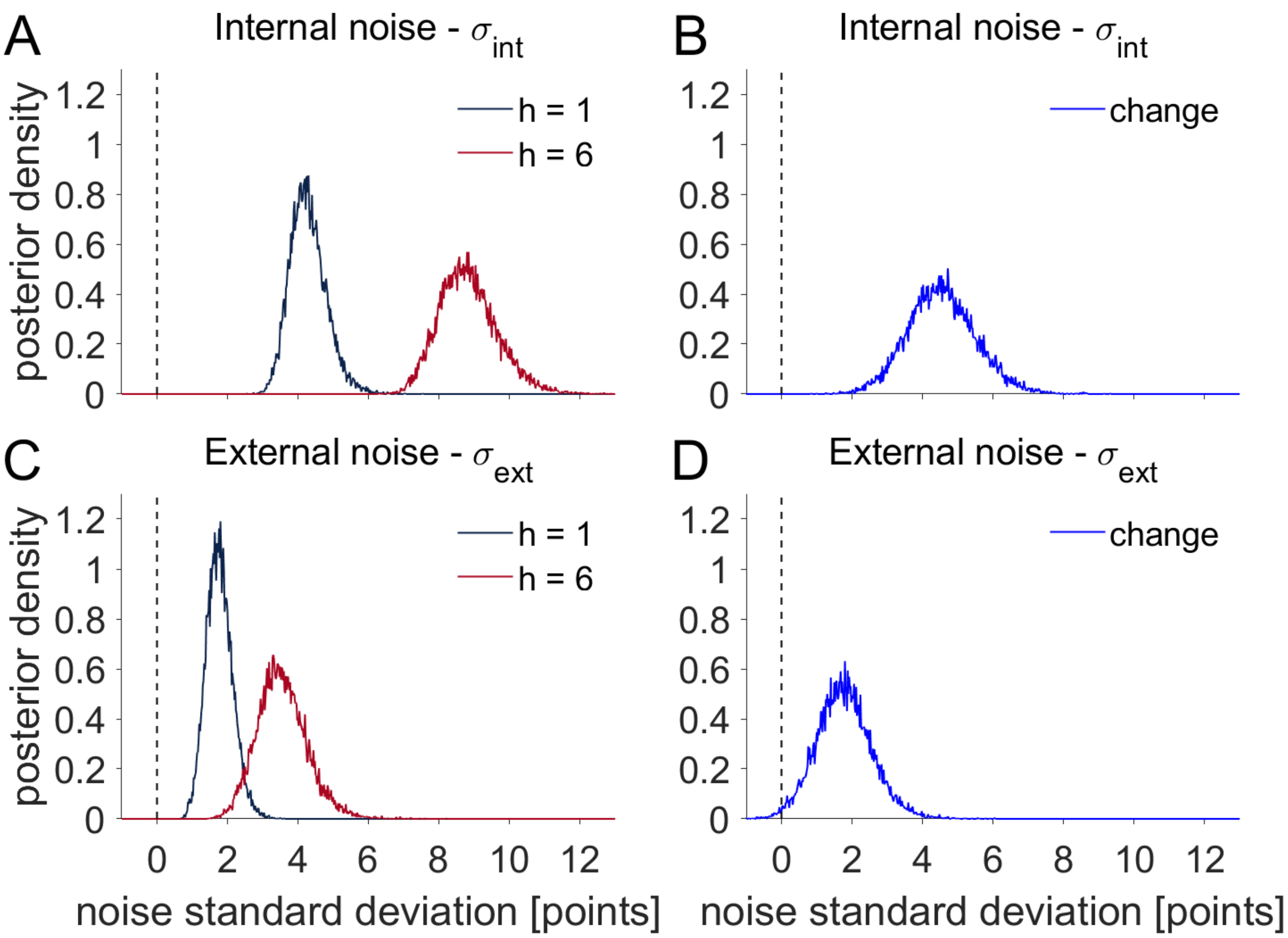


Figure 4: Posterior distributions over the group-level means of the external and internal noise standard deviation. Both internal and external noises are nonzero, and internal noise has a much greater magnitude. Posterior distributions over the group-level means of the change of external and internal noise standard deviation. Internal noise increases significantly with horizon, external noise increases with horizon as well.

Discussion

In this paper, we investigated whether random exploration is driven by internal noise, putatively arising in the brain, or external noise, arising from the environment. Using a version of the Horizon Task with repeated games, we found that evidence for both types of noise in explore-exploit decisions. In addition, we see that both internal and external noise increase with horizon, but that the horizon effect is much larger for internal noise. Taken together our results suggest that random exploration, i.e. the use and adaptation of decision noise to drive exploration, is driven primarily by internal noise.

Perhaps the main limitation of this work is in the interpretation of the different types of noise as being internal and external. 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, our estimate of external noise is likely a lower bound. Likewise, our estimate of internal noise is likely an upper bound as these 'missing' sources of external noise would be interpreted as internal noise in our model. Despite this, it seems hard to imagine that these additional noise sources could be enough to account for the large differences between internal and external noise that we found in Figure 4, where internal noise is 2-3 times the size of external noise.

Taken at face value, the horizon-dependent increase in internal noise is consistent with the idea that random exploration is driven by intrinsic variability in the brain. This is in line with work in the bird song literature in which song variability during song learning has been tied to neural variability arising from specific areas of the brain (Brainard and Doupe, 2002, Kao et al., 2005). In addition, this work is consistent with a recent report from Ebitz et al. (2017) in which the behavioral variability of monkeys in an 'explore' state was also tied to internal rather than external sources of noise.

Whether such a noise-controlling area exists in the human brain is less well established, but one candidate theory (Aston-Jones and Cohen, 2005) suggests that norepinephrine (NE) from the locus coeruleus may play a role in modulating internal levels of noise. Indeed, manipulation of the NE system has been found to change behavioral variability in both humans and other animals in a variety of tasks (Keung et al., 2018, Tervo et al., 2014). In addition there is some evidence that NE plays a direct role in random exploration (Warren et al., 2017), although this finding is complicated by other work showing no effect of NE drugs on exploration (Jepma et al., 2012, Nieuwenhuis et al., 2005)

More generally, our finding that internal noise dominates behavioral variability over external noise, is consistent with findings of Drugowitsch et al. (2016). In particular these authors show that randomness in behavior arises from imperfections in mental inference, that happen inside the brain, rather than in peripheral processes such as sensory processing and response selection. This suggests that most noise in behavior is generated internally and that this may arise from computational errors in computing the correct strategy. In the context of the Horizon Task, such computational errors would likely be larger in the long horizon condition as the correct course of action in these cases is much harder to compute.

Methods

Participants

80 participants (ages 18-25, 43 male, 37 female) from the University of Arizona undergraduate subject pool participated in the experiment. 20 were excluded on the basis of performance, using the same exclusion criterion as in (Wilson et al., 2014). This left 60 for the main analysis. Note that including the 20 badly performing subjects did not change the main results (Supplementary Figures 5-7)

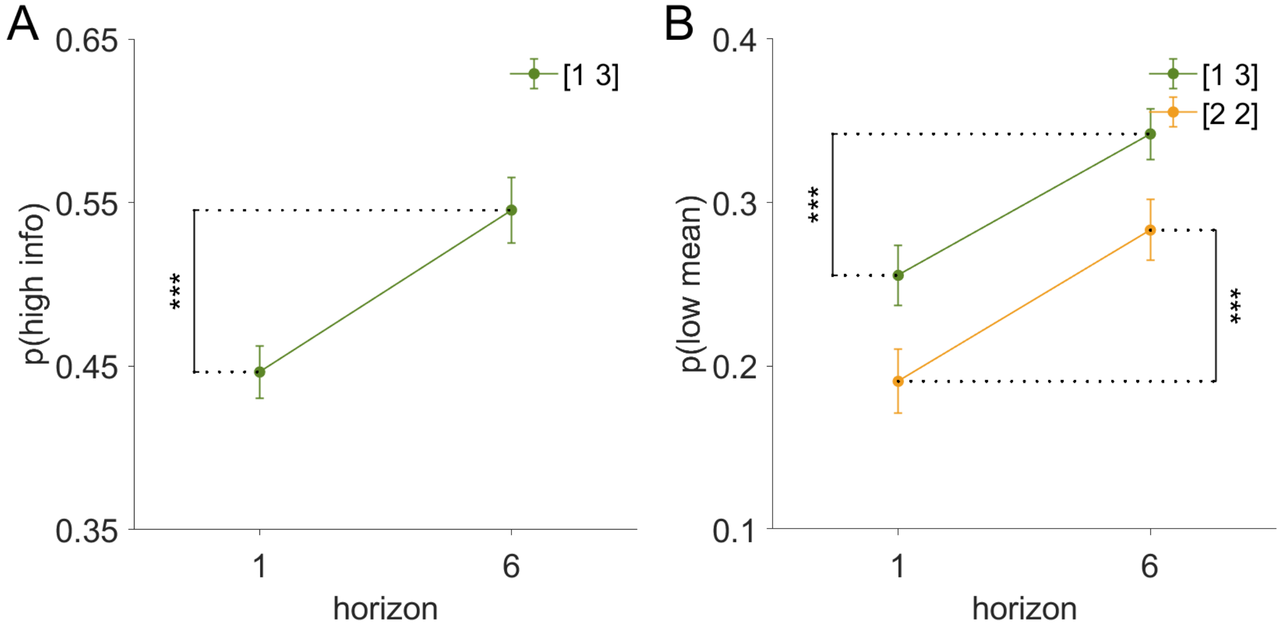


Figure 5: Both directed and random exploration increase with horizon. Choice inconsistency also increases with horizon for both [1 3] and [2 2] conditions.

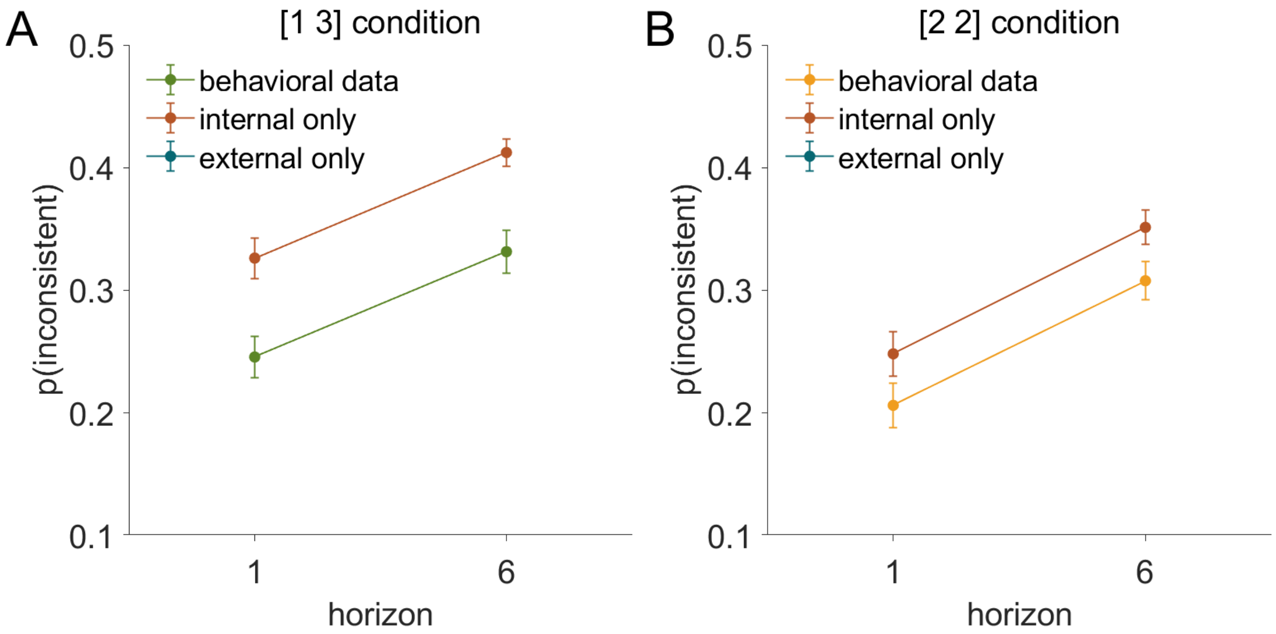


Figure 6: Both external and internal noise contribute to the choice variability in random exploration. For both [1 3] and [2 2] condition, there is a significant difference between people's behavior and predicted choice inconsistency assuming that only external noise exists where people should behave identically in repeated games. Also, there is a significant difference between people's behavior and predicted choice inconsistency assuming that only internal noise exists where people treat repeated games independently.

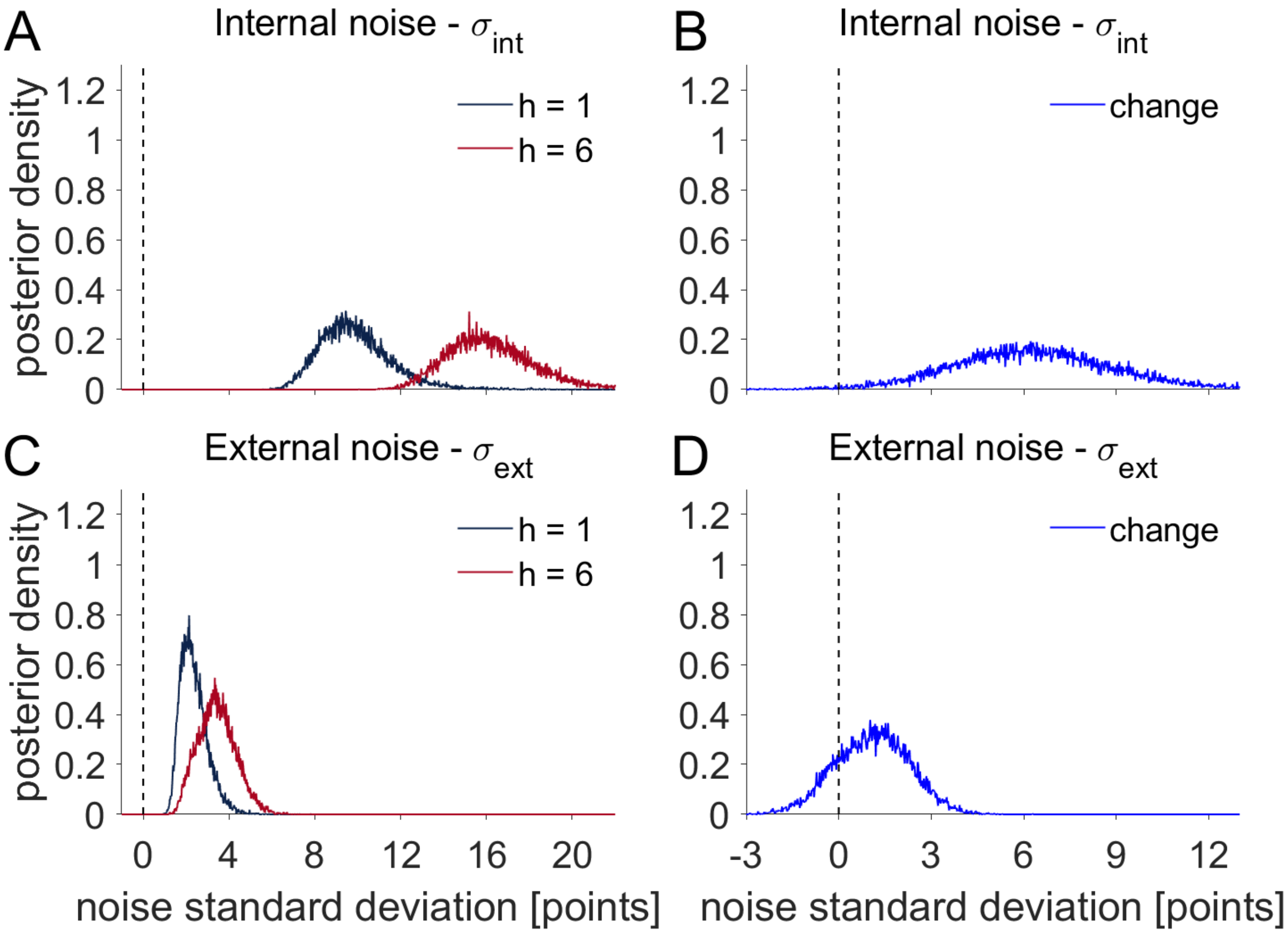


Figure 7: Posterior distributions over the group-level means of the external and internal noise standard deviation. Both internal and external noises are nonzero, and internal noise has a much greater magnitude. Posterior distributions over the group-level means of the change of external and internal noise standard deviation. Internal noise increases significantly with horizon, external noise increases with horizon as well.

Task

The task was a modified version of the Horizon Task (Wilson et al., 2014). 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 Gaussian 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).THIS IS ONLY THE TASK TIME, NOT INCLUDING BASELINE PUPIL OR THE TWO CURIOSITY SURVEY OR RA SETUP TIME

As in the original paper, 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 will show up indicating the horizon for that game. For the first 4 games - the instructed games, we highlight the box on one of the bandits to instruct the participant to choose that option, they have to press the corresponding key to reveal the outcome. From the 5^{th} 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 they could press either the left arrow key or right arrow key to indicate their choice of left or right bandit. The score feedback was presented for 300ms. The task was programmed using Psychtoolbox in MATLAB (Brainard, 1997, Pelli, 1997). (See Figure 8)

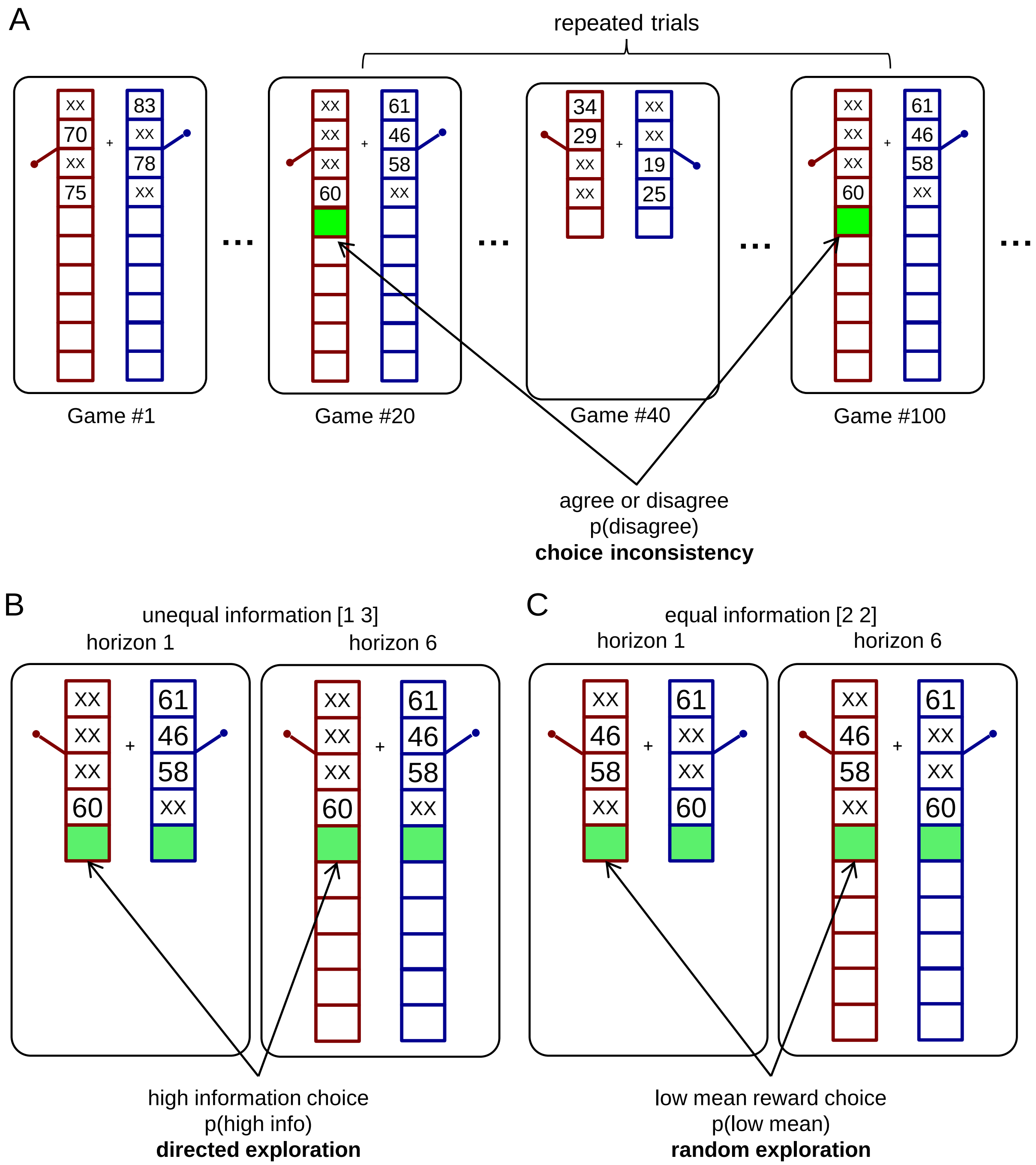


Figure 8: Time line of a horizon 6 game. The game begins with four forced choice trials in which participants are instructed which option to play. After these forced trials they are free to choose between the two options for the rest of the game (6 trials in this case). I WILL SEND YOU THE POWERPOINTS TOO

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). (See Figure 8)

Data and code

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 internal and external noise. In particular, we assume that in repeated games, external noise remains the same whereas internal 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

$$\Delta Q = \Delta R + A \Delta I + b + n_{ext} + n_{int} \tag{2}$$

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, -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_{ext} and n_{int} are external noise and internal 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_{ext} denotes the external, external noise, which is identical on the repeat versions of each game; and n_{int} denotes internal 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 external noise, σ_{ext}, and the standard deviation of the internal noise, σ_{int} (Table 1, Figure 9). 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.

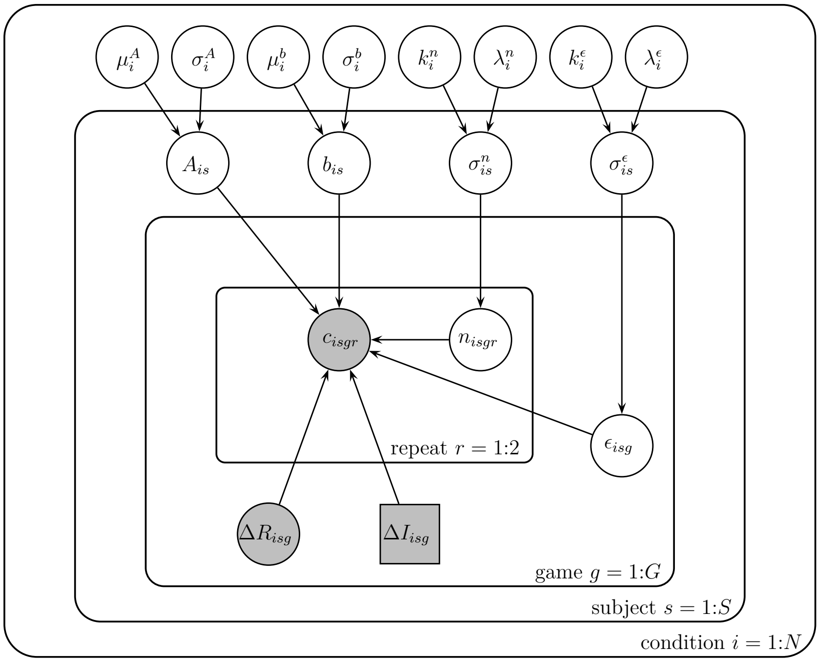
Parameter	Prior	Hyperparameters	Hyperpriors
information bonus, A_{is}	$A_{is} \sim$ Gaussian(μ_i^A, σ_i^A)	$\theta_i^A = (\mu_i^A, \sigma_i^A)$	$\mu_i^A \sim$ Gaussian(0, 100) $\sigma_i^A \sim$ Exponential(0.01)
spatial bias, b_{is}	$b_{is} \sim$ Gaussian(μ_i^b, σ_i^b)	$\theta_i^b = (\mu_i^b, \sigma_i^b)$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

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

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

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

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

Subject specific parameters

$A_{is} \sim$ Gaussian(μ_i^A, σ_i^A)

$B_{is} \sim$ Gaussian(μ_i^B, σ_i^B)

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

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

Stimulus driven noise for repeated game

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

Free noise for each game

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

Observed choices

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

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

Figure 9: Hierarchical Bayesian model

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 the choices to see whether we can recover the parameters.

Results of this parameter recovery procedure are shown in Figure 10. As is clear from this figure, parameter recovery is good for all parameters apart from the bias, which is likely due to this parameter being so close to zero (Figure 10 Panel C, D). The recovery for the noise parameters, σ_{ext} and σ_{int}, 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 internal noise than external noise because we effectively have half as many trials for external noise since we are only generating one sample of external noise for each repeated game pair. Overall, we are able to recover both external and internal noises using our model to a satisfactory extent.

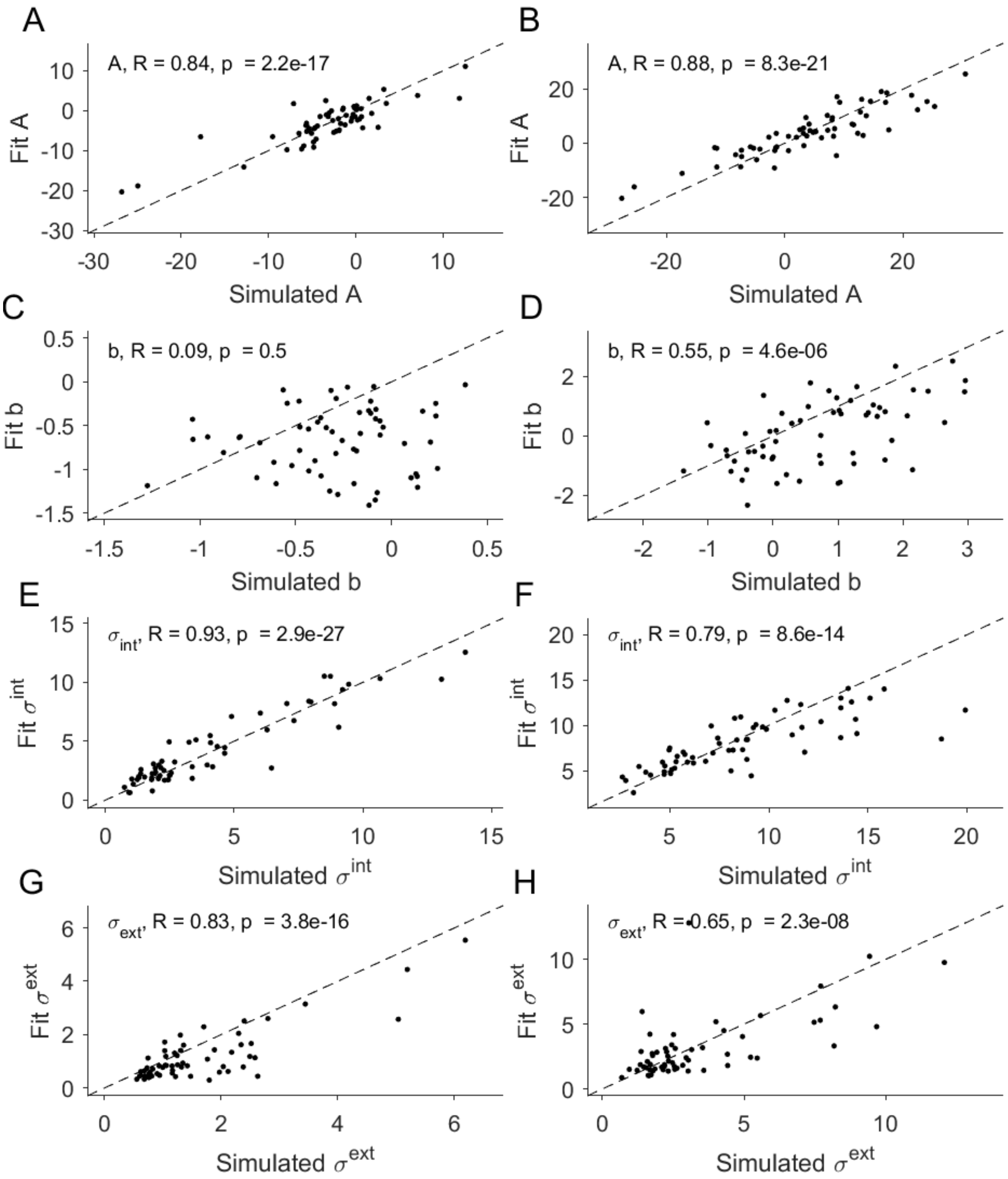


Figure 10: Parameter recovery over the subject-level means of information bonus(A), spatial bias(b), internal noise variance(σ_{int}) and external noise variance(σ_{ext})

THIS PLOT CHANGES EVERY TIME I RUN IT. THE R^2 VALUE FOR INFO BONUS AND INTERNAL NOISE ARE SIMILAR ALL THE TIME, BUT FOR EXTERNAL NOISE IT RANGES FROM 0.2 TO 0.7, FOR BIAS, FROM 0 TO 0.3

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.

B. W. Brunton, M. M. Botvinick, and C. D. Brody. Rats and humans can optimally accumulate evidence for decision-making. Science, 340(6128):95–98, Apr 2013.

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.

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, 2014. 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.

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.