PLOS Computational Biology

Separating random and deterministic sources of computational noise in explore-exploit decisions

--Manuscript Draft--

Manuscript Number:	PCOMPBIOL-D-24-00833R1
Full Title:	Separating random and deterministic sources of computational noise in explore-exploit decisions
Short Title:	Separating random and deterministic noise in random exploration
Article Type:	Research Article
Keywords:	explore-exploit; behavioral variability; random exploration; Decision making
Corresponding Author:	Siyu Wang The University of Arizona Tucson, Arizona UNITED STATES OF AMERICA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	The University of Arizona
Corresponding Author's Secondary Institution:	
First Author:	Siyu Wang
First Author Secondary Information:	
Order of Authors:	Siyu Wang
	Robert C. Wilson
Order of Authors Secondary Information:	
Abstract:	Human decision making is inherently variable. While this variability is often seen as a sign of suboptimal behavior, both theoretical work in machine learning and empirical human studies suggest 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 can encourage us to explore options we might otherwise ignore. In line with this idea, several studies have found evidence that people increase their behavioral variability when it is valuable to explore. A key question, however, is whether this variability in so-called '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 exploreexploit task in which, unbeknownst to them, participants are presented with the exact same choice twice, we provide a partial answer to this question. By modeling behavior in this task, we were able to estimate a lower bound on the amount of variability that is deterministically driven by the stimulus and an upper bound on the amount of variability that is random. Using this approach, we find evidence that at least 14% of the variability in random exploration in our studied task can be accounted for by deterministic processing of the stimulus. Conversely, this suggests that up to 86% of the variability is truly 'random', although it is still possible that this variability is driven by deterministic factors not related to the stimulus. Finally, our results suggest that both deterministic and random sources of variability change proportionally to each other as the value of exploration increases, suggesting that a common noise gating mechanism may be at play in random exploration.
Suggested Reviewers:	Anne Collins, PhD Associate Professor, UC Berkeley: University of California Berkeley annecollins@berkeley.edu She is an expert in computational modeling of behavior.
	Vincent Costa, PhD Assistant Professor, ONPRC: Oregon Health & Science University Oregon National

	Primate Research Center costav@ohsu.edu He is an expert in explore-exploit decision making.
	Becket Ebitz, PhD Professeur adjoint, University of Montreal: Universite de Montreal r.becket.ebitz@umontreal.ca He is an expert in explore-exploit decision making.
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Government Employee Are you or any of the contributing authors an employee of the United States government? Manuscripts authored by one or more US Government employees are not copyrighted, but are licensed under a CC0 Public Domain Dedication, which allows unlimited distribution and reuse of the article for any lawful purpose. This is a legal requirement for US Government employees. This will be typeset if the manuscript is accepted for publication.	Yes - One or more authors are employees of the U.S. government.
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Computational Biology* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate.	Yes

Funded studies Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?	
Please add funding details. as follow-up to "**Financial Disclosure** Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Computational Biology* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?"	This work is supported by National Institute on Aging grants awarded to Robert C Wilson (R01 AG061888, R56 AG061888). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Please select the country of your main research funder (please select carefully as in some cases this is used in fee calculation).	UNITED STATES - US

as follow-up to "**Financial Disclosure** Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Computational Biology* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?"	
Competing Interests On behalf of all authors, disclose any competing interests that could be perceived to bias this work. This statement will be typeset if the manuscript is accepted for publication. Review the instructions link below and PLOS Computational Biology's competing interests policy to determine what information must be disclosed at submission.	The authors have declared that no competing interests exist.
Data and Code Availability From the time of publication, Authors are	Behavioral data as well as MATLAB codes to recreate the main figures from this paper will be made available upon publication in https://github.com/wangxsiyu/RW_RandomDeterministicNoise.git

required to make fully available and without restriction all data and computational code underlying their findings. Please see our PLOS Data Policy page for detailed policy information, and our Code Sharing page for specific information on code sharing. A **Data Availability Statement**, detailing where the data (and code, if applicable) can be accessed, is required at first submission. Insert your Data Availability Statement in the box below. The statement you provide **will be published in the article**, if accepted. Please see the Data Reporting section of our submission guidelines for instructions on what you need to include in your Data Availability Statement. If the data and code are all contained in your submission files, please state: *All relevant data are within the manuscript and its Supporting Information files.* **PLOS allows rare exemptions to address legal and ethical concerns.** If you have legal or ethical restrictions, please detail these in your Data Availability Statement below for the Journal team to consider.	

We would like to thank the reviewers for carefully reading our manuscript and providing detailed and useful comments. We have addressed all comments and believe that the additional details and analyses have strengthened the manuscript.

Reviewer's Responses to Questions

Comments to the Authors:

Please note here if the review is uploaded as an attachment.

Reviewer #1: This manuscript describes a cognitive modeling study of human choice behavior in a well-known task (the Horizon task) developed by the last author. In this study, the authors ask whether part of the decision noise that is captured by the standard cognitive model used to fit human choices is not genuinely random (unpredictable), but rather deterministic (predictable). To split decision noise in the Horizon task into random and deterministic terms, the authors apply a repeated-trial approach that has been described and applied in tightly connected contexts to show that at least 14% of random exploration is due to deterministic biases that affect choices in the same way across repetitions of the same trial. More interestingly, the authors report that not only random, but also deterministic sources of noise increase with horizon length, an effect that the authors discuss in terms of a decrease in the decision gain of the reward term (rather than an increase in the decision gain of the information bonus term).

I found the manuscript to be very well written and structured in terms of analyses and results, the authors adequately motivate their study in the introduction and the (repeated-trial) approach they have chosen to use to address their research question. The methods also appear well suited to provide statistical support for their findings, and their discussion of the results (in particular the joint increase of random and deterministic noise components with horizon length) is interesting from a cognitive perspective. I nevertheless have a few comments below which the authors should address in my opinion to make the manuscript stronger and better reflect the existing literature that has used in recent years the exact same approach - in very similar contexts - to decompose choice variability into random and deterministic components.

Comment 1.1: * The current version of the manuscript is missing recent references (copied below) that describe (ref. 1 and 2) and apply/discuss (ref. 3 and 4) the same (repeated-trial) approach to cognitive problems that are tightly connected to the one studied here. These references should be cited in the revised manuscript to provide additional theoretical (and empirical) background about this bias-variance separation approach. They would also provide additional findings that can be used to discuss the findings obtained by the authors in the present study:

1/ Wyart V, Koechlin E (2016) Choice variability and suboptimality in uncertain environments. Current Opinion in Behavioral Sciences 11, 109-115. doi:10.1016/j.cobeha.2016.07.003

2/ Wyart V (2018) Leveraging decision consistency to decompose suboptimality in terms of its ultimate predictability. Behavioral and Brain Sciences 41, e248. doi:10.1017/S0140525X18001504 - Commentary on Rahnev D, Denison RN (2018) Suboptimality in perceptual decision making. Behavioral and Brain Sciences 41, e223.

doi:10.1017/S0140525X18000936

3/ Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V (2019) Computational noise in reward-guided learning drives behavioral variability in volatile environments. Nature Neuroscience 22(12), 2066-2077. doi:10.1038/s41593-019-0518-9

4/ Findling C, Wyart V (2021) Computation noise in human learning and decision-making: origin, impact, function. Current Opinion in Behavioral Sciences 38, 124-132. doi:10.1016/j.cobeha.2021.02.018

Response 1.1:
We thank the reviewer for bringing up these relevant references. Reference 3 was already cited in the original manuscript. We have added the other references to the introduction and the discussion.

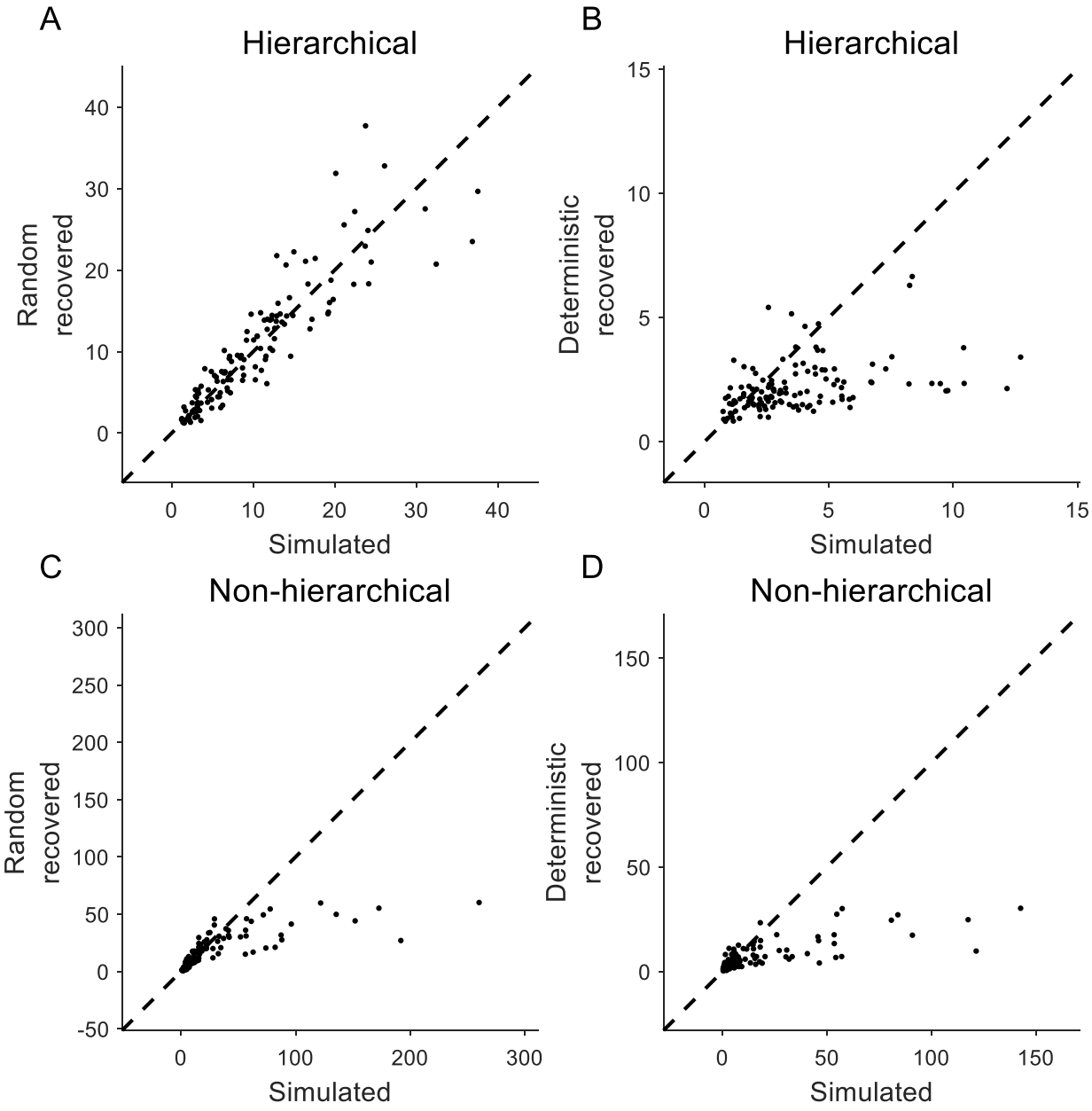
Comment 1.2: * The authors have applied parameter recovery and posterior predictive checks to check whether their fitting procedure is capable of estimating parameter values, and of predicting the key features of the observed human choice behavior. These two procedures have been described as critically important by Wilson and Collins (2019, eLife) and by Palminteri, Wyart and Koechlin (2017, Trends in Cognitive Sciences - missing reference which should ideally be cited where posterior predictive checks are first described). They reveal that the deterministic noise term is underestimated by the fitting procedure, and that the best-fitting model overestimates p(low mean) and p(inconsistent) in the [2,2] condition - which is the most 'basic' condition without information bonus.

The authors do not discuss these biases of the fitting procedure, but it would be important to understand why it is the case. Could it be due to the hierarchical fitting approach used by the authors? Why did the authors choose this hierarchical fitting approach over and above a simpler, independent (subject-wise) fitting approach? The theoretical merits of a hierarchical fitting approach are clear and obvious, but could the authors employ the simpler subject-wise fitting approach to check that the same biases of the fitting procedure remain present (and therefore that they are not triggered by the hierarchical fitting approach)?

Response 1.2:
Non-hierarchical fits
The reason we chose the hierarchical approach was due to the difficulty of estimating deterministic noise which requires many trials across participants to properly converge. In our model, instead of using softmax to output a choice probability, to separately estimate deterministic and random noises, we had to sample both types of noises and had to sample a frozen noise for each repeated pair of games. This made it more challenging for the MCMC to converge with limited data, compared to fitting more commonly used softmax-based models.

In fact, if we use a non-hierarchical approach, the mean estimates of the subject-level noises remain at very large values due to the large variance in the posterior. We assumed a non-

informative prior that spans a large range of possible noise values. Since the non-hierarchical fits rely only on the limited number of trials from each individual subject, and the amount of data per subject is not enough for the model to converge to meaningful values, the posteriors remain broad and we get unreasonably large noise estimates.



Bias in parameter recovery

Nevertheless, as requested by the reviewer, we performed parameter recovery on non-hierarchical fits to see if it resolves the bias in parameter recovery. In the above figure, the top row shows parameter recovery for hierarchical fits (note that noise estimate is between 0-40), and the bottom row is for non-hierarchical fits (note that noise estimate is up to 300), left column is for random noise, and right column is for deterministic noise. As you can see in the

above figure, non-hierarchical fitting procedure does not fix the underestimation of deterministic noise. In fact, it also underestimates random noise.
We think the underestimation of deterministic noise is partially due to limited data. To properly estimate large noises, more data is needed. As we only have half as many trials for deterministic noise as for random noise, we don't have enough data to reliably estimate large values of deterministic noise.

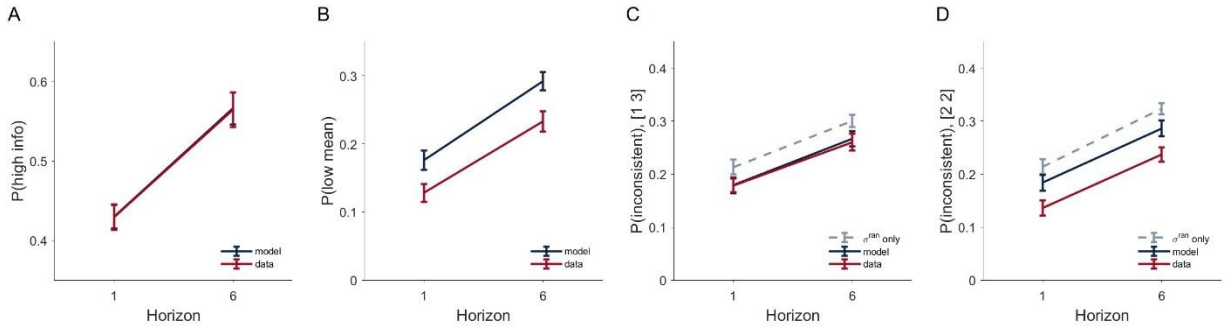
We acknowledged this limitation and stated in our manuscript that our method provides a lower bound for deterministic noise (as opposed to a faithful recovery).

Bias in posterior checks

We cited the reference suggested by the reviewer for posterior checks.

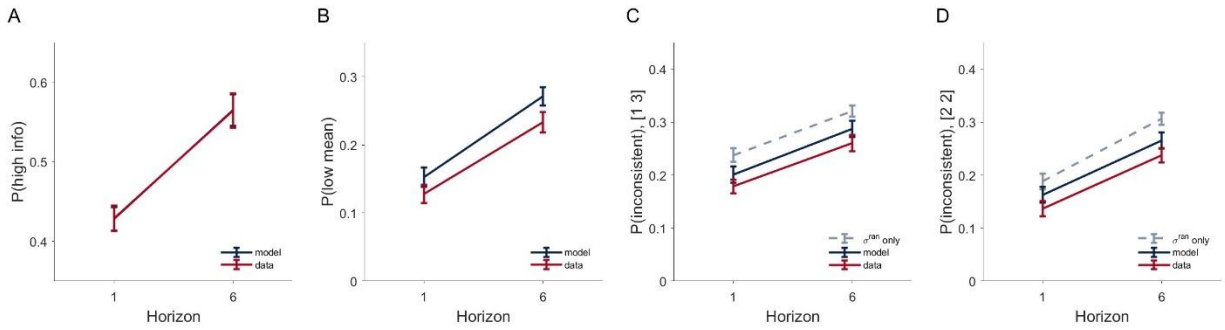
For the overestimation of p(low mean) and p(consistent) in [2 2] condition in posterior checks, we have identified two factors that led to the mismatch:

0. For comparison, we attach here the posterior check figure from the initial submission.

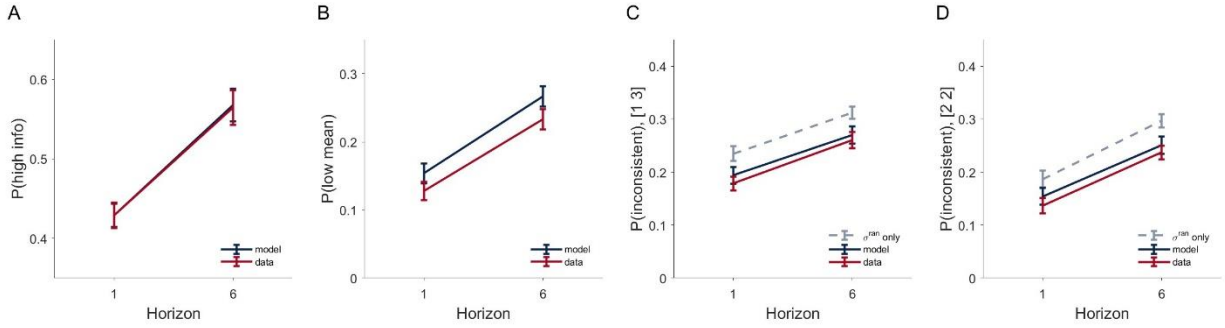


1. In our model, we assumed that the variances of deterministic and random noises are from a constant distribution for both the [1 3] and [2 2] information conditions. Since our model overestimates p(low mean), it suggests that we are overestimating noise in [2 2] condition. It is possible that people have higher noises in [1 3] condition compared to [2 2] condition, by assuming the [1 3] and [2 2] condition share the same noise distribution, it will lead us to overestimate noise in [2 2] condition.

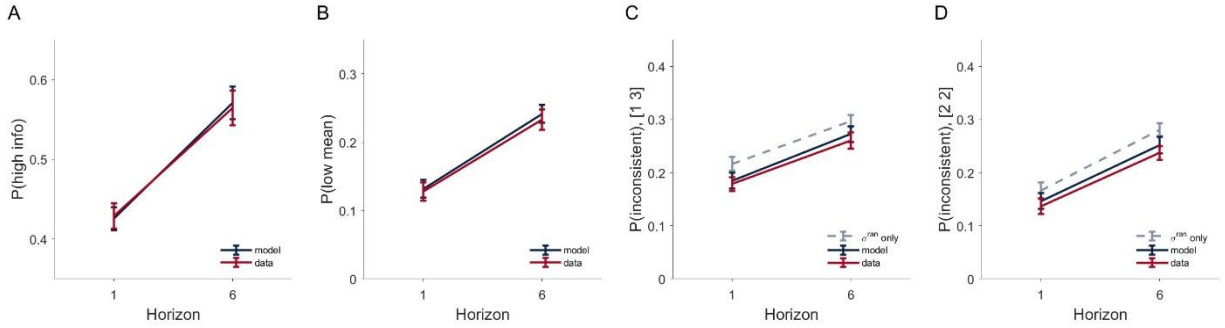
   To fix this, we fit a variant of our model in which we separately estimate the variances of random and deterministic noises in [1 3] and [2 2] conditions. Indeed, we observe higher overall noise level for [1 3] condition. When performing posterior checks with this model, the mismatch between data and model becomes smaller in the [2 2] condition.



2. In our model, the subject-level noises are assumed to follow a gamma distribution (to ensure positiveness), the posteriors are right skewed (maximal likelihood estimation or mode is smaller than the mean), in the original analysis, we simulated data from each participant using the mean of the subject-level posterior for both deterministic and random noises, however, because of the skewness, simulating data with the "mean" is nosier than simulating with the true distribution. Simulating from the true distribution requires taking expected value over all possible noise values, for ease of implementation, we simulated data by taking random samples from the posterior distribution (instead of using the mean). The simulation was repeated 50 times and then averaged, below is what we get. Indeed, the mismatch between data and model further decreased. (we adopted this method in the main text, as this is the closest to the data generation process in the model)



If we use the mode, which corresponds to the maximal likelihood estimation estimate, we can get the model simulation to closely match the data:



Comment 1.3: * The authors appear to take for granted that the joint increase of random and

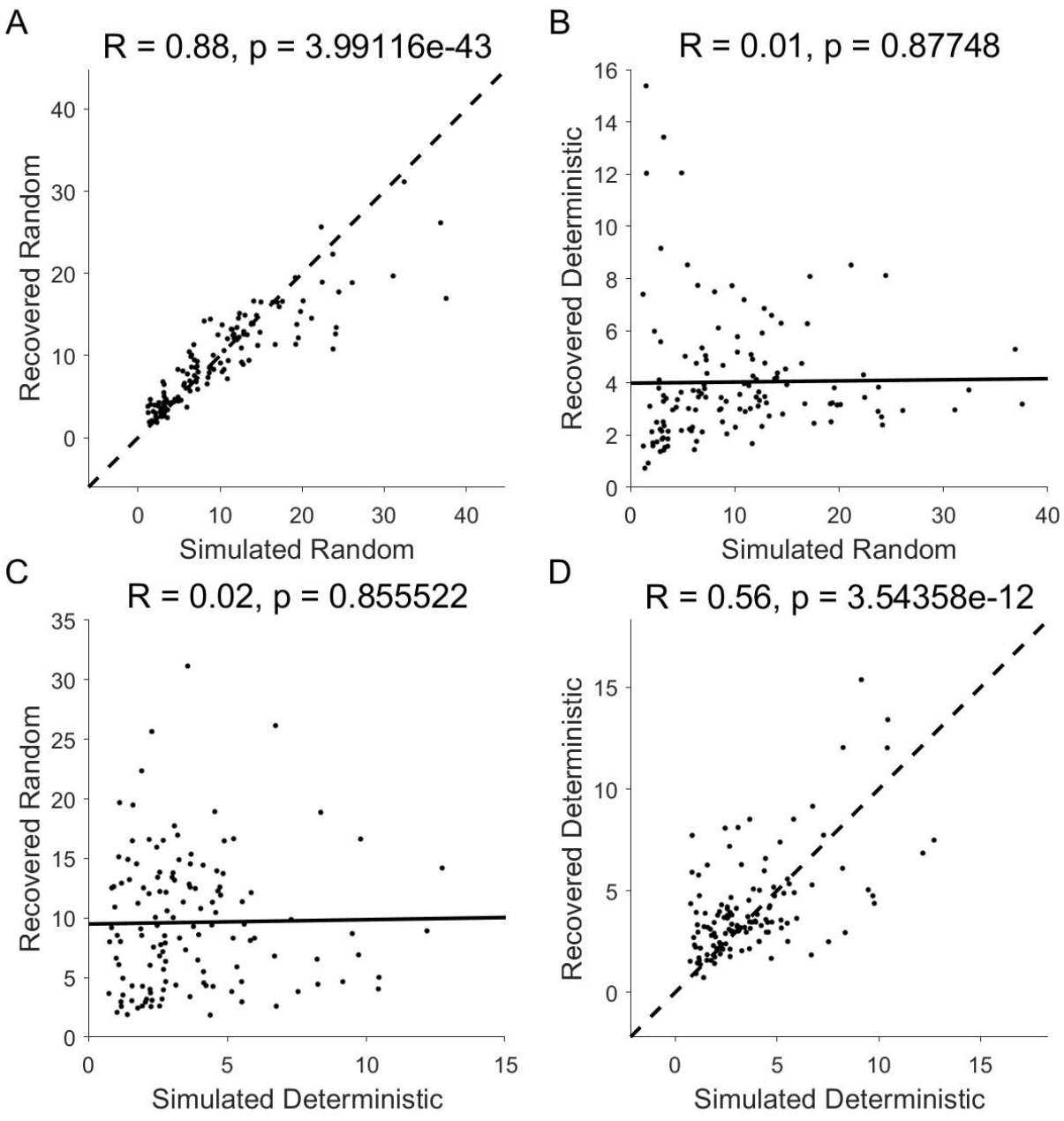
deterministic sources of noise with horizon length is genuine and not caused by a limitation of the fitting procedure. I tend to agree with their interpretation, but it would be very useful to provide additional empirical evidence in the main text that this joint increase is indeed genuine. The parameter recovery approach illustrated in Figure 4 should include panels of figures found in the Supplementary Materials which show that the fitting procedure is capable of correctly recovering arbitrary combinations of random and deterministic sources of noise. A simpler and more compact test, which the authors should perform and plot as a new panel of Figure 4, would be to plot the confusion matrix arising from the parameter recovery procedure (Figure 4 only shows what corresponds to the diagonal of the confusion matrix). It is indeed critically important that simulated ground-truth values of random noise do not correlate with best-fitting values of deterministic noise (and vice versa).

This would require to fit choice behavior using a non-hierarchical, subject-wise fitting approach, or to complexify the hierarchical fitting approach to estimate the covariance between random and deterministic sources of noise. Both control analyses would be valid, and I let the authors choose whichever approach they find most appropriate for their data. This would provide empirical evidence (already available to the authors since they have performed a parameter recovery analysis) that random and deterministic sources of noise can indeed be reliably separated, and therefore that the joint increase of the two forms of noise with horizon length is genuine. This type of control analysis have been performed in a recent study, in case this is helpful:

Lee JK, Rouault M, Wyart V (2023) Adaptive tuning of human learning and choice variability to unexpected uncertainty. Science Advances 9, add0501. doi:10.1126/sciadv.add0501

Response 1.3:

To show that the joint increase of random and deterministic sources of noise is not caused by a limitation of the fitting procedure, following the reviewer's advice, we calculated the correlation between ground-truth values of random noise, and best-fitting values of deterministic noise (and vice versa). Ground-truth values are shuffled best-fit parameters.



As expected, ground-truth random values do not correlate with recovered deterministic noises, showing that the increase of deterministic noise with horizon is genuine and not a by-product of increase of random noise, and vice versa.

Alternatively, in the supplemental materials, we also performed an analysis in which we simulated data where we have ground truth about whether random or deterministic noises change with the horizon condition. We fit our model to the simulated data and our model can faithfully detect the existence of horizon changes in both random and deterministic noises only when they exist, for simulated data. (See Supplemental Fig. S16)

Reviewer #2: In the manuscript entitled 'Separating random and deterministic sources of computational noise in explore-exploit decisions', Siyu Wang and Robert C. Wilson present an extension of the Horizon task by Wilson and colleagues (2014) to investigate whether random exploration in human decision-making is driven by stochastic processes in the brain or by some unobserved deterministic process. The task is extended so as to disentangle deterministic noise from random noise by presenting a situation where, unbeknownst to them, participants are presented with the exact same choice twice. This enables the authors to estimate a lower bound on the amount of variability that is deterministically driven by the stimulus and an upper bound on the amount of variability that is random. They found evidence that at least 14% of the variability in random exploration in their task can be accounted for by deterministic processing of the stimulus.

Comment 2.0: The topic of this research is very interesting, timely and of importance to the community. The maths behind the computational models and the analyses seem correct and are elegantly developed. But at the end of the reading I am left only partially satisfied. The starting important question is: where does the 'random' noise identified by Wilson et al. 2014 come from? But the article arrives to the conclusion that there's around 14% of explainable random noise on a task where there are important limitations to the protocol (see detailed comments below), and without really finding explanations of where do those 14% of deterministic noise come from, not why participants' choices tend to be repeated. This would require to look at when random (or deterministic) choices occur. For example, when the average rewards displayed by the bandits are close (this is taken into account in the model), when the uncertainty displayed on one of the two bandit arms is higher than for the other (risk seeking or risk averse behavior), when there is an attractor point (a higher reward) for the choice that is sub-optimal. The authors themselves admit this in the discussion: 'As a result, from both a conceptual and methodological perspective, it is possible that the remaining 86% of the decision noise that is not stimulus-driven noise, could be deterministic. ' I wish they'd develop more hypotheses on this.

Response 2.0: We sincerely thank the reviewer for carefully reading our manuscript and giving us detailed and constructive feedback.

Before we address the reviewer's comments point by point, we would like to make a general comment and clarify the scope and focus of our paper. While we understand the reviewer's "partial satisfaction" that we did not attempt to explain the exact source of the 14% of deterministic noise and the 86% of random noise, we would like to point out that this is almost intentional and is considered an advantage of our method. Instead of making hypotheses about the exact task-specific sources of deterministic and random noises, we took an alternative approach in this paper and developed a novel Bayesian approach which allowed us to separate deterministic noise from random noise without explicitly specifying or knowing their sources.

Although we have some ideas about where the 14% of deterministic noise might come from (e.g., motor sequence patterns of key presses during the forced-play trials), we leave this

pursuit of task-specific modeling the deterministic noise for future work, since it will not affect the main message of our paper.

Overall, I feel that some control analyses and ways to discard alternative interpretations are needed.

ANALYSES

Comment 2.1: I am intrigued by the negative information bonus A in the model-based analyses for Horizon=1 (Figure 2D). Is it significantly different from 0? If yes, does this mean that participants are even avoiding uncertainty (risk aversiveness) in that case? What would be the implications of this?

Response 2.1: Yes, the negative information bonus in horizon 1 is significantly different from 0. This phenomenon was previously reported and interpreted in the original Horizon Task paper (Wilson et al., 2014). One of the main contributions of the Wilson et al., 2014 paper, was to use the horizon manipulation to separate "risk preference" (negative information bonus in Horizon = 1 which reflects uncertainty aversion) from "directed exploration" (changes in information bonus between Horizon 1 and 6 which reflects information-driven exploration).

Comment 2.2: I do not fully understand how the plotted values for the 'pure random noise prediction' (i.e., 'random noise only' in the figures) and 'deterministic noise only' in Fig. 3, Suppl. Fig. S2 and S3 were computed. If these correspond to theoretical values for the choice inconsistency for the purely deterministic and purely random noise cases, as formalized pages 11 and 12, then I don't understand why these values have standard deviations and vary so much between figures: in the [2 2] condition of Figure 3, the plotted mean of random noise only are <0.2 for Horizon 1 and <0.3 for Horizon 6, while in the [2 2] condition of Suppl. Fig. S3, they are >0.2 for Horizon 1 and >0.3 for Horizon 6. I expect that the simulated data in the [2 2] condition of Suppl. Fig. S3 are significantly different from the 'pure random noise prediction' (i.e., 'random noise only') in the [2 2] condition of Figure 3. Isn't it a problem and shouldn't the authors solve it here?

Response 2.2: I am happy to clarify this:

1. How was "pure random noise prediction" computed?

   "pure random noise" refers to the assumption that participants treat the repeated games independently (there is zero deterministic noise). Under this assumption, considering a single game, if we know the probability that a participant chooses option A, then the probability of making consistent/same choices in repeated games would be:

$$P(consistent) = p(choose\ A\ twice\ or\ choose\ B\ twice)$$
$$= p(choose\ A)^2 + p(choose\ B)^2$$
$$= p(choose\ A)^2 + (1 - p(choose\ A))^2$$

   If we define option A to be the option that has a lower mean reward from the first four forced-choice trials, then the above formula becomes:

$$P(consistent) = p(low\ mean)^2 + (1 - p(low\ mean))^2$$

The "pure random noise prediction" refers to the theoretical prediction of p(consistency) using the above formula. When considering all games, we predict p(consistent) based on the empirical percentage of choosing the low mean option, p(low mean), based on each participant's behavior.

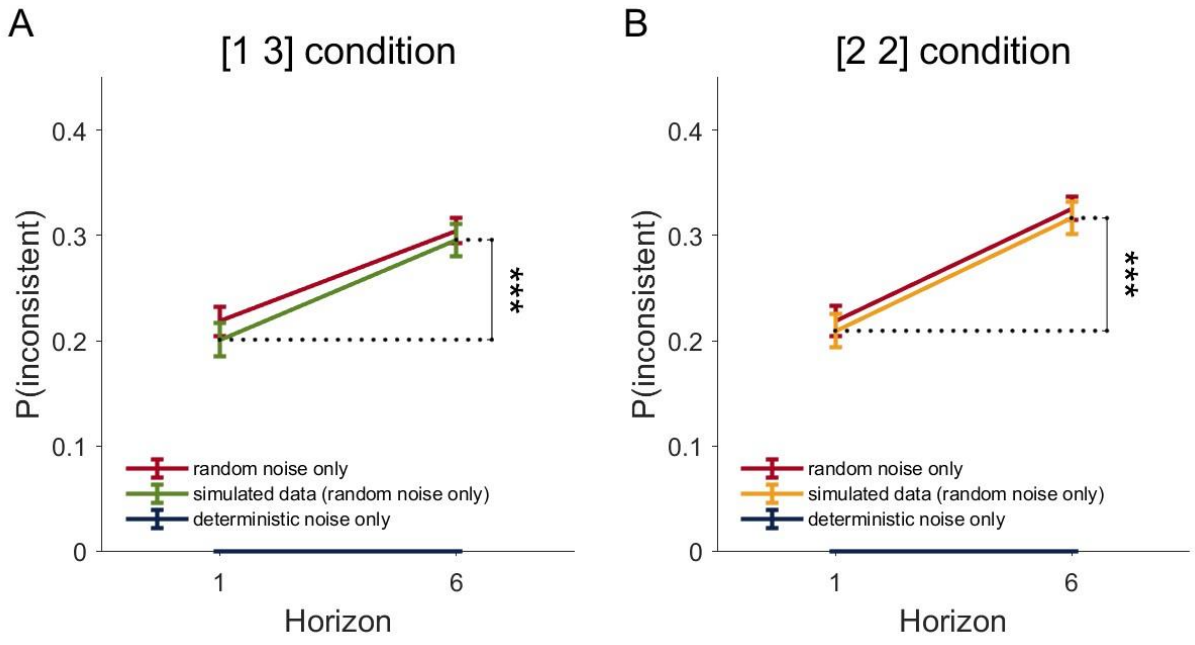
2. Why do these theoretical values have error bars and vary between figures?

   Since we use p(low mean) from each participant's behavior, we calculate the theoretical p(consistent) for each participant, and the error bar is the standard deviation of the predicted p(consistent) across participants.

   The reason Figure 3 and Figure S3 have very different y values is because Figure 3 is from data (where both deterministic and random noises presumably exist), and Figure S3 is from our simulation from a model that has only random noise and zero deterministic noise.

3. Explanation of Supplementary Fig. S3:

   The point of Figure S3 is to empirically validate our theoretical calculation. This is not a posterior check analysis, we did not simulate from the best-fit parameters. Instead, we simulated data from a model that assumes 0 deterministic noise and all random noise. Our claim that deterministic noise exists comes from the fact that the data line significantly deviates from the theoretical pure random noise line in Figure 3. As a control, if we simulate data from a model that only has random noise, we expect the data line to be not statistically different from the theoretical predicted line, and that is exactly what we see in Fig S3 (pasted below).



Comment 2.3: I think the model validation analyses in supplementary data are very useful and well-performed. Nevertheless, shouldn't the spatial bias term be kept in all model versions to make them comparable? Could the authors quantitatively show how much the spatial bias term contributes to explaining participants' behavior? In Suppl. Fig. S7, it seems difficult to recover the spatial bias parameter with the parameter recovery method. Why is that so?

Response 2.3: Spatial bias was in fact kept in all model versions. We included spatial bias as it was included in the standard Horizon Task behavioral model proposed in Wilson et al., 2014.

In practice, participants do not have spatial biases significantly different from 0. Since spatial bias term was small (around 0), it was relatively more difficult to recover. In parameter recovery analysis, we simulated data with the best-fit parameters from the data, so the simulated bias terms were small (close to 0) to begin with, making the recovery difficult.

Comment 2.4: Moreover, it would be useful to give the reader a quick grasp of the summarized results by showing a model recovery matrix (as in Wilson & Collins 2019) with all nested versions of the full model (those in Table S1). Does the full model win when the simulations are generated by the full model? Does a reduced model without random noise win when the simulations are generated by the very same model? Conversely, does a reduced model without deterministic noise win when the simulations are generated by the very same model?

Response 2.4: Model recovery analysis as in Wilson & Collins (2019) would require an estimate of the data likelihood for each model. While this is doable for most cognitive models, we want to point out that this is not feasible for our model.

In a traditional cognitive model, usually a choice probability p is computed based on input (e.g., reward history), and choice is sampled based on this probability. However, our model is not probabilistic, and is binary instead:

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

Instead of using a softmax to model randomness in behavior (which only has random noise), in order to separate random and deterministic noises, we had to sample n_{det} and n_{ran} trial-by-trial using a MCMC procedure, and for each trial, choice is 1 if $\Delta Q > 0$, and choice is 0 if $\Delta Q < 0$. As a result, there is no easy way to get a data likelihood estimate and a model recovery analysis in the traditional way is not possible.

Instead, to show that our model is capable of correctly measuring both random and deterministic noises and their horizon-dependent changes, we performed the supplemental analysis shown in Supplementary Fig. S16 (also attached below). In this analysis, we simulated data from the 6 variants of the reduced models, and fit the full model to data generated from each of the reduced models.

Model	Deterministic noise	Random noise
$\sigma^{ran}_{horizon}, \sigma^{det}_{horizon}$	Horizon dependent	Horizon dependent
$\sigma^{ran}_{horizon}, \sigma^{det}$	Fixed	Horizon dependent
$\sigma^{ran}, \sigma^{det}_{horizon}$	Horizon dependent	Fixed
$\sigma^{ran}, \sigma^{det}$	Fixed	Fixed
$\sigma^{ran}_{horizon}$	Horizon dependent	None
$\sigma^{det}_{horizon}$	None	Horizon dependent

Table S1: Variants of the model.

The idea is that the full model should only detect the existence and horizon-dependent changes of random/deterministic noises when the data-generating model have them. For example, when we simulate from a model that deterministic noise is fixed and random noise increases with Horizon, and fit our full model to the simulated data, we expect to see that the recovered deterministic noise does not change with horizon, but random noise does. And that's exactly what we see (panels E-H).

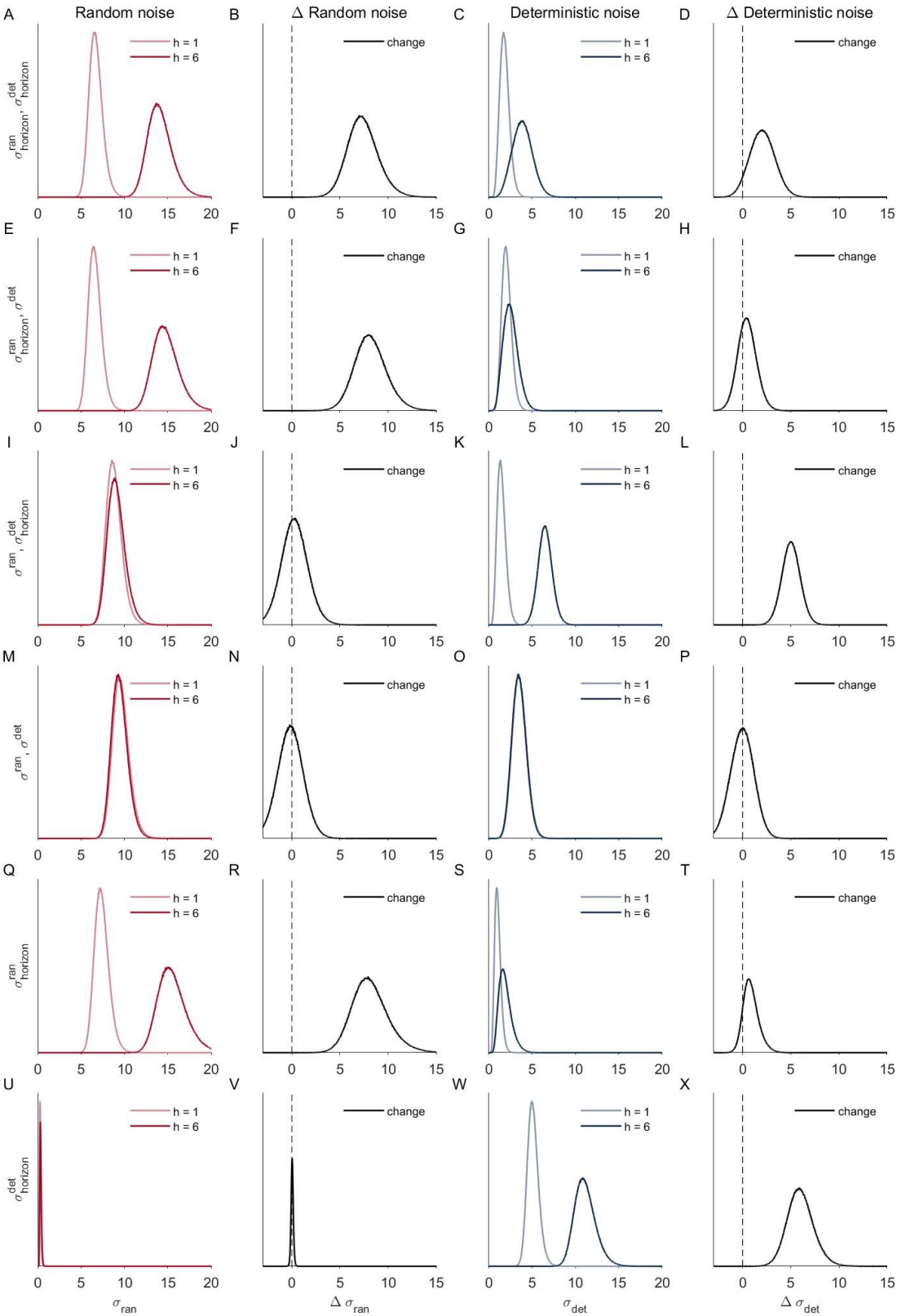


Figure S16. Our model qualitatively captures whether deterministic and random noise are present or not and whether either types of noise is dependent on horizon. A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is

horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

Comment 2.5: Do I understand correctly that in the full model as well as in the reduced one, the random noise (sigma ran) and the deterministic noise (sigma det) come into play only during the first choices that are repeated during two games, and not during other first choices nor during 2nd, 3rd, etc. choices? Or instead are these two noise terms contributing to all decisions? Could the authors make this clearer in the manuscript, please?

Response 2.5: Yes. Since the first free choice (the first 4 choices are forced, this would be the 5th choice in both H = 1 and H = 6 games) is the only choice that is comparable between Horizon = 1 and Horizon = 6 conditions, all of our analyses were performed on the first free choice. While these two noise terms should contribute to all decisions, our paper only modeled the first free choice.

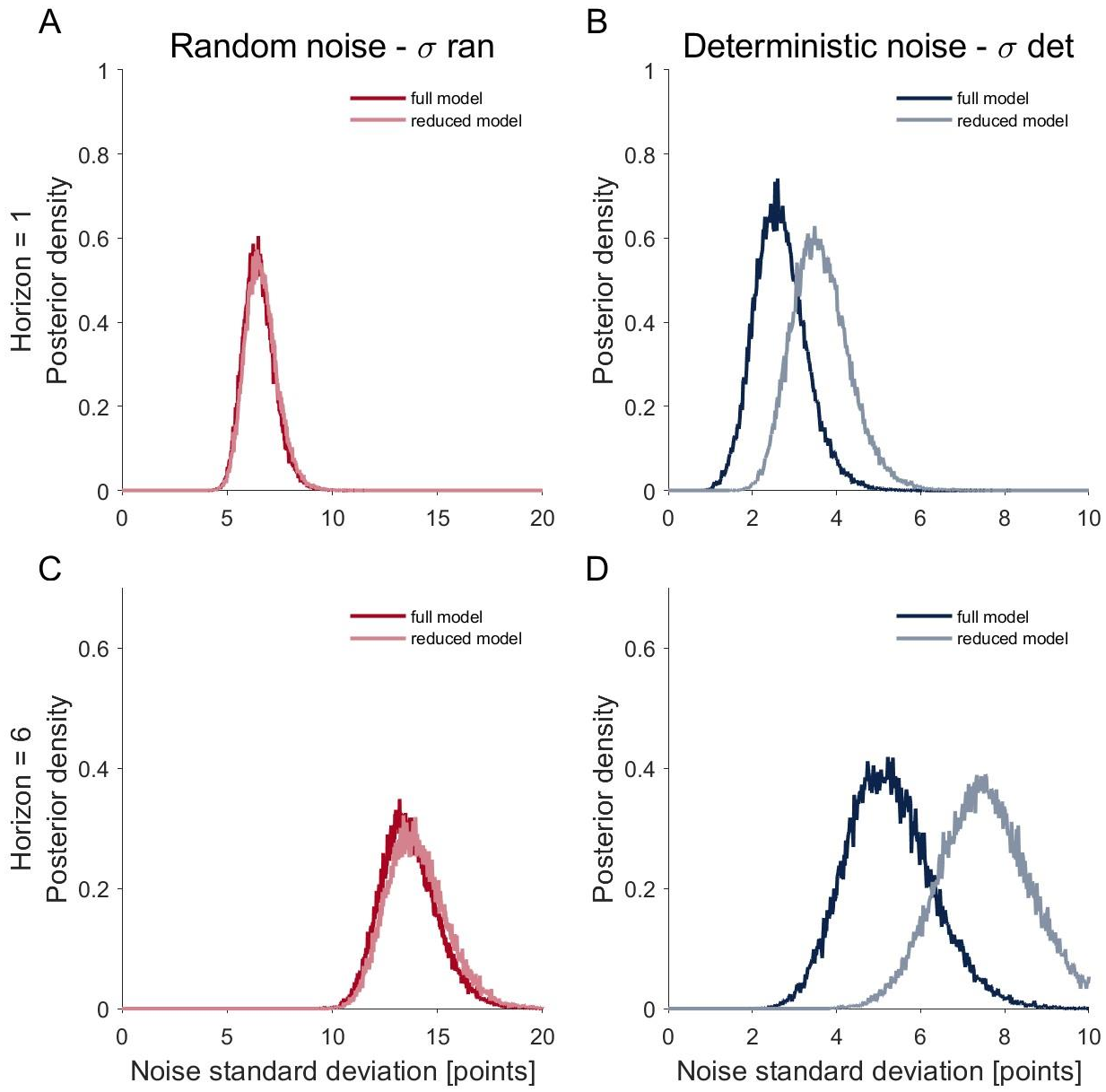
In the results section, we had this sentence when describing the task
"*By contrasting behavior between short and long horizon conditions on the very first free-choice trial, when all else is equal, the Horizon Task allows us to quantify how behavior changes, when it is more valuable to explore*."

When describing the methods, we modified the following paragraph, hopefully it is clearer now:
"…we focus on just the first free-choice trial in each game, where the only thing that differs between the horizon conditions is the number of choices that participants will make in the future. Subsequent choices in Horizon 6 games were not analyzed."

Comment 2.6: Could the authors add a few sentences in the supplementary information to clarify that if the random noise increases in the reduced model (the one without information bonus), it leads to less choice consistency between repeated games, and that conversely if the deterministic noise increases, it leads to more choice consistency?

Does the deterministic noise in the reduced model capture and replace the effect that the information bonus produces in the full model?

Response 2.6: In general, increasing random noise leads to higher p(low mean) and lower p(consistent) between repeated games, and increasing deterministic noise leads to higher p(low mean) and higher p(consistent).

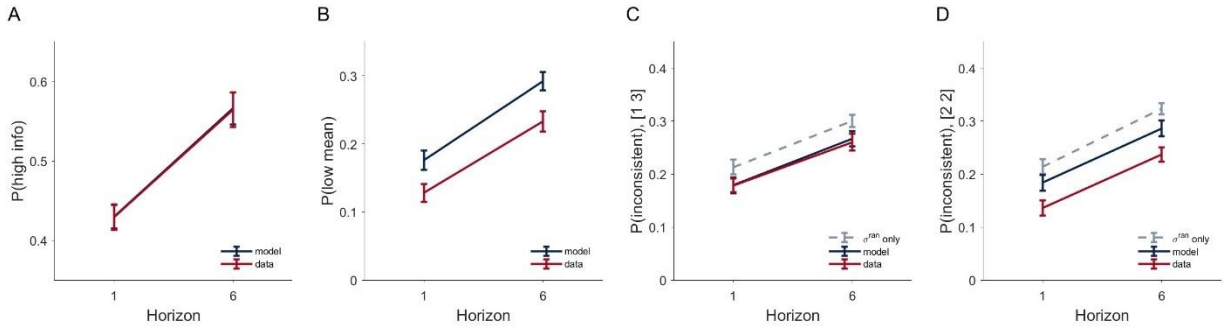


When removing information bonus from the full model, we created a case where we know the reduced model is missing an important source of deterministic noise (explained by information bonus). As a result, we expect to see an increase in deterministic noise and that's what we saw (Supplemental Fig. S4, also attached above). The increased deterministic noises show that in the reduced model, there are higher unexplained variances (in the reduced model) that are predictable from the stimulus, compared to the full model. And the increased deterministic noise reflects the missing of the "information bonus" term.

Comment 2.7: The 'posterior predictive check' analysis is very nice and still rarely performed in the literature. Nevertheless, why are the model and data so different for p(low mean) (Figure 7B)? Is the difference significant? How can this be explained and interpreted?

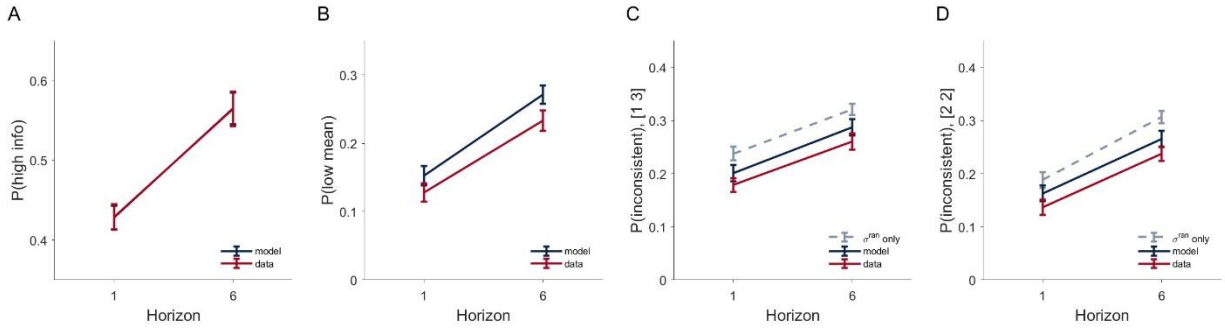
Response 2.7: We thank the reviewer for asking us to examine deeper the quantitative mismatch between model and data in posterior checks. For the overestimation of p(low mean) and p(consistent) in [2 2] condition in posterior checks, we have identified two factors that led to the mismatch:

0. For comparison, we attach here the posterior check figure from the initial submission.



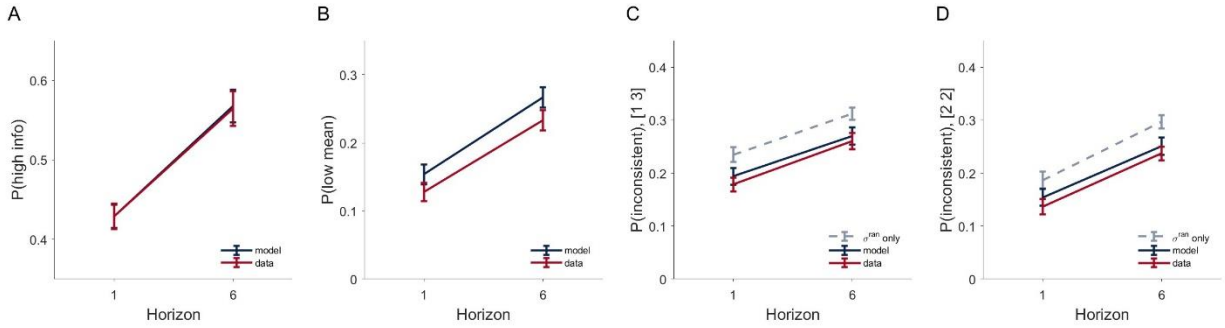
1. In our model, we assumed that the variances of deterministic and random noises are from a constant distribution for both the [1 3] and [2 2] information conditions. Since our model overestimates p(low mean), it suggests that we are overestimating noise in [2 2] condition. It is possible that people have higher noises in [1 3] condition compared to [2 2] condition, by assuming the [1 3] and [2 2] condition share the same noise distribution, it will lead us to overestimate noise in [2 2] condition.

   To fix this, we fit a variant of our model in which we separately estimate the variances of random and deterministic noises in [1 3] and [2 2] conditions. Indeed, we observe higher overall noise level for [1 3] condition. When performing posterior checks with this model, the mismatch between data and model becomes smaller in the [2 2] condition.

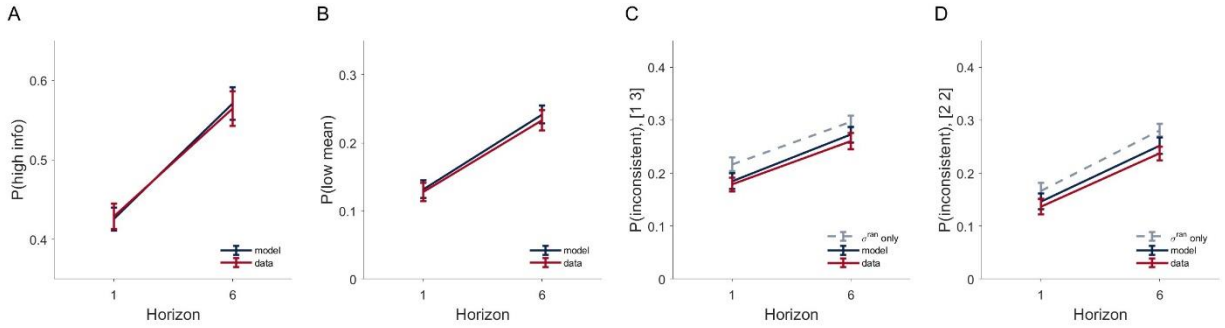


2. In our model, the subject-level noises are assumed to follow a gamma distribution (to ensure positiveness), the posteriors are right skewed (maximal likelihood estimation or mode is smaller than the mean), in the original analysis, we simulated data from each participant using the mean of the subject-level posterior for both deterministic and random noises, however, because of the skewness, simulating data with the "mean" is nosier than simulating with the true distribution. Simulating from the true distribution requires taking expected value over all possible noise values, for ease of implementation, we simulated data by taking random samples from the posterior distribution (instead of using the mean). The simulation was repeated 50 times and then averaged, below is what we get. Indeed, the

mismatch between data and model further decreased. (we adopted this method in the main text, as this is the closest to the data generation process in the model)



If we use the mode, which corresponds to the maximal likelihood estimation estimate, we can get the model simulation to closely match the data:



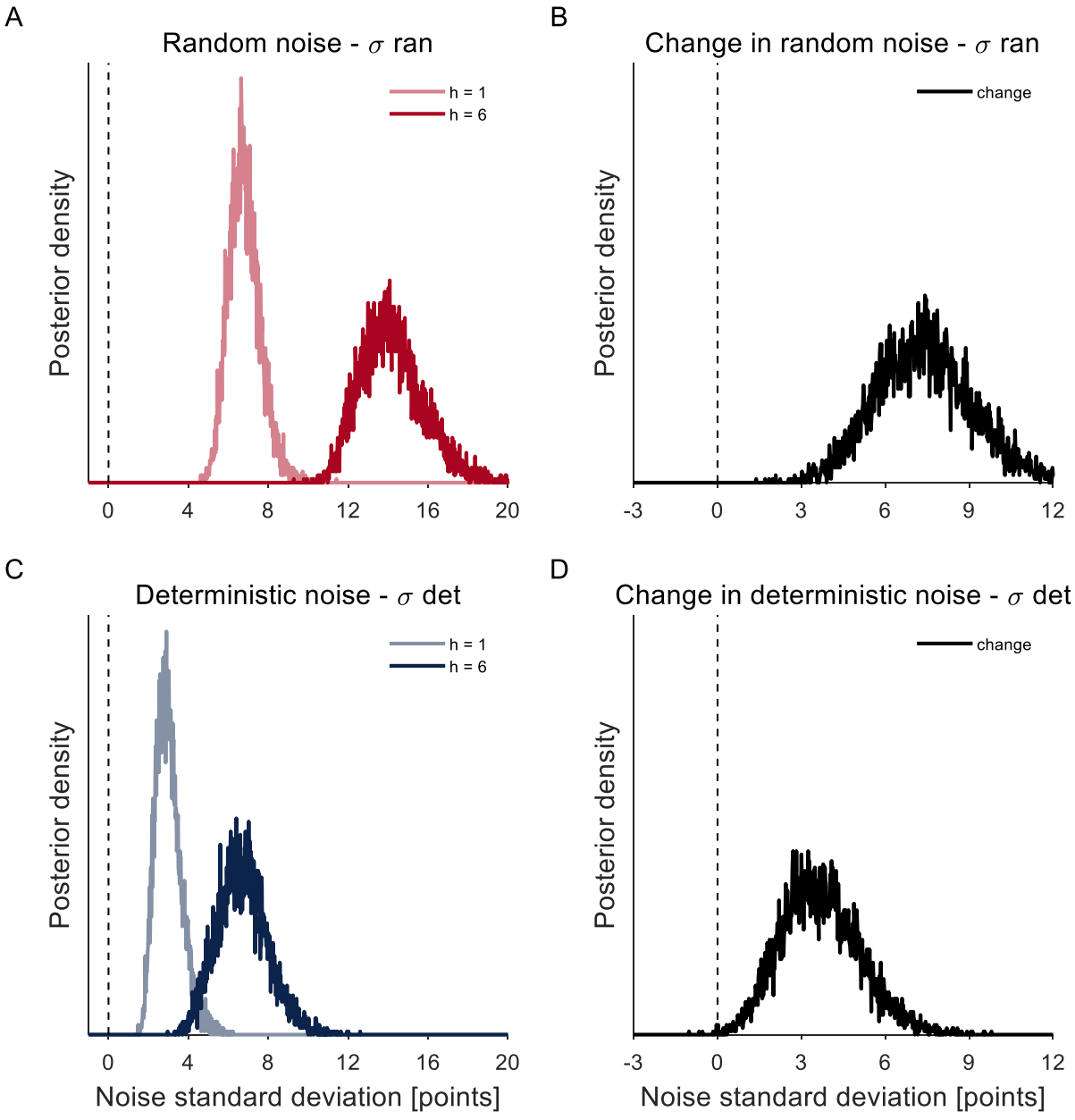
Comment 2.8: I'll go into more details about risk averse and risk seeking behavior. In the model, discretizing information amounts is an approximation that may have important consequences in terms of explanatory power and interpretation of the results. In particular, setting a value of 0 for information in a situation [2 2] is bit unsatisfying: even if the four elements have been drawn from distributions with the same variance, we may have two values close together on the left and two far apart on the right. For example, if you have two identical values on one side and two different values on the other, it's very clear that there's information to be found on one side but not on the other. This is not taken into account in the model and falls under the heading of 'random' noise, which sounds somehow a bit absurd. Changing this in the model wouldn't require much effort. For example, replacing -1, 0, 1 by the variance differences. The variance of a single element is 0, so in [1 3] and [3 1] we go from -1 or 1 to the value of the variance of the 3 elements, with a plus or minus sign in front. And in the case [2 2] we go from 0 to a difference in variance that is probably small but potentially non-zero. This wouldn't change the model much, but it would directly change the interpretation (what if the 14% came from there?). Otherwise, it gives me the impression that the authors are a bit over-interpreting their results of a model that may not be ideal, and that the value of 14% doesn't mean much.

Response 2.8: We understand the reviewer's comment.

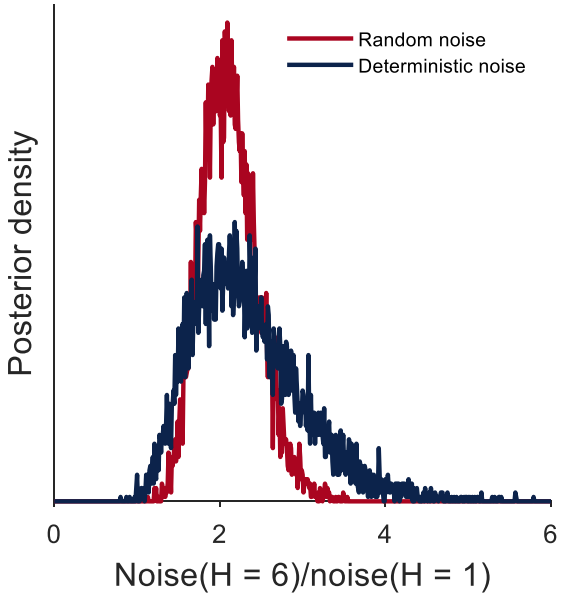
1. We acknowledge that our model is simplified in the sense that we discretize information amounts and fail to account for information difference in [2 2] condition. We chose to present this simplified model in the main paper, as this model was the standard model used for the Horizon task, and was the model proposed in the original Horizon task paper (Wilson et al., 2014).

2. We want to point out that the information difference in [2 2] condition which we fail to capture in our model, falls in the category of “deterministic noise” as opposed to “random noise”, as the amount of information difference in [2 2] condition for each game would be identical between repeated games.

3. Per reviewer’s request, we have implemented the requested version of the model, in which we explicitly consider uncertainty in both [1 3] and [2 2] options by defining dI to be the variance differences between bandits. Our main findings did not change when using this model. Specifically, we see that both deterministic and random noises increase with horizon.



We also see that random and deterministic noises increase with horizon at similar rates using this alternative model:



In this model (model VAR), actually 18.1% of variances is explained by deterministic noise. Note that this is higher than the 14% for the original model with dI = -1, 0 or 1. This shows that the 14% of deterministic noise can not be explained simply by the variance differences in [2 2] condition. The fact that deterministic noise is higher in model VAR might suggest that model VAR is a worse fit to behavior compared to the original model. While the difference in variances theoretically accounts better for information differences in [2 2] conditions, it is possible that it numerically does not fit the data as well.

If we assume a uniform prior over all possible reward values and use the example trials to compute the variance of the posterior for each bandit. The number of plays (1, 2, or 3) has a much bigger influence than the specific reward outcomes. The variances of the posteriors are actually tightly clustered around 3 numbers (for 1 sample, 2 samples and 3 samples, respectively), regardless of which 1, 2 or 3 rewards were shown. This suggests that assuming dI = -1, 0 and 1 did not miss much in capturing the information difference between options in this particular task.

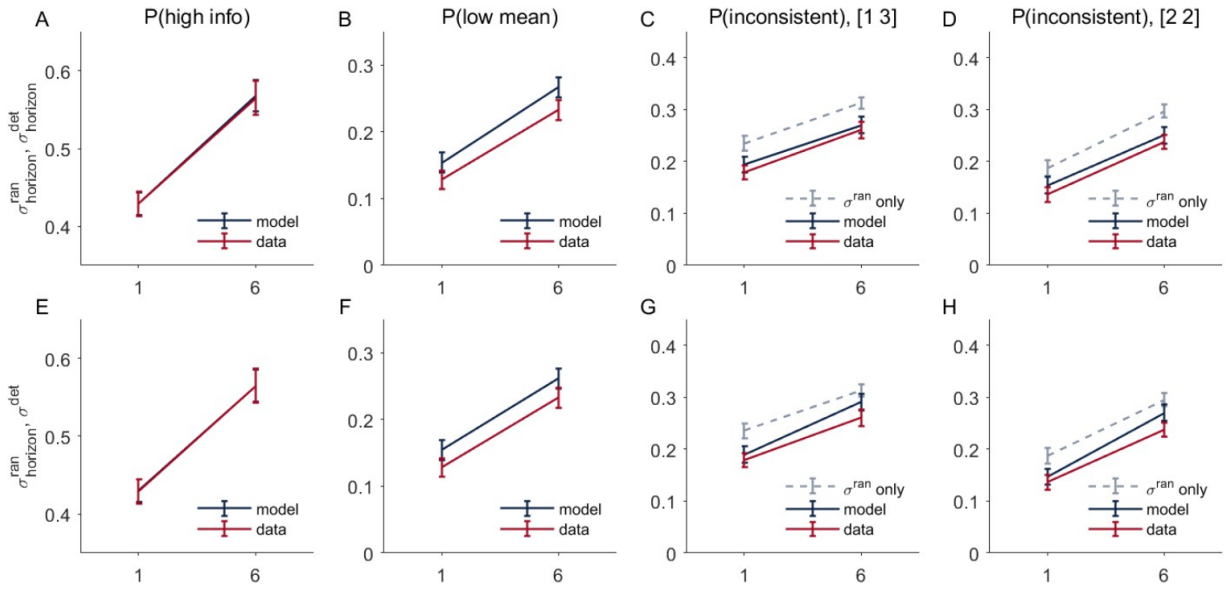
Comment 2.9: There is a contradiction between the sentence 'these reduced models fail to capture all qualitative patterns (Supplementary Figure S13)' in the main article, and the sentence 'As shown in Figure S13, only one of these alternative models, where random noise is horizon dependent but deterministic noise is not, can capture the full qualitative pattern of behavior.' in Supplementary Information. The main original should clarify that one of the alternative models does capture qualitative patterns. Moreover, the authors should clarify which quantitative measure they used to decide 'not as good' in sentence 'However, the quantitative fit to the data is not as good (Figure S13)' in Supplementary Information. To me it seems in Suppl. Fig. S13 that the difference between the two models (Fig. S13 A-D vs. Fig. S13 E-

H) and the data is not significant. Again, I think that a model recovery analysis would be needed here.

Response 2.9: We apologize for the ambiguity in our language. To clarify, all qualitative patterns include:

1. p(high info) and p(low mean) increase with horizon (Wilson et al., 2014 findings)
2. p(consistent) increase with Horizon
3. p(consistent) is statistically different from the theoretical predicted value of p(low mean)

We want to clarify that only the full model captures all of these patterns. The best alternative model (Fig. S17 E-H) is quite close, as it captures patterns 1 and 2, but does not fully capture pattern 3 in Horizon 6.



For the full model, p(consistent) differs from the theoretical random noise prediction (gray dotted line) statistically, whereas in the best alternative model (row 2 here), p(consistent) is only different from the random noise prediction in about half of the simulations in Horizon 6 (out of 50 simulations), due to the inability to increase deterministic noise with Horizon in this model.

Comment 2.10: In the Discussion section, the demonstrations of how a change in reward processing could affect random and deterministic noise should be acknowledged as similar to the demonstration of how a change in reward processing could affect random noise in Cinotti et al. 2019 Scientific Reports, where it is written that when 'all Q-values are downscaled in the same proportion as the reward [and] When these values are plugged into the softmax process, the result is exactly equivalent to a decrease of the inverse temperature, again in the same proportion.'

Response 2.10: We have added the reference to the discussion section.

Comment 2.11: What if the participants had sometimes a good memory of having already been

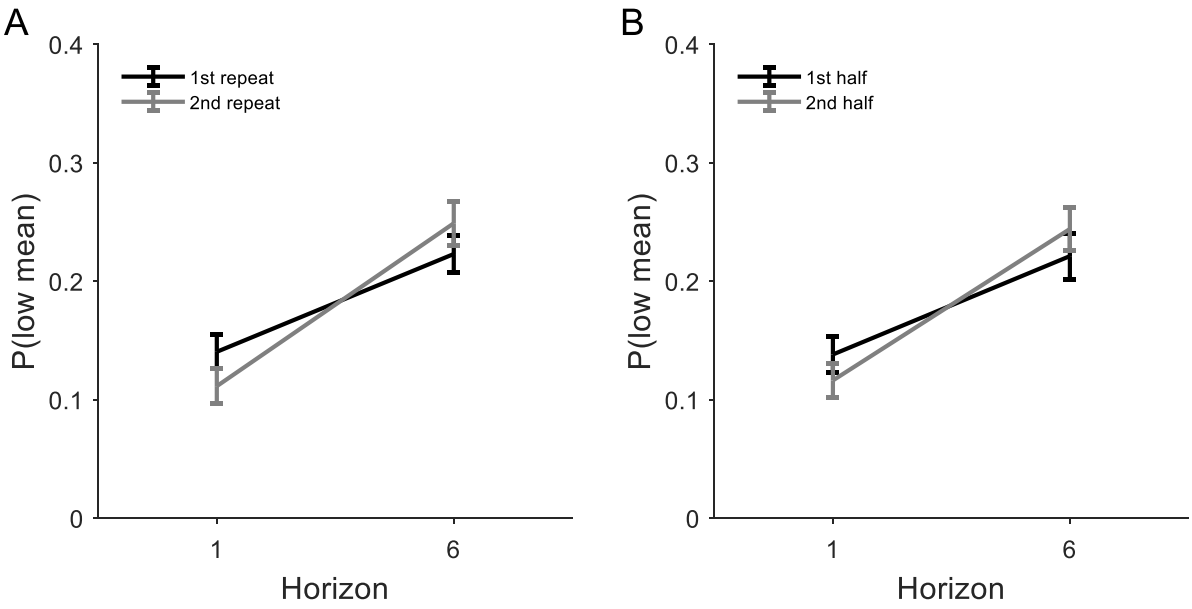
confronted with the same game + the bandit they had previously chosen, and deterministically decide to pick the other bandit so at to see which payout they obtain in this case? This would be deterministic exploration policy at the game level rather than at the bandit level, in contrast to the authors' interpretation as non-stimulus-driven random noise in explore-exploit decisions. The authors argue that two identical games in their task are 'separated by several minutes in time so as to avoid detection'. But how could they unsure that the repetition has never been detected? What would be the interpretation of their results if let's say at least a proportion of game repetitions had been detected by the participants? It seems to me that a way to address this problem would be to redo the task and ask 2 questions after each game's first free choice: have you already encountered the same game before? If yes, which bandit had you chosen the previous time? On the one hand, this would prompt people to know that there are game repetitions, which would increase their vigilance towards this feature and would increase their detection probability, on the other hand, this would enable to separate repeated games for which participants' accurately remembered their previous choice from those where they failed to remember. Another solution, so as not to bias participants' responses during the task would be to ask them to fill a questionnaire after, where they are asked whether they think they have encountered twice the same situation. This would be less ideal than the proposed variant of the task, but at least give an idea whether this was a problem or not in the present task. Have the authors asked the participants such questions during a post-task questionnaire?

Response 2.11: Unfortunately, we did not formally ask whether participants detected a repeat. However, we did interview participants after the experiment and asked them about what strategies they used during the task, and what they thought the experiment was about, and none of the participants mentioned noticing a repeat.

The reviewer is correct that our model does not consider game-level deterministic strategy.

1. Game-level strategies like memorizing repeated games will show up as random noise in our current model, just like other deterministic processes that are not a deterministic function of the current game stimulus.
   Because of this, we made clear that our method provides a lower bound of deterministic noise.
2. We checked p(low mean) separated by the 1st repeat versus the 2nd repeat (panel A below), the idea is that if people can detect repeats and have a tendency to explore the alternative in the second repeat, they should on average pick the low mean option more in the 2nd repeat. But that's only true in Horizon 6, in horizon 1 they show the opposite.

   As a control, we also did the same analysis but separated trials by the 1st half of the experiment versus the 2nd half of the experiment, regardless of whether it's the 1st repeat or 2nd repeat (panel B). It looks very similar to Panel A. It is likely that the difference we see are due to early vs late phase of the experiment, and likely not due to repeats.



3. Conceptually, our model could naturally be extended to account for game-level deterministic strategy. Instead of treating "reward history within a game" as stimulus, the whole history of rewards across games can be treated as the stimulus. But this is beyond the scope of the current paper.

We have added this limitation to the discussion section.

LITERATURE

Comment 2.12: In the abstract, I suggest to replace 'recent work suggests that variability can actually be adaptive' by 'a long body of machine learning work suggests that variability can actually be adaptive'.

In the introduction, to be fair with the existing computational literature on adaptive decision noise, after the sentence 'It has recently been shown that humans appear to use random exploration and can increase decision noise when it is more beneficial to explore (Findling et al., 2019, Gershman, 2018, Wilson et al., 2014)', I suggest the authors add the following: , as has also been suggested in computational models of animal behavior (Doya 2002 Neural Networks; Khamassi et al., 2013 Progress in Brain Research).

Response 2.12: We have updated the introduction and abstract accordingly.

Comment 2.13: Typos
Page 8, a logistic distributions.
Page 12, in the both the.
Page 19, a fixed random motion stimuli -> stimulus.
Page 25, in the reference by Beck et al. 2012, there is a duplication of bibliographical

information. Same thing for Findling et al., Tomov et al., Musall et al., Hogeveen et al., Ebitz et al., and Costa et al.
Page 8 of Suppl. Info. (Section 2.3) to recovery parameters -> to recover.
Response 2.13: We have corrected these typos, thanks for pointing these out.

Separating random and deterministic sources of computational noise in explore-exploit decisions

Siyu Wang[1] and Robert C. Wilson[1,2,3]

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

[2]Neuroscience and Physiological Sciences Graduate Interdisciplinary Program, University of Arizona, Tucson AZ, USA

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

April 22, 2025

Abstract

Human decision making is inherently variable. While this variability is often seen as a sign of suboptimal behavior, both theoretical work in machine learning and empirical human studies suggest 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 can encourage us to explore options we might otherwise ignore. In line with this idea, several studies have found evidence that people increase their behavioral variability when it is valuable to explore. A key question, however, is whether this variability in so-called '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. By modeling behavior in this task, we were able to estimate a lower bound on the amount of variability that is deterministically driven by the stimulus and an upper bound on the amount of variability that is random. Using this approach, we find evidence that at least 14% of the variability in random exploration in our studied task can be accounted for by deterministic processing of the stimulus. Conversely, this suggests that up to 86% of the variability is truly 'random', although it is still possible that this variability is driven by deterministic factors not related to the stimulus. Finally, our results suggest that both deterministic and random sources of variability change proportionally to each other as the value of exploration increases, suggesting that a common noise gating mechanism may be at play in random exploration.

Author Summary

Human decisions often seem random. Even simple decisions like what food to order at a restaurant can be difficult to predict ahead of time. This randomness in our decisions can be beneficial, effectively allowing us to explore new options. One outstanding question is where the randomness in our decisions comes from. Sometimes, our seemingly random decisions are driven by predictable external factors, like what the guest at the next table ordered could influence what we order. Other times, our decisions are not driven by external factors but are instead made by random thoughts within our brain. In this work, we developed a computational method that quantifies the extent to which the apparent randomness in our decisions can be explained by deterministic sources of variability in the external stimuli, or random variability unexplained by the stimuli. We found evidence that randomness in exploratory decisions can be explained by both random (up to 86%) and deterministic (more than 14%) sources of variability. Moreover, our results suggest that both sources of variability are adaptive, which enables humans to explore more when it is more beneficial to explore. The joint adaptation of random and deterministic noises also suggests a common noise-gating mechanism for exploration.

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 computational complexity, there is considerable interest in how humans and animals solve the explore-exploit dilemma in practice (Mehlhorn et al., 2015, Schulz and Gershman, 2019, Wilson et al., 2021).

One particularly effective strategy for solving the explore-exploit dilemma is choice randomization (Bridle, 1990, Thompson, 1933, Watkins, 1989), also known as random exploration. In this strategy, high value 'exploit' options are not always chosen and exploratory choices are sometimes made by chance. From a modeling perspective, random exploration works by adding 'decision noise' to the value of the options such that sub-optimal exploratory options can sometimes have a higher total score (i.e., value + noise) than the exploit option and get chosen. 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 decision noise when it is more beneficial to explore (Gershman, 2018, Wilson et al., 2014), as has also been suggested in computational models of animal behavior (Doya, 2002, Khamassi et al., 2013). In one of these tasks, known as the Horizon Task (Wilson et al., 2014), 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 will be more likely to explore the new restaurant. Using such a horizon manipulation it has been shown 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 research, however, is that it is difficult to tell whether what we have called 'decision noise' actually reflect a noise process. From a modeling perspective, decision noise as

defined in previous research essentially quantifies the extent to which behavior cannot be explained by a computational model. A missing deterministic component from the model could give rise to variability in behavior that might appear to be random noise. 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 the 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. To dissociate a missing deterministic component from a true random process, choice consistency between repeated decisions can be utilized to decompose behavioral variability into predictable deterministic components and unpredictable random components (Findling et al., 2019, Findling and Wyart, 2021, Wyart, 2018, Wyart and Koechlin, 2016).

In this paper, we investigate the extent to which the apparent randomness in random exploration can be explained by deterministic processing of the stimulus (which we refer to as 'deterministic noise') vs other processes, including deterministic processing that is unrelated to the stimuli as well as truly stochastic processes (which we refer to as 'random noise'). To distinguish between these two types of noise, we modify the Horizon Task (Wilson et al., 2014) to have people face the exact same explore-exploit choice twice. If the decision is a purely deterministic function of the stimulus (i.e., decision noise is purely deterministic noise), then people's choices should be identical for both decisions, since the stimulus is the same both times. Conversely, if the decision is a purely random function of the stimulus (i.e., decision noise is purely random noise), then people's choices will be different 50% of the time, since the random noise is different each time. In between these two extremes of purely deterministic and purely random drivers of behavioral variability, the extent to which people's decisions are consistent between the two decisions can be used to estimate the amount of deterministic and random noise.

In the following, we analyze behavior on the repeated decisions version of the Horizon Task in both a model-free and model-based manner. Our model-free analysis estimates the extent to which people's behavior is consistent across repeated versions of the same decision. By measuring how this choice consistency changes as a function of horizon, this model-free analysis offers qualitative insight into the extent to which behavioral variability is driven by deterministic vs random noise. Our model-based analysis uses a computational model of the explore-exploit decision in the Horizon Task that incorporates both noise processes. By fitting this model to the behavioral data, this model-based analysis allows us to quantify the relative size of the two sources of noise and how they change in the service of exploration.

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' in explore-exploit decisions (Figure 1). In this task, participants make a series of choices between two slot machines, or 'one-armed bandits', that pay out probabilistic rewards. They are asked to choose between the two bandits to maximize the total rewards. One bandit always has a higher mean payout than the other. Participants need to try each bandit a few times to learn about the distribution of payout from that bandit. 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.

The task is organized in games (Figure 1A). The mean payout of the two bandits are held fixed within a game and reset between games. Each game consists of either 5 or 10 trials. The first four trials of each game are 'forced-choice' trials. In the first four trials, participants are instructed about which bandit to choose, this allows us to manipulate what information from both bandits participants receive before they make their first free choice between the two bandits. From the 5th trial, participants make free choices between the two bandits. Participants have either 1 or 6 free choices to make.

The Horizon Task has two key features that together allow it to quantify explore-exploit behavior. The first of these features is the time horizon — the number of decisions participants will make in the future. By changing this horizon from short (1 free-choice trial) to long (6 free-choice trials), the Horizon Task allows us to control the relative value of exploration and exploitation. Just like the restaurant example in the introduction, when the horizon is short, participants should be more likely to exploit the option they believe to be best, because this leads to the highest payoff in the short term. Conversely, when the horizon is long, participants should be more likely to explore at first, because this allows them to gather information to make better choices later on. By contrasting behavior between short and long horizon conditions *on the very first free-choice trial*, when all else is equal, the Horizon Task allows us to quantify how behavior changes, when it is more valuable to explore.

The second key feature of the Horizon Task are the 4 forced-choice trials at the start of each game that allow us to control exactly what participants know about the two bandits before they make their choice. In these forced-choice trials, participants are instructed which of the bandits to play allowing us to control how much information they have about each of the options. The forced-choice trials are used to set up one

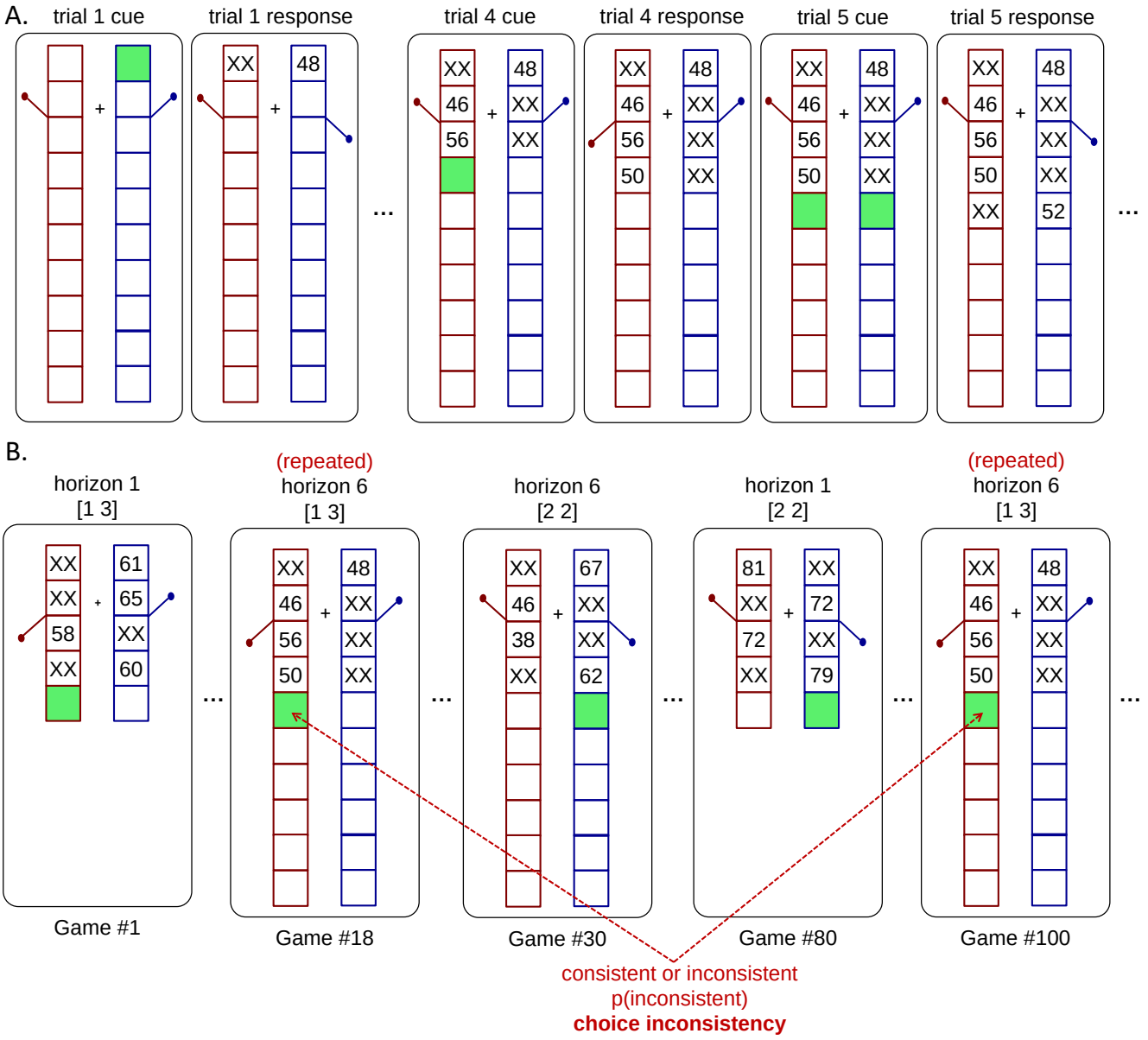


Figure 1: Schematic of the experiment. (A) Dynamics of an example horizon 6 game. Here the first four trials are forced trials in which participants are instructed which option to play. After the forced trials, participants are free to choose between the two options for the remainder of the game. (B) Example repeated games over the course of the experiment. On average, participants play more than 150 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 two information conditions: an 'unequal information' or [1 3] condition, in which participants play one bandit once and the other three times, and an 'equal information' or [2 2] condition, in which participants play both bandits twice.

Relative to the original Horizon Task, the key modification in this paper is to give people 'repeated games' (Figure 1B), in which they see the exact same set of forced-choice plays twice in two separate games separated by several minutes in time so as to avoid detection. By repeating the forced-choice 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, then they will make the same decision in both games and their choices will be consistent. Conversely, if their behavior is not driven by a deterministic function of the stimulus, then their choices on the repeated games will be inconsistent some fraction of the time. The extent to which participants' choices are consistent on the repeated versions of the games allow us to quantify the extent to which the variability in their behavior was driven by a deterministic process vs a random noise process.

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 using both a model-free and model-based approach (Wilson et al., 2014). In both analyses, we focus on just the first free-choice trial in each game, where the only thing that differs between the horizon conditions is the number of choices that participants will make in the future. Subsequent choices in Horizon 6 games were not analyzed.

Model-free analysis

In the model-free analysis, we quantify random and directed exploration using simple choice probabilities. Random exploration is quantified as the probability of choosing the option that has the lower average payout in the forced-choice plays in the equal, or [2 2], condition, $p(\text{low mean})$. The idea here is that, in the equal condition, the optimal strategy is to compute the mean payout for each bandit from the forced-choice plays and then always choose the option with the highest mean. When participants do not choose the option with the higher mean, the assumption is that this is due to some kind of 'decision noise', making the probability of choosing the low mean option a measure of behavioral variability. In this view, random

exploration corresponds to an increase in $p(\text{low mean})$ with horizon, which is exactly what we see in the data (Figure 2A; t(64) = 7.99, p $<$ 0.001).

Directed exploration is quantified as the probability of choosing the more informative option $p(\text{high info})$ in the unequal, or [1 3], condition. The more informative option is the option played once during the forced-choice plays as choosing this option gives relatively more information (doubling the number of samples from 1 to 2) than choosing the option played three times (only increasing the number of sample by a third, from 3 to 4). In this view, directed exporation corresponds to an increase in $p(\text{high info})$ with horizon, which is exactly what we see in the data (Figure 2B; t(64) = 6.92, p $<$ 0.001).

Model-based analysis

Another approach to understanding behavior in the Horizon Task is to use a computational model (Wilson et al., 2014). In this case, we model participants' choices on the first free-choice trial by assuming they make decisions by computing the difference in value (or utility) ΔQ between the right and left options, choosing right when $\Delta Q > 0$ and left otherwise. Specifically, we write

$$\Delta Q = \Delta R + A\Delta I + b + n \tag{1}$$

where, the experimentally controlled variables are $\Delta R = R_{right} - R_{left}$, the difference between the mean of rewards shown on the forced-choice trials, and ΔI, the difference in 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 in the [1 3] condition, and in [2 2] condition, ΔI is 0.

Here, n denotes decision noise, which, in this version of the model is a combination of deterministic and random noise. n is assumed to come from a logistic distribution with mean 0 and standard deviations σ.

The free parameters of this model are: the information bonus A, which controls the level of directed exploration; the noise standard deviation, σ, which controls the level of random exploration, and the spatial bias, b, which determines the extent to which participants prefer the option on the right. These free parameters are fit separately for each participant in each horizon condition, allowing us to test whether directed and random exploration increase with horizon. Consistent with previous research, we find that this is indeed the case (Figure 2C; t(64) = 5.35, p $<$ 0.001. Figure 2D; t(64) = 3.54, p $<$ 0.001).

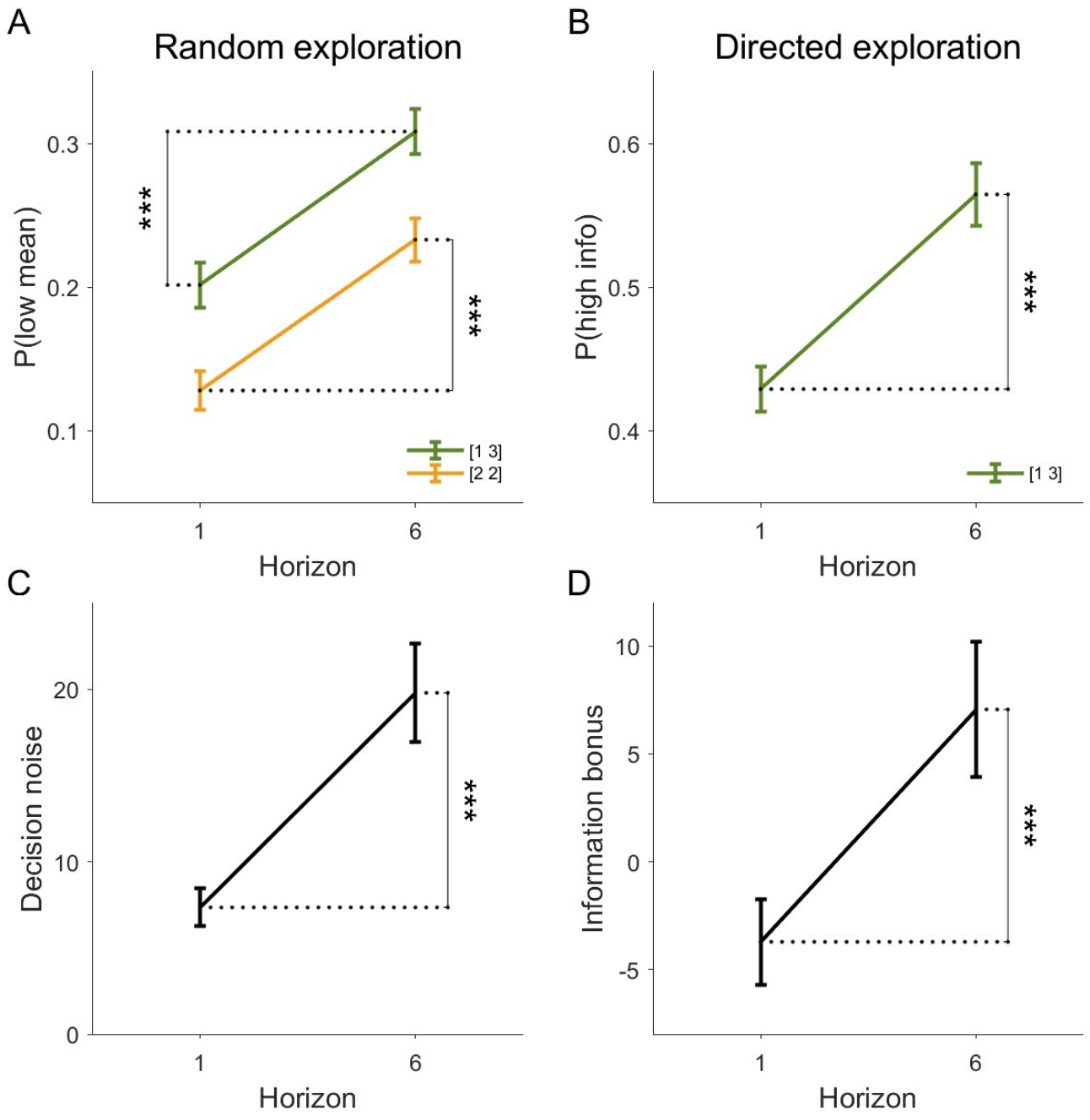


Figure 2: Replication of previous findings that people use both random and directed exploration in this task. (A) model-free measure of behavioral variability, $p(\text{low mean})$, increases with horizon. (B) model-free measure of information seeking, $p(\text{high info})$, increases with horizon. (C) model-based measure of behavioral variability, decision noise σ, increases with horizon. (D) model-based measure of information seeking, information bonus A, increases with horizon.

Taken together, our model-free and model-based analyses agree with previous findings showing in-

creased behavioral variability and increased information seeking in the long horizon condition, consistent with humans using random and directed exploration (Figure 2, Supplementary Figure S1). However, for random exploration, this previous analysis cannot distinguish between deterministic and random sources of noise. For this we analyze the extent to which people's choices are consistent on the repeated games.

Model-free analysis of repeated games suggests that random exploration involves 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 entirely driven by a deterministic process and is also driven by random noise, participants' choices should be more inconsistent across the repetitions of the game. Moreover, if decision noise is purely random noise, meaning there is no unobserved deterministic process, we will show that we can actually predict the expected level of choice inconsistencies across repetitions of games by accounting for the known deterministic processes and assuming that the random noise process is independent in repetitions of the game.

To quantify choice inconsistency we computed the frequency with which participants made different responses for pairs of repeated games (Figure 3, Supplementary Figure S2). Using this measure we found that participants made inconsistent choices in both the unequal ([1 3]) and equal ([2 2]) information conditions ($p(\text{inconsistent}) > 0$), 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 (For [1 3] condition, t(64) = 5.41, $p < 0.001$; for [2 2] condition, t(64) = 6.26, $p < 0.001$), suggesting that at least some of the horizon dependent noise is not a deterministic function of the stimulus, but rather random noise.

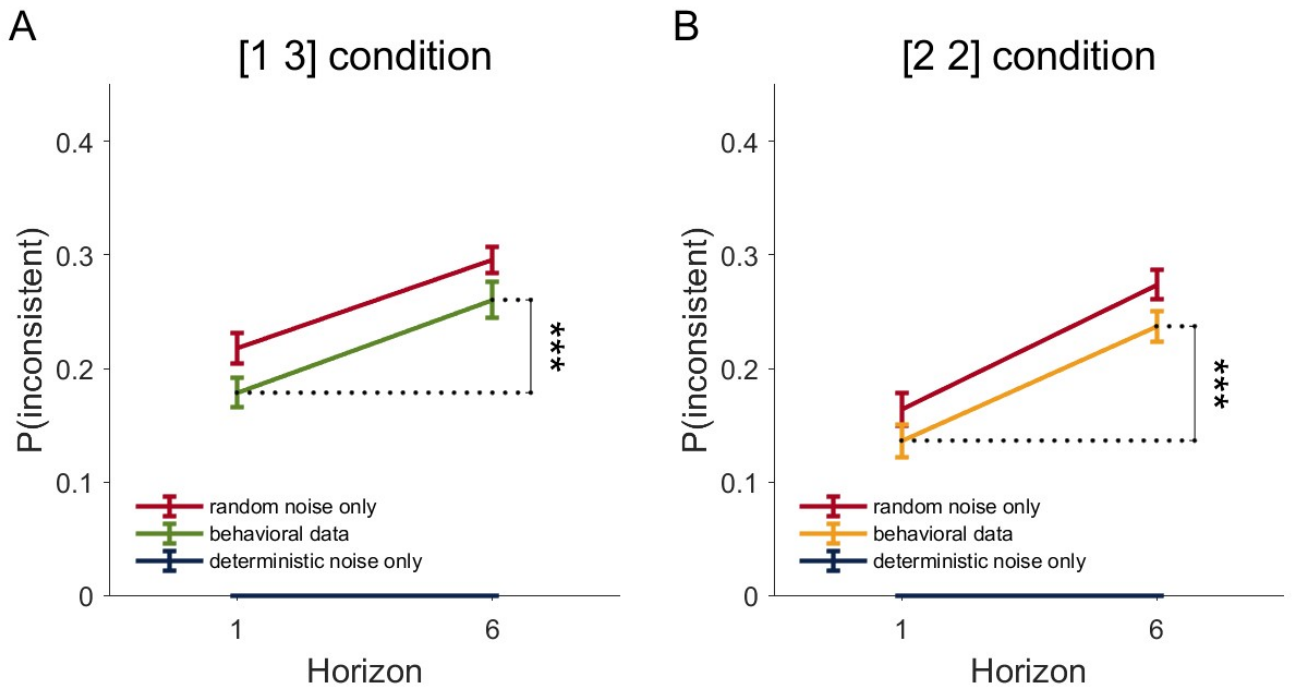


Figure 3: Model-free analysis suggests that both deterministic and random noise contribute to the choice variability in random exploration. For both the [1 3] (A) and [2 2] (B) condition, people show greater choice inconsistency in horizon 6 than horizon 1. However, the extent to which their choices 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. Since repeated decisions mean that participants either choose the low-mean option twice, or choose the high-mean option twice, we could predict the choice inconsistency, $p(\text{inconsistent})$, based on the probability of choosing the low mean option, $p(\text{low mean})$, as

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

Furthermore, to account for the fact that $p(\text{low mean})$ is a function of reward difference ΔR between

the two bandits and the information condition I, we estimated the conditional probability:

$$p(\text{inconsistent}|\Delta R, I) = 2p(\text{low mean}|\Delta R, I)(1 - p(\text{low mean}|\Delta R, I))$$

Then based on the likelihood that each condition (ΔR and I) occurs in the task $\rho(\Delta R, I)$, we have

$$p(\text{inconsistent}) = \sum_{\Delta R, I} \rho(\Delta R, I) p(\text{inconsistent}|\Delta R, I)$$

As shown in Figure 3, people's behavior falls in between the pure deterministic noise prediction and the pure random noise prediction. Specifically, behavior is different from the pure random noise prediction in both the [1 3] condition (t(64) = 4.83, $p < 0.001$ for horizon 1, t(64) = 3.12 $p = 0.003$ for horizon 6) and the [2 2] condition (t(64) = 3.92, $p < 0.001$ for horizon 1, t(64) = 3.71, $p < 0.001$ for horizon 6). Likewise, behavior is different from pure deterministic noise prediction in both the [1 3] condition (t(64) = 13.72, $p < 0.001$ for horizon 1, t(64) = 16.71, $p < 0.001$ for horizon 6) and the [2 2] condition (t(64) = 9.55, $p < 0.001$ for horizon 1, t(64) = 17.93, $p < 0.001$ for horizon 6). As a negative control of our method for estimating $p(\text{inconsistent})$ for purely random noise, we simulated choices using a decision model that only includes random noise (Equation 2), and found that $p(\text{inconsistent})$ in this simulated data is not different from our pure random noise prediction in all horizon and uncertainty conditions ($p > 0.05$, Supplementary Figure S3). Together, our results suggest that both random noise and deterministic noise contribute to the choice variability in random exploration. However, the relative contribution from each of these types of noise, as well as how each type of noise changes with horizon, are difficult to discern.

Model-based analysis provides a lower-bound estimate of deterministic noise and an upper-bound estimate of random noise

To more precisely quantify the contribution of deterministic noise and random 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 (Equation 1) 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'). 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

To model participants' choices on the first free-choice trial, we use a modified version of Equation 1.

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

where, as before ΔR, is the the difference in mean rewards shown on the forced-choice trials, ΔI, is the difference in information, A is the information bonus, and b is the spatial bias. New in Equation 2 are the terms n_{det} and n_{ran}. 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 repeated plays and changes every game. n_{det} and n_{ran} are assumed to come from logistic distributions with mean 0, and standard deviations σ_{det} and σ_{ran}.

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 $n^i_{ran} \neq n^j_{ran}$. While an increase in either random or deterministic noises could lead to higher $p(\text{low mean})$, an increase in random noise predicts higher $p(\text{inconsistent})$ while an increase in deterministic noise predicts lower $p(\text{inconsistent})$.

We used hierarchical Bayesian analysis to fit the parameters of the model (see Figure 8 for a 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.

We also fit a series of reduced and alternative models to the data. This includes reduced models that assume only deterministic or random noises. We also fit models in which the standard deviation of random and deterministic noises σ_{det} and σ_{ran} are estimated separately for [1 3] and [2 2] information conditions. Lastly, we fit a model with an alternative definition of ΔI that ΔI is defined to be difference between the variances of rewards shown on the forced-choice trials. Results on these model variants are presented in the Supplementary Materials.

Model validation

To be sure that our fit parameter values were meaningful and to understand the limits of our model, we evaluated our model extensively using simulated data. This allowed us to quantify whether deterministic

and random noise can be identified under ideal conditions where the behavior is generated by the model with known parameters. Full details are presented in the Supplementary Materials section 2.

In this section we focus on our results for parameter recovery (Wilson and Collins, 2019). In a parameter recover analysis, behavioral data is simulated by the model with known parameters and then this simulated behavioral data is fit with the model to quantify the extent to which fit parameters match the input simulated parameters — that is, whether the simulated parameters can be recovered.

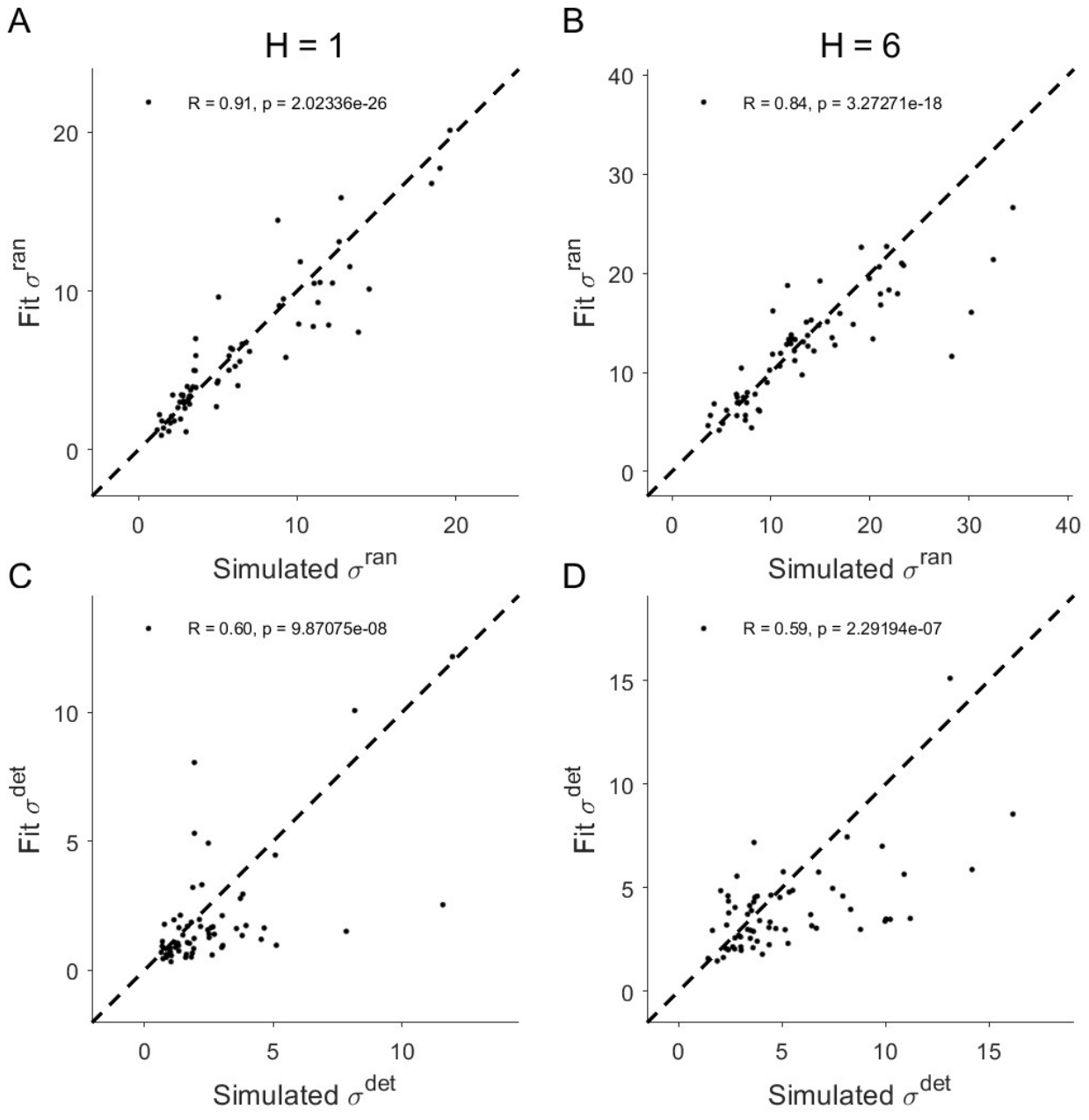


Figure 4: Parameter recovery analysis for random (A,B) and deterministic (C,D) noises in the two horizons.

Parameter recover in this task was good for this model (Figure 4, Supplementary Figure S7, S8), with fit values for σ_{ran} and σ_{det} showing strong correlations with their simulated values in both horizon (H) conditions (For σ_{ran}, $R = 0.91(\text{H} = 1)$ and $0.84(\text{H} = 6), p < 0.001$, For σ_{det}, $R = 0.60(\text{H} = 1)$ and $0.59(\text{H} = 6), p < 0.001$). However, while the relationship was near perfect for random noise ($\frac{\text{Recovered } \sigma_{ran}}{\text{Simulated } \sigma_{ran}} = 1.01$), there was a systematic bias to underestimate the level of deterministic noise by about 32% ($\frac{\text{Recovered } \sigma_{det}}{\text{Simulated } \sigma_{det}} = 0.68$). Despite this underestimation of deterministic noise in both horizon conditions, the difference in deterministic noise between horizons is much better captured (see Supplementary Materials section 2.2). This is because the underestimation of deterministic noise is partially canceled out when the difference is taken between horizon conditions. In addition, we see better parameter recovery for random noise than deterministic noise. This is likely because we effectively have half as many trials for deterministic noise. In particular, while we generate two samples of random noise for each repeated game pair, we only generate one sample of deterministic noise, which by definition is the same in both of the repeated games.

In addition to the conventional subject-level parameter recovery analysis presented here, we also performed parameter recovery analysis that examined how faithful the full posterior distribution of group-level parameters can be recovered in simulated data (Supplementary Figure S5, S6, S9, S10). Qualitatively, we also showed that our way of modeling deterministic noise is capable of capturing known deterministic processes intentionally omitted from the full model (Supplementary Figure S4). Full details of these additional analysis are presented in the Supplementary Materials.

Overall, we were able to detect both deterministic and random noises using our model. Because random noise is modeled as non-stimulus-driven noise, it can reflect both true stochastic random noise and possible deterministic noises which do not depend on the stimuli. Thus conceptually our random noise estimate provides an upper bound for the true 'random noise' induced by intrinsic stochastic processes in the brain. Thus, our model provides a lower bound for deterministic noise and an upper bound for random noise.

Model-based results

Posterior distributions over the group-level means of the deterministic and random noise standard deviation σ_{det} and σ_{ran} are shown in Figure 5 and Supplementary Figure S11. Consistent with our model-free results, we see that both random and deterministic noise are non-zero. Numerically, random noise is about 2-3 times larger than the deterministic noise. By computing the posterior distribution of $\sigma_{det}^2/(\sigma_{det}^2 + \sigma_{ran}^2)$, our data suggests that 14.25% of the variability in random exploration is accounted for by deterministic

noise ([4.90%, 28.81%], 95% CI). In addition, we find that both random and deterministic noise increase with horizon. This increase was larger for random noise (mean = 7.13, 100% of samples showed an increase in random noise with horizon) than deterministic noise (mean = 2.59, 98.64% of samples showed an increase in deterministic noise with horizon). But intriguingly, the relative increase in both types of noise was similar (Figure 6). That is, when we compute the relative increase in deterministic noise with horizon, $\sigma^{det}_{horizon6}/\sigma^{det}_{horizon1}$, it is very similar to the relative increase in random noise with horizon $\sigma^{ran}_{horizon6}/\sigma^{ran}_{horizon1}$.

Similar results are found in other variants of our model. Results for a model that estimates random and deterministic noises separately for [1 3] and [2 2] conditions, and a model with an alternative definition of information bonus dI, can be found in the Supplemental Materials (Supplemental Figure S12, S13).

To ensure that the joint increase of random and deterministic noises is genuine and not an artifact from the fitting procedure, we computed the correlation between ground-truth values of random noise, and best-fitting values of deterministic noise (and vice versa), and they do not correlate (Supplementary Figure S14). Furthermore, we simulated data from a series of reduced models with known random and deterministic noise values in which either random or deterministic noise does not change with Horizon, and fit our model to the simulated data. Our model detects and only detects a change in random/deterministic noise with horizon, when the change is present in the model that simulates the data (Supplementary Figure S15, S16).

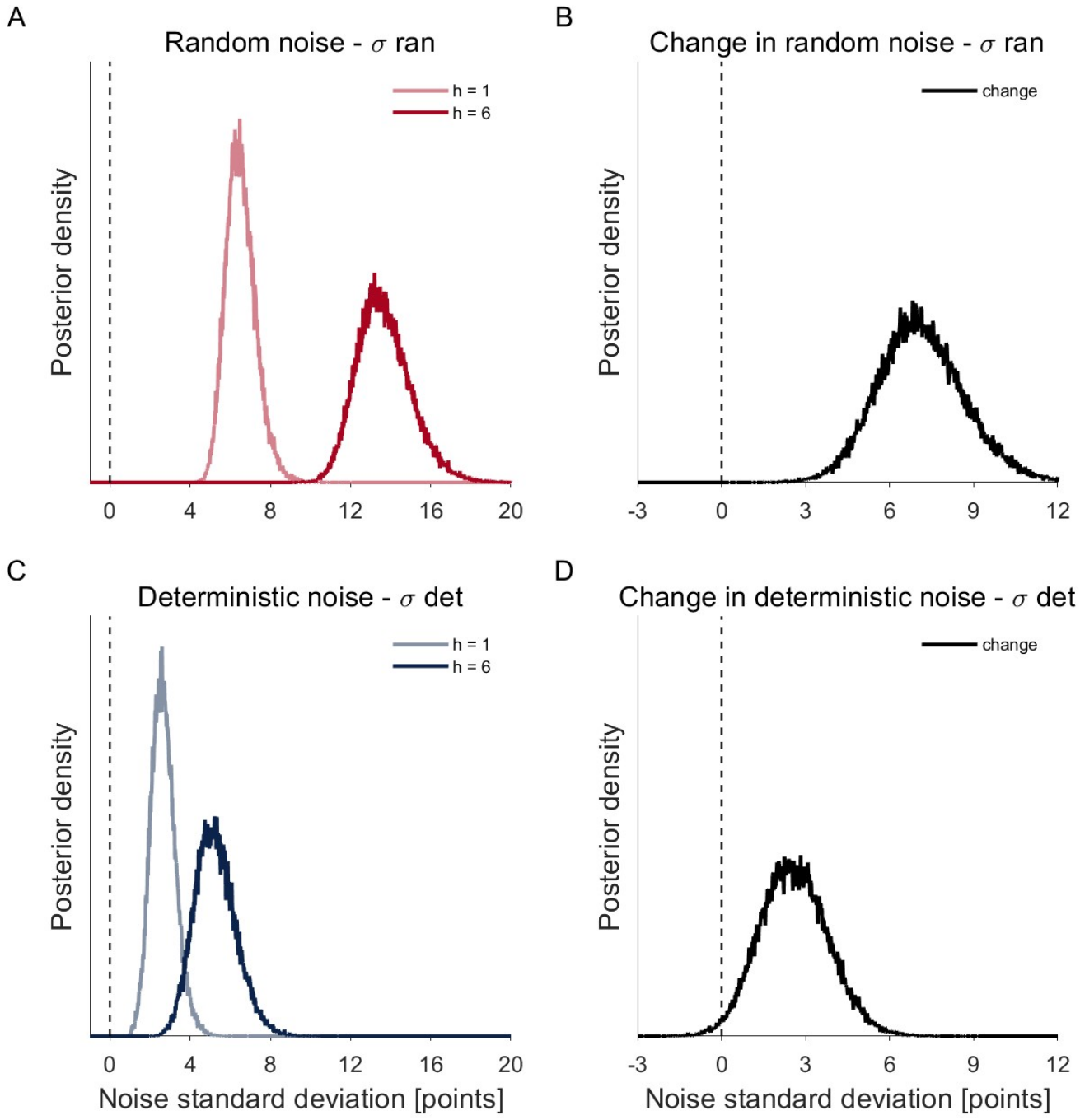


Figure 5: Model based analysis showing the posterior distributions over the group-level mean of the standard deviations of random and deterministic noise. Both random (A, B) and deterministic (C,D) noises are nonzero (A, C) and increase with horizon (B, D).

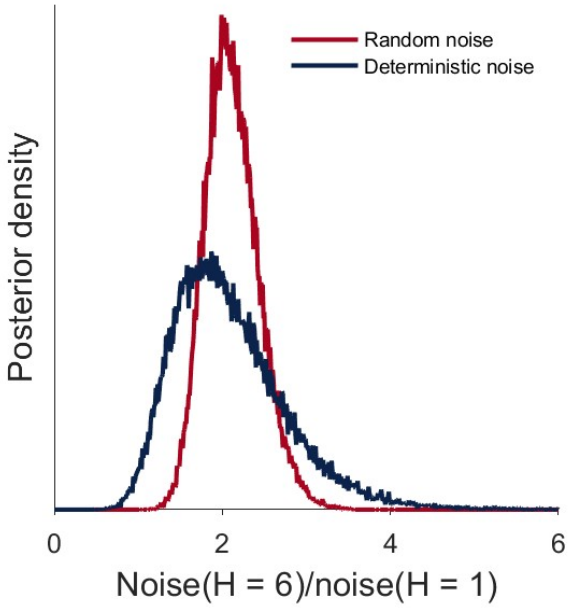


Figure 6: Model based analysis showing the posterior distributions over the ratio of the group-level mean of the standard deviations of random and deterministic noise between horizon 6 and horizon 1 respectivelly. The ratios in the standard deviations of noises between horizon 6 and horizon 1 are similar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

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 (Palminteri et al., 2017, 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) and for p(inconsistent), but the model slightly overestimate p(low mean) in [2 2] conditions. This has to do with the skewness of the subject-level posterior distribution (see Supplemental Materials).

As a control, we also applied posterior predictive checks on alternative models that consider only deterministic or only random noise, and these reduced models fail to capture all qualitative patterns (Supplementary Figure S17, S18). Full details of this analysis can be found in Supplementary Materials section 3.5.

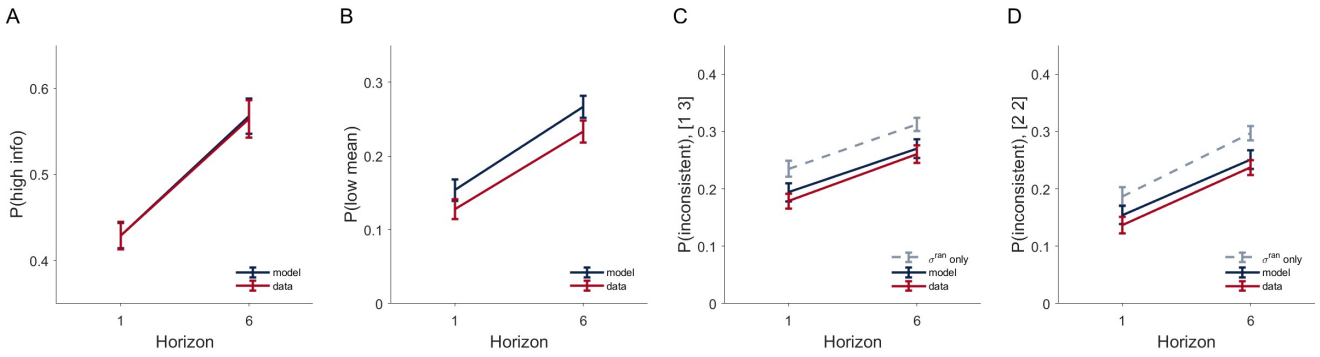


Figure 7: Our model accounts for all qualitative patterns of the data, namely, (A) p(high info) and (B) p(low mean) increase as a function of horizon, p(inconsistent) increases as a function of horizon for both [1 3] (C) and [2 2] (D) conditions and it lies between the pure random and pure deterministic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

Discussion

In this paper, we investigated whether random exploration is really random or whether it is 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 around 14% of the overall variability in people's behavior.

One interpretation for this low level of deterministic noise is that most of the variability in random exploration is truly random. Such a random noise interpretation, would be consistent with recent work showing that variability in perceptual decisions may be driven by imperfections in mental inference (Drugowitsch et al., 2016). In this view, apparently random behavior is not due to sensory processing or response selection, but to suboptimal computations in the brain. Although suboptimal inference is different from simply adding random noise to neural circuitry(Beck et al., 2012), as long as the suboptimality in neural computation is not a deterministic function of the stimuli, it is a form of random noise in our definition. Indeed, a strong interpretation of this hypothesis would suggest that randomness in explore-exploit behavior is

due to imperfect inference about the correct course of action. 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 (Wilson et al., 2020).

Although the random noise interpretation is theoretically appealing, our approach, while an improvement on previous methods, is not without limitations. Most important is that our measure of 'random' noise is only an upper bound on the true level of randomness and that, in principle, the random decision noise could be lower. Specifically, in our model, what we labeled random noise was really 'non-stimulus-driven variability'. While this non-stimulus-driven variability could be driven by truly random stochastic processes, it could also be driven by deterministic processing that is unrelated to the stimuli in the task. For example, such deterministic noise could be driven by differences in where people look, or for how long they look, or by whether they were fidgeting or scratching their nose (Musall et al., 2019). Another limitation is that deterministic noise is defined based on the stimulus within a game, between-game deterministic strategies and stimuli from previous games (e.g., memory of previously seen game) were also treated as random noises in our model, although our model could be potentially extended by considering deterministic noise over a sequence of stimuli across games (Wyart, 2018, Wyart and Koechlin, 2016). In addition to this conceptual limitation in measuring deterministic noise, parameter recovery simulations suggest that our estimation method also slightly underestimates deterministic noise (see Figure 4, Supplementary Figure S5, S6). As a result, from both a conceptual and methodological perspective, it is possible that the remaining 86% of the decision noise that is not stimulus-driven noise, could be deterministic.

Like the random noise account, the deterministic noise account is also in line with previous work in which neural variability can be accounted for by fluctuations in sensory inputs. For example, MT neurons were shown to have a reproducible temporal modulation in response to a fixed random motion stimulus (Bair and Koch, 1996). In other words, 'irrelevant' features in the stimuli are represented in a reliable way in the brain that could drive downstream choices in a predictable way.

Regardless of whether we interpret the noise as random or deterministic, a key finding in this paper is that both types of noise change with horizon. Such a horizon increase is a hallmark of an exploratory process and suggests that the modulation of deterministic and random processes may underlie random exploration. Moreover, the fact that the horizon change in the two types of noise are proportional to each other (Figure 6) suggests a possible mechanism for random exploration: a reduction in the strength with which reward drives the choice.

We show first that a change in noise is mathematically equivalent to a change in reward signal strength

in our decision model (see also Cinotti et al. (2019)).. To show how a change in reward processing could affect random and deterministic noise, consider the simple decision model we introduced in Equation 2. In this model, choice is determined by the sign of the difference in utility ΔQ between the two options, where

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

Now imagine a case where the reward signal is scaled by a factor β. In this case, ΔQ becomes

$$\Delta Q = \beta\Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{4}$$

Because the choice only depends on the sign of ΔQ, scaling ΔQ by a factor of $1/\beta$ will not change the behavior of the model. Thus, if we divide both sides of the above equation by β we get

$$\Delta Q/\beta = \Delta R + A\Delta I/\beta + b/\beta + n_{det}/\beta + n_{ran}/\beta \tag{5}$$

which is equivalent to a scaling of both deterministic and random noise by the same factor $1/\beta$. Thus, one interpretation of our result that both deterministic and random noise change across horizons with the same ratio, is that this reflects a change in reward processing. That is, the reward signal is reduced in the longer horizon condition (smaller β in horizon 6 than horizon 1).

Such a reduction in the strength of reward coding in exploration, is consistent with our recent work using a drift diffusion model (DDM) to model explore-exploit decisions (Feng et al., 2021). In the drift diffusion model, changes in behavioral variability can be driven by changes in the decision threshold (smaller threshold = more noise) or changes in the signal-to-noise ratio with which reward is encoded (lower SNR = more noise). By fitting both choices and response times, we were able to distinguish between these two accounts showing that the majority of the horizon-change in variability was driven by changes in SNR and not threshold. However, this model could not determine whether the changes in SNR were driven by signal or noise. By showing that the change in deterministic and random noise have approximately the same ratio, the present work suggests that this SNR change is driven by changes in reward-signal processing, not noise. Of course, to truly see whether changes in signal or noise are driving random exploration will require more direct measurements of neural processing such as with neuroimaging and electrophysiology (Costa et al., 2019, Ebitz et al., 2018, Hogeveen et al., 2022, Tomov et al., 2020)

Materials and Methods

Ethics statement

Human subject protocols were approved by the University of Arizona institutional review board (IRB # 1411567117). Written informed consent was given by all participants prior to participating in the study.

Participants

80 participants (ages 18-25, 37 male, 43 female) from the University of Arizona undergraduate subject pool participated in the experiment. 15 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 by making a completely random choice between the two options and excluded participants with a likelihood greater than 0.1%, in other words, participants who didn't show an accuracy significant above chance with $p < 0.001$ were excluded in the analysis. This left 65 for the main analysis. Note that including the 15 badly performing subjects did not change the main results (Supplementary Figures S1, S2, S11)

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 153.7 games (minimum = 90 games, maximum = 192 games) and the whole task lasted between 12.37 and 32.15 minutes (mean 22.78 minutes). Participants played an average of 65.3 repeated pairs of games (minimum = 30 repeated pairs, maximum = 79 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 'forced-choice' 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 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).

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 deterministic noise from 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

$$\Delta Q = \Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{6}$$

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 of 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. 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} and n_{ran} are deterministic noise and random noise respectively. 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 (Table 1): 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}. 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 (Figure 8). 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.

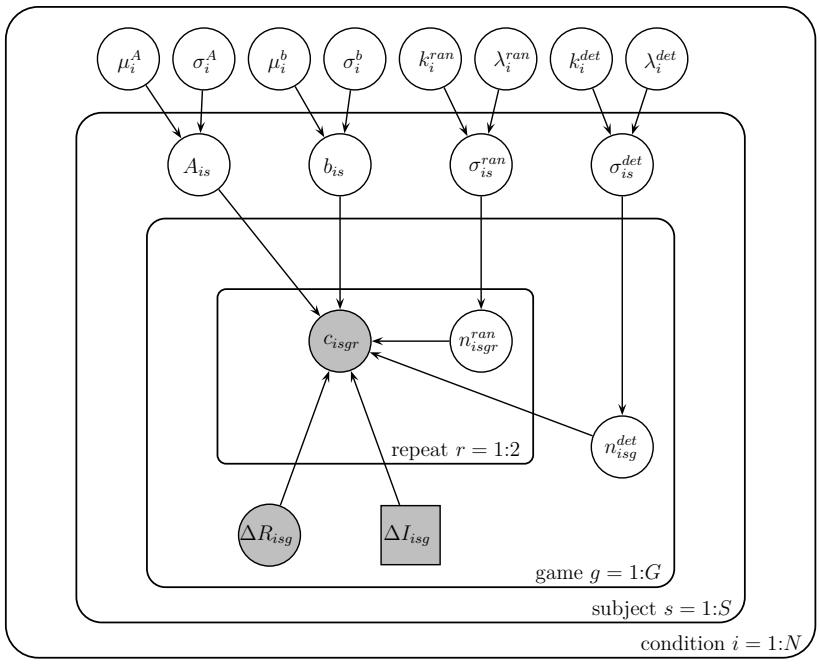
Parameter	Prior	Hyperparameters	Hyperpriors
information bonus, A_{is}	$A_{is} \sim$ Gaussian(μ_i^A, σ_i^A)	$\theta_i^A = (\mu_i^A, \sigma_i^A)$	$\mu_i^A \sim$ Gaussian(0, 100) $\sigma_i^A \sim$ Exponential(0.01)
spatial bias, b_{is}	$b_{is} \sim$ Gaussian(μ_i^b, σ_i^b)	$\theta_i^b = (\mu_i^b, \sigma_i^b)$	$\mu_i^b \sim$ Gaussian(0, 100) $\sigma_i^b \sim$ Exponential(0.01)
deviation of deterministic noise, σ_{isg}^{det}	$\sigma_{is}^{det} \sim$ 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}	$\sigma_{is}^{ran} \sim$ 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 10 independent Markov chains to generate 50000 samples from the posterior distribution over parameters (5000 samples per chain). Each chain had a burn in period of 5000 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

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

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

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

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

Subject specific parameters

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

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

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

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

Deterministic noise for repeated game

$n_{isg}^{det} \sim \text{Logistic}(0, \sigma_{is}^{det})$

Random noise for each game

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

Observed choices

$\Delta Q_{isgr} \leftarrow \Delta R_{isg} + A_{is}\Delta I_{isg} + b_{is} + n_{isgr}^{ran} + n_{isg}^{det}$

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

Figure 8: Schematic of the hierarchical Bayesian model using notation of Lee and Wagenmakers (2014b)

Data and code

Behavioral data as well as MATLAB codes to recreate the main figures from this paper will be made available upon publication at `https://github.com/wangxsiyu/RW_RandomDeterministicNoise.git`

Supporting information

S1 Fig. Replication of previous findings with data from all participants (i.e. no exclusions). (A) model-free measure of behavioral variability, $p(\text{low mean})$, increases with horizon. (B) model-free measure of information seeking, $p(\text{high info})$, increases with horizon. (C) model-based measure of behavioral variability, decision noise σ, increases with horizon. (D) model-based measure of information seeking, information bonus A, increases with horizon.

S2 Fig. Model-free analysis with data from all participants (i.e. no exclusions) 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..

S3 Fig. Model-free analysis with simulated choices from a model that has only random noise validates our prediction of p(inconsistent) for pure random noise. The extent to which simulated choices are inconsistent completely overlaps with our pure random noise prediction($p > 0.05$). This suggests that when choice inconsistency lies below the pure random noise prediction indeed provides evidence that deterministic noise exists in random exploration (Figure 3).

S4 Fig. Deterministic noise can recover known deterministic processes that's intentionally omitted by the model. In the reduced model where the deterministic effect of uncertainty condition is omitted from the model, deterministic noise is higher compared to the full model that accounts for the effect of uncertainty. Random noise remains unchanged between the two models.

S5 Fig. Hyperprior recovery. Parameter recovery over the posterior distribution of random and deterministic noise standard deviations σ_{det} and σ_{ran}. Solid lines are true posterior used to simulate choices. Lighter color shades represent the re-fitted posterior to the simulated choices. Our model fitting procedure faithfully recovers the non-stimulus-driven random noise (A, B), but systematically underestimates deterministic noise in both horizons (D, E). The horizon differences in random noise is also faithfully recovered (C). The horizon differences in deterministic noise is also underestimated but not significant (F).

S6 Fig. Frequentist coverage analysis. Parameter recovery over the mean estimates of random and deterministic noise standard deviations σ_{det} and σ_{ran}. Solid lines are true posterior used to simulate choices, dashed black line is the mean of the true posterior. Histograms represent the mean estimates of the respective parameters in the refitting to the simulated data. (A) and (B) are random noise at H = 1 and H = 6, respectively. (C) is the random noise differences between horizons. (D) and (E) are deterministic noise at H = 1 and H = 6, respectively. (F) is the deterministic noise differences between horizons.

S7 Fig. Parameter recovery. Parameter recovery over the subject-level means of information bonus, A, spatial bias, b, random noise standard deviation, σ_{ran}, and deterministic noise standard deviation, σ_{det}, for horizon 1 (left column) and horizon 6 (right column) games.

S8 Fig. Parameter recovery (200 repetitions). Same as Figure S7, except that the recovered parameters were averaged across 200 repetitions and then compared to the original parameters.

S9 Fig. Parameter recovery with 0 random noise or 0 deterministic noise. Parameter recovery over the posterior of random noise standard deviation, σ_{ran}, and deterministic noise standard deviation, σ_{det}, for purely random noise (top row) and purely deterministic noise (bottom row) games.

S10 Fig. Parameter recovery on arbitrary combinations of random and deterministic noises. A. Recovered posterior distributions of random noise. B. Recovered posterior distributions of deterministic noise. For both A and B, from the top row to the bottom row, the true noise standard deviation that is used in the simulations go from 0 to 10. The y limit of each panel is 4 (+/- 2 from the true value). Our model did a relatively good job in recovering all combinations of deterministic and random noises.

S11 Fig. Model based analysis with data from all participants (i.e. no exclusions) 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.

S12 Fig. Model based analysis from a model that estimates random and deterministic noises separately for [1 3] and [2 2] conditions. The posterior distributions over the group-level mean of the standard deviations of random and deterministic noise. Both random (A, E) and deterministic (C,G) noises are nonzero (A, C, E, G) and change with horizon (B, D, F, H). However, random noise has both a greater magnitude overall (A, E) and a greater change with horizon (B, F) than deterministic noise. Moreover, both random and deterministic noises have a greater magnitude in [1 3] compared to [2 2] conditions.

S13 Fig. Model based analysis from a model that uses variance differences as dI. 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.

S14 Fig. Parameter recovery for shuffled data. To show that the joint increase of random and deterministic sources of noise is not caused by a limitation of the fitting procedure, we calculated the correlation between ground-truth values of random noise, and best-fitting values of deterministic noise (and vice versa). Ground-truth values are shuffled best-fit parameters. As expected, ground-truth random values do not correlate with recovered deterministic noises, showing that the increase of deterministic noise with horizon is genuine and not a by-product of increase of random noise, and vice versa.

S15 Fig. Model based analysis with reduced models. Each row is one model. These models varied in whether deterministic σ^{det} and random noise σ^{ran} are present or not and whether either types of noise is dependent on horizon (subscript denotes the dependence on horizon).

S16 Fig. Hyperprior recovery of reduced models. Our model qualitatively captures whether deterministic and random noise are present or not and whether either types of noise is dependent on horizon. A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S17 Fig. Posterior checks for reduced models Model comparison. A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S18 Fig. Posterior checks for reduced models (maximal likelihood estimation) Model comparison (using maximal likelihood estimation). A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S1 Table. Variants of the model.

S1 File. Supplemental Materials containing additional analyses.

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.

Wyeth Bair and Christof Koch. Temporal Precision of Spike Trains in Extrastriate Cortex of the Behaving Macaque Monkey. *Neural Computation*, 8(6):1185–1202, 1996. ISSN 08997667. doi: 10.1162/neco.1996.8.6.1185.

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

J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–9, 2012. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2012.03.016. URL `https://www.ncbi.nlm.nih.gov/pubmed/22500627`. Beck, Jeffrey M Ma, Wei Ji Pitkow, Xaq Latham, Peter E Pouget, Alexandre eng R01 EY020958/EY/NEI NIH HHS/ R01EY020958/EY/NEI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2012/04/17 Neuron. 2012 Apr 12;74(1):30-9. doi: 10.1016/j.neuron.2012.03.016.

D. H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

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`.

François Cinotti, Virginie Fresno, Nassim Aklil, Etienne Coutureau, Benoît Girard, Alain R. Marchand, and Mehdi Khamassi. Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, 9(1):6770, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-43245-z. URL https://doi.org/10.1038/s41598-019-43245-z.

V. D. Costa, A. R. Mitz, and B. B. Averbeck. Subcortical substrates of explore-exploit decisions in primates. *Neuron*, 103(3):533–545 e5, 2019. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2019.05.017. URL https://www.ncbi.nlm.nih.gov/pubmed/31196672. Costa, Vincent D Mitz, Andrew R Averbeck, Bruno B eng Z99 MH999999/ImNIH/Intramural NIH HHS/ ZIA MH002928/ImNIH/Intramural NIH HHS/ ZIA MH002928-09/ImNIH/Intramural NIH HHS/ Research Support, N.I.H., Intramural 2019/06/15 Neuron. 2019 Aug 7;103(3):533-545.e5. doi: 10.1016/j.neuron.2019.05.017. Epub 2019 Jun 10.

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.

Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4):495–506, 2002. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(02)00044-8. URL https://www.sciencedirect.com/science/article/pii/S0893608002000448.

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.

R. B. Ebitz, E. Albarran, and T. Moore. Exploration disrupts choice-predictive signals and alters dynamics in prefrontal cortex. *Neuron*, 97(2):475, 2018. ISSN 1097-4199 (Electronic) 0896-6273 (Linking). doi: 10.1016/j.neuron.2018.01.011. URL https://www.ncbi.nlm.nih.gov/pubmed/29346756. Ebitz, R Becket Albarran, Eddy Moore, Tirin eng Published Erratum 2018/01/19 Neuron. 2018 Jan 17;97(2):475. doi: 10.1016/j.neuron.2018.01.011.

Samuel F. Feng, Siyu Wang, Sylvia Zarnescu, and Robert C. Wilson. The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, 11(1): 3077, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82530-8. URL https://doi.org/10.1038/s41598-021-82530-8.

C. Findling, V. Skvortsova, R. Dromnelle, S. Palminteri, and V. Wyart. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nat Neurosci*, 22(12):2066–2077, 2019. ISSN 1097-6256. doi: 10.1038/s41593-019-0518-9. 1546-1726 Findling, Charles Skvortsova, Vasilisa Dromnelle, Rémi Palminteri, Stefano Orcid: 0000-0001-5768-6646 Wyart, Valentin Orcid: 0000-0001-6522-7837 Journal Article Research Support, Non-U.S. Gov't United States 2019/10/30 Nat Neurosci. 2019 Dec;22(12):2066-2077. doi: 10.1038/s41593-019-0518-9. Epub 2019 Oct 28.

Charles Findling and Valentin Wyart. Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, 38:124–132, 2021. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2021.02.018. URL `https://www.sciencedirect.com/science/article/pii/S2352154621000401`. Computational cognitive neuroscience.

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

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

J. Hogeveen, T. S. Mullins, J. D. Romero, E. Eversole, K. Rogge-Obando, A. R. Mayer, and V. D. Costa. The neurocomputational bases of explore-exploit decision-making. *Neuron*, 110(11):1869–1879 e5, 2022. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2022.03.014. URL `https://www.ncbi.nlm.nih.gov/pubmed/35390278`. Hogeveen, Jeremy Mullins, Teagan S Romero, John D Eversole, Elizabeth Rogge-Obando, Kimberly Mayer, Andrew R Costa, Vincent D eng P51 OD011092/OD/NIH HHS/ P30 GM122734/GM/NIGMS NIH HHS/ ZIA MH002929/ImNIH/Intramural NIH HHS/ P20 GM109089/GM/NIGMS NIH HHS/ ZIA MH002928/ImNIH/Intramural NIH HHS/ R01 MH125824/MH/NIMH NIH HHS/ Research Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, U.S. Gov't, Non-P.H.S. 2022/04/08 Neuron. 2022 Jun 1;110(11):1869-1879.e5. doi: 10.1016/j.neuron.2022.03.014. Epub 2022 Apr 6.

Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Chapter 22 - medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In V.S. Chandrasekhar Pammi and Narayanan Srinivasan, editors, *Decision Making*, volume 202

of *Progress in Brain Research*, pages 441–464. Elsevier, 2013. doi: https://doi.org/10.1016/B978-0-444-62604-2.00022-8. URL https://www.sciencedirect.com/science/article/pii/B9780444626042000228.

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.

Katja Mehlhorn, Ben Newell, Peter Todd, Michael Lee, Kate Morgan, Victoria Braithwaite, Daniel Hausmann, Klaus Fiedler, and Cleotilde Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2, 07 2015. doi: 10.1037/dec0000033.

S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nat Neurosci*, 22(10):1677–1686, 2019. ISSN 1546-1726 (Electronic) 1097-6256 (Print) 1097-6256 (Linking). doi: 10.1038/s41593-019-0502-4. URL https://www.ncbi.nlm.nih.gov/pubmed/31551604. Musall, Simon Kaufman, Matthew T Juavinett, Ashley L Gluf, Steven Churchland, Anne K eng R01 EY022979/EY/NEI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2019/09/26 Nat Neurosci. 2019 Oct;22(10):1677-1686. doi: 10.1038/s41593-019-0502-4. Epub 2019 Sep 24.

Stefano Palminteri, Valentin Wyart, and Etienne Koechlin. The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433, 2017. ISSN 1364-6613. doi: 10.1016/j.tics.2017.03.011. URL https://doi.org/10.1016/j.tics.2017.03.011. doi: 10.1016/j.tics.2017.03.011.

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

Eric Schulz and Samuel J. Gershman. The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55:7–14, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2018.11.003. URL https://www.sciencedirect.com/science/article/pii/S0959438818300904. Machine Learning, Big Data, and Neuroscience.

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

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.

M. S. Tomov, V. Q. Truong, R. A. Hundia, and S. J. Gershman. Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nat Commun*, 11(1):2371, 2020. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-020-15766-z. URL https://www.ncbi.nlm.nih.gov/pubmed/32398675. Tomov, Momchil S Truong, Van Q Hundia, Rohan A Gershman, Samuel J eng R01 MH109177/MH/NIMH NIH HHS/ S10 OD020039/OD/NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2020/05/14 Nat Commun. 2020 May 12;11(1):2371. doi: 10.1038/s41467-020-15766-z.

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.

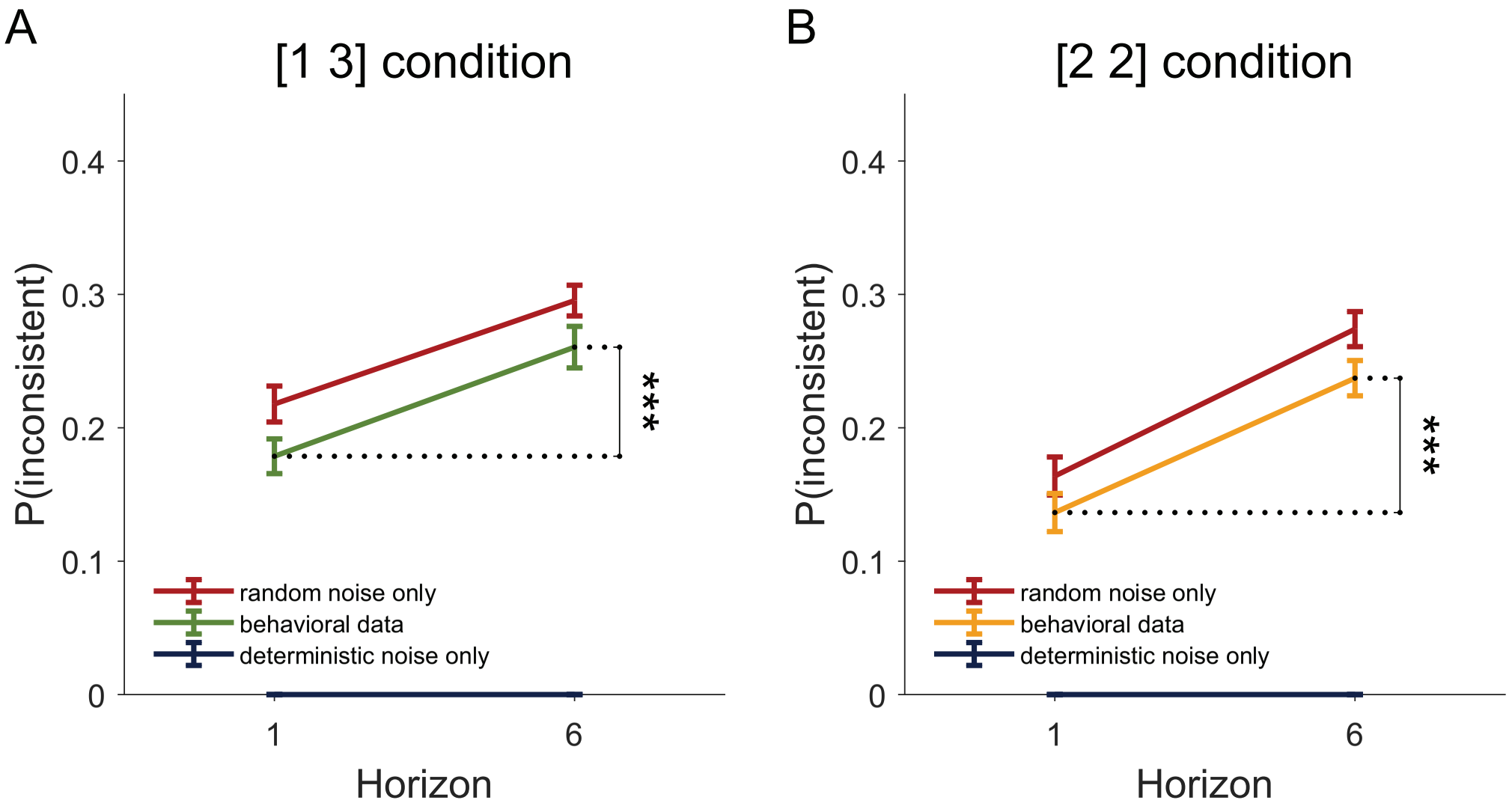
Robert C. Wilson and Anne G.E. Collins. Ten simple rules for the computational modeling of behavioral data. *eLife*, 2019. ISSN 2050084X. doi: 10.7554/eLife.49547.

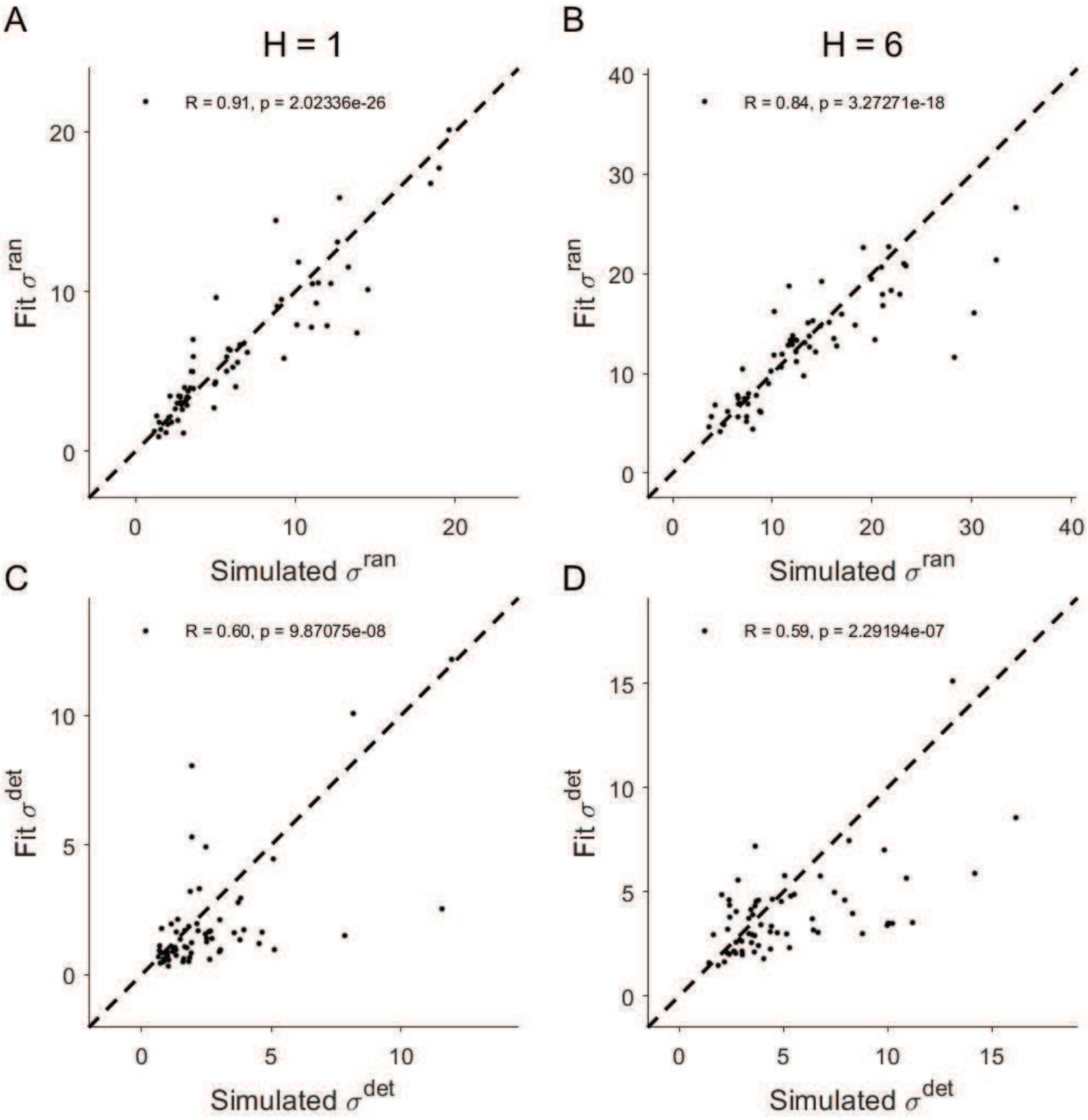
Robert C Wilson, Siyu Wang, Hashem Sadeghiyeh, and Jonathan D Cohen. Deep exploration as a unifying account of explore-exploit behavior. Feb 2020. doi: 10.31234/osf.io/uj85c. URL https://doi.org/10.31234/osf.io/uj85c.

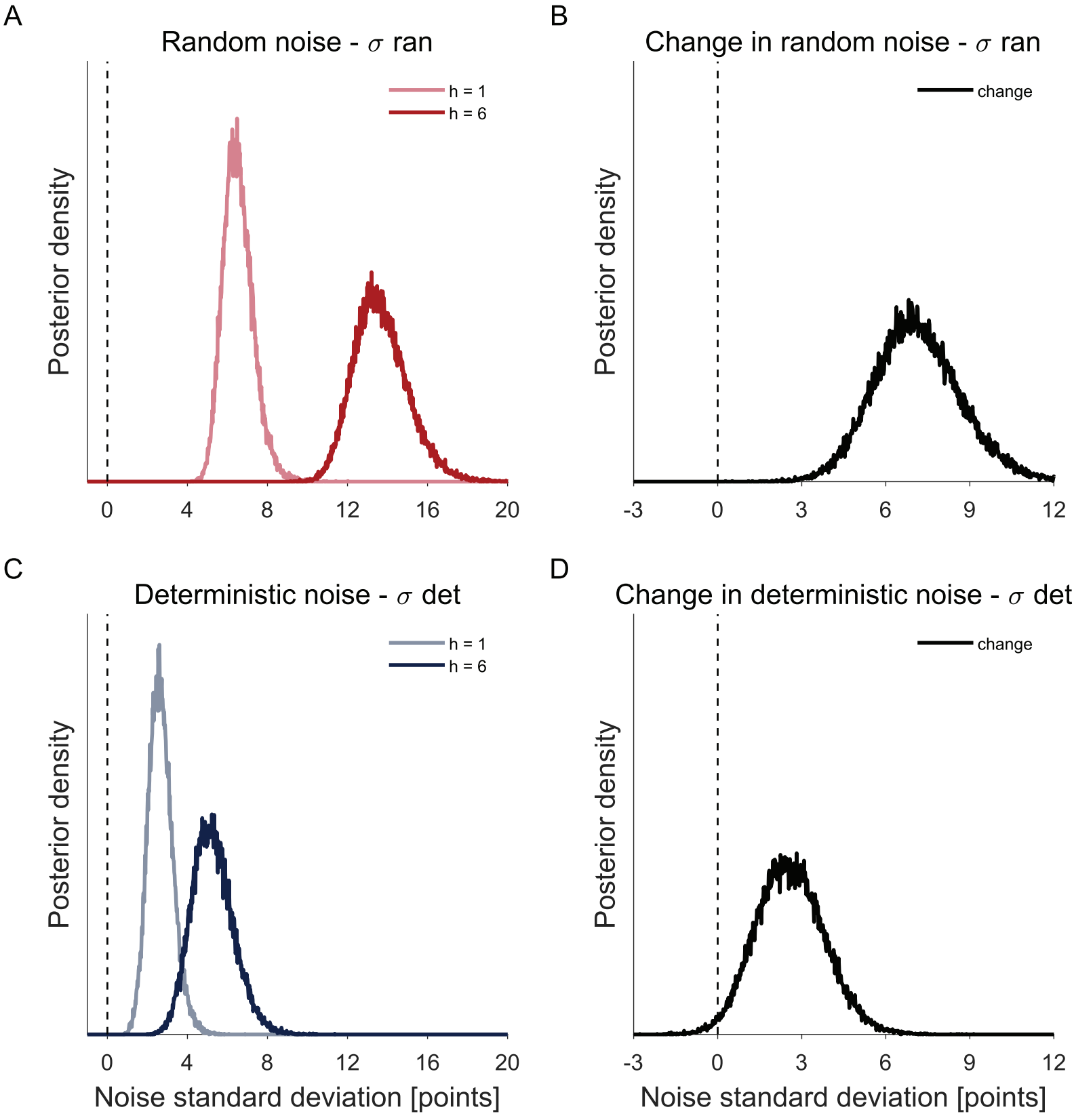
Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38: 49–56, 2021. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2020.10.001. URL https://www.sciencedirect.com/science/article/pii/S2352154620301467. Computational cognitive neuroscience.

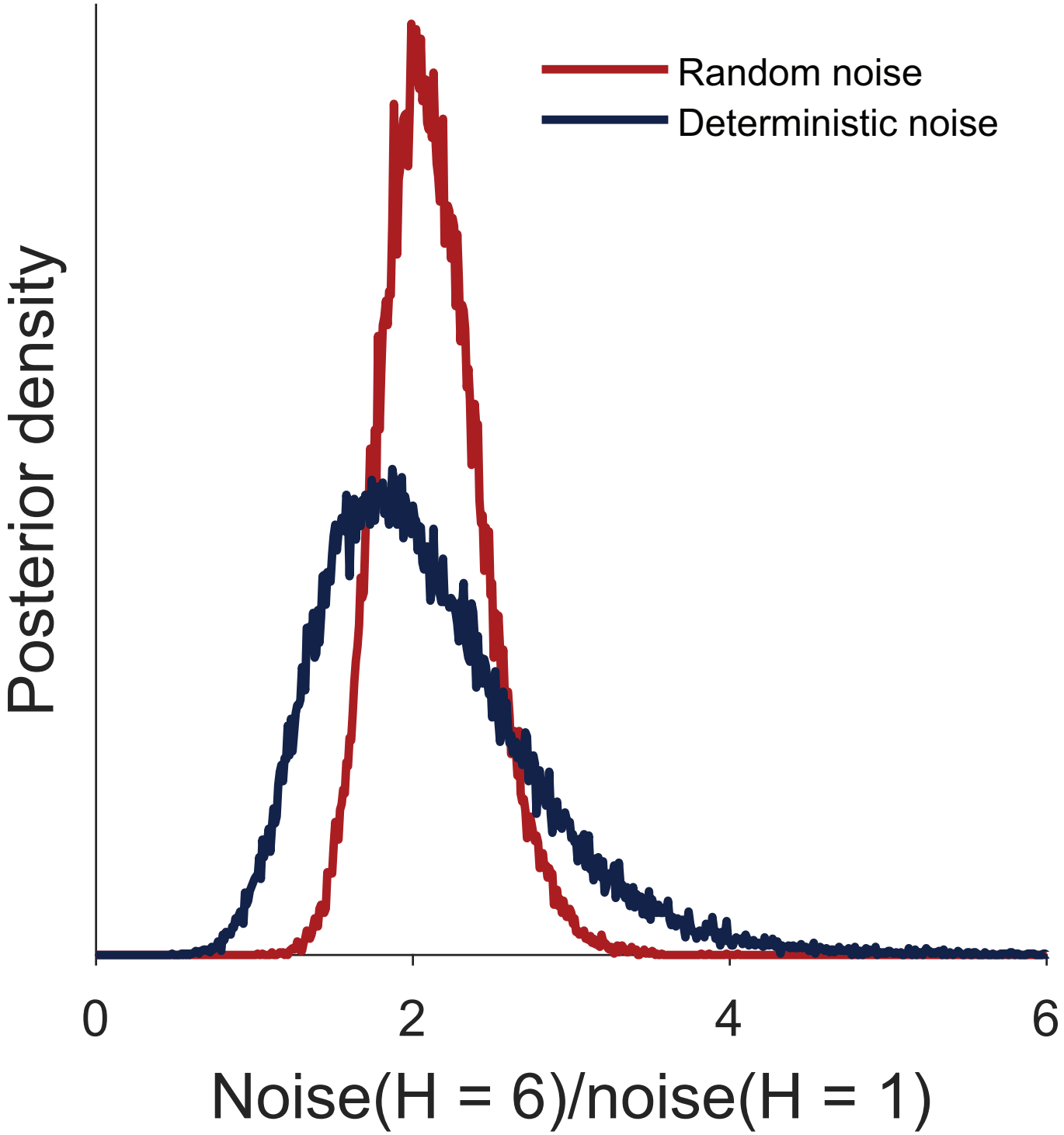
Valentin Wyart. Leveraging decision consistency to decompose suboptimality in terms of its ultimate predictability. *Behavioral and Brain Sciences*, 41:e248, 2018. doi: 10.1017/S0140525X18001504.

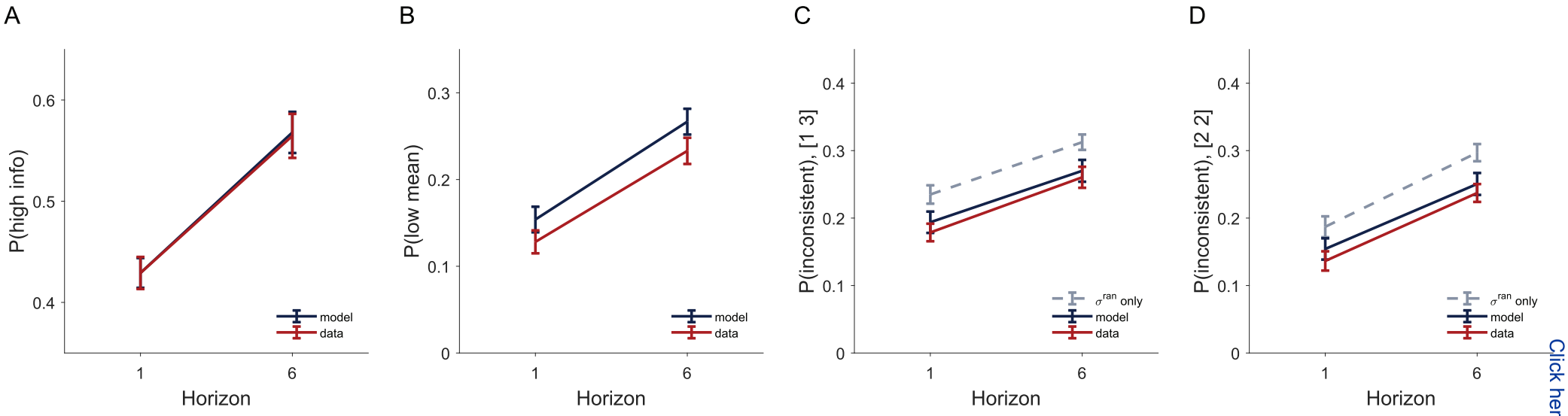
Valentin Wyart and Etienne Koechlin. Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, 11:109–115, 2016. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2016.07.003. URL https://www.sciencedirect.com/science/article/pii/S235215461630136X. Computational modeling.

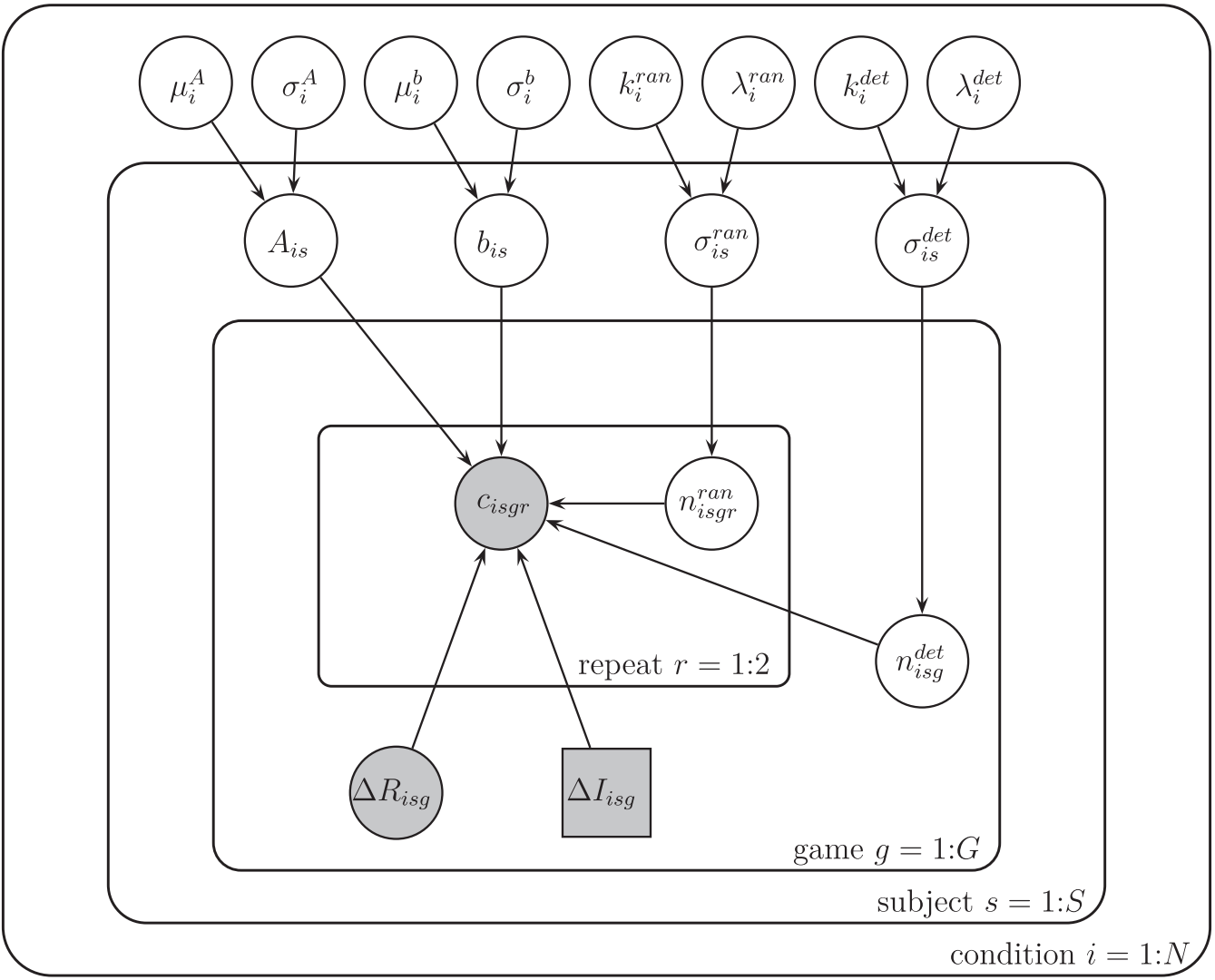












Priors

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

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

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

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

Subject specific parameters

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

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

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

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

Deterministic noise for repeated game

$n_{isg}^{det} \sim \text{Logistic}(0, \sigma_{is}^{det})$

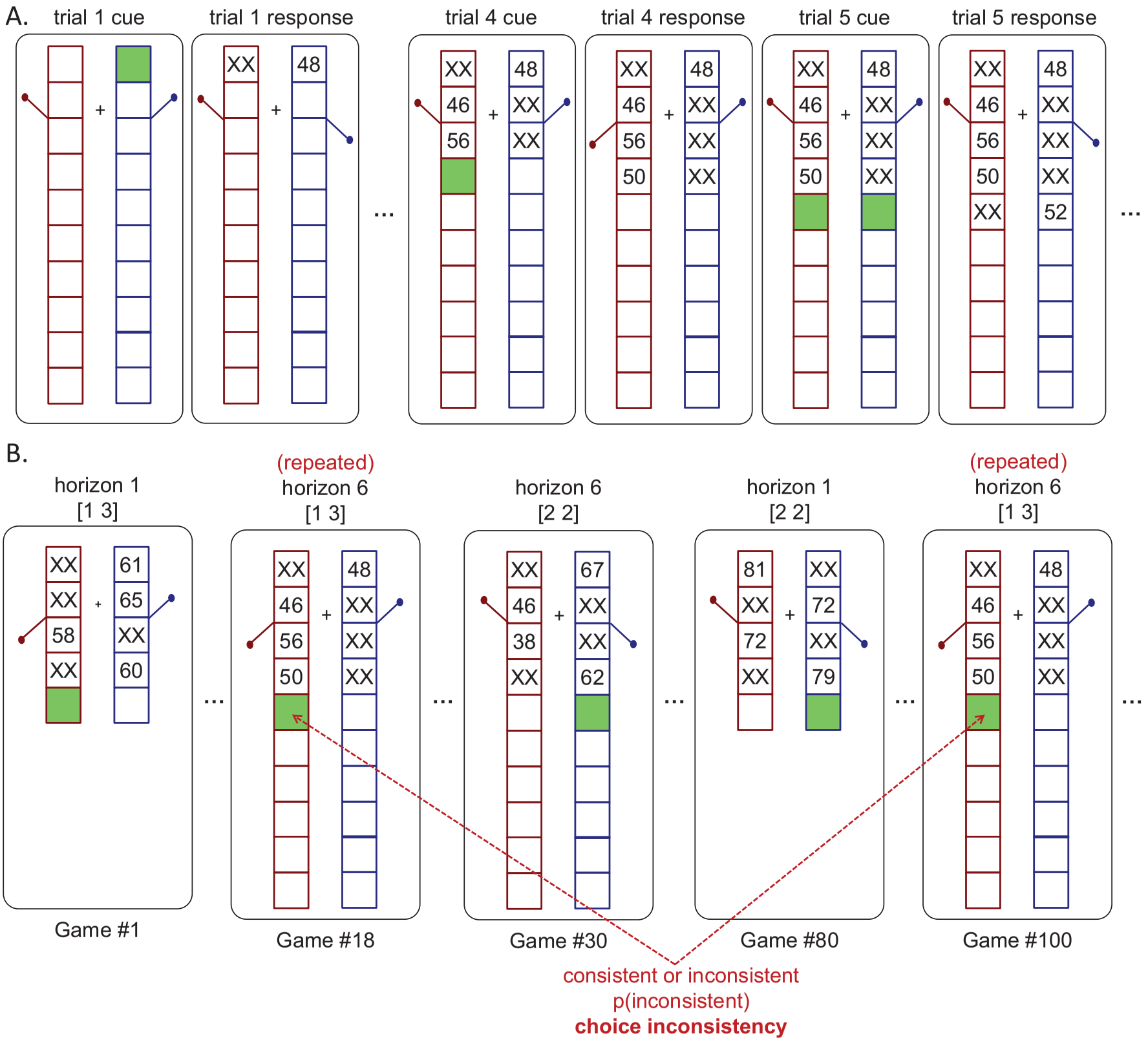
Random noise for each game

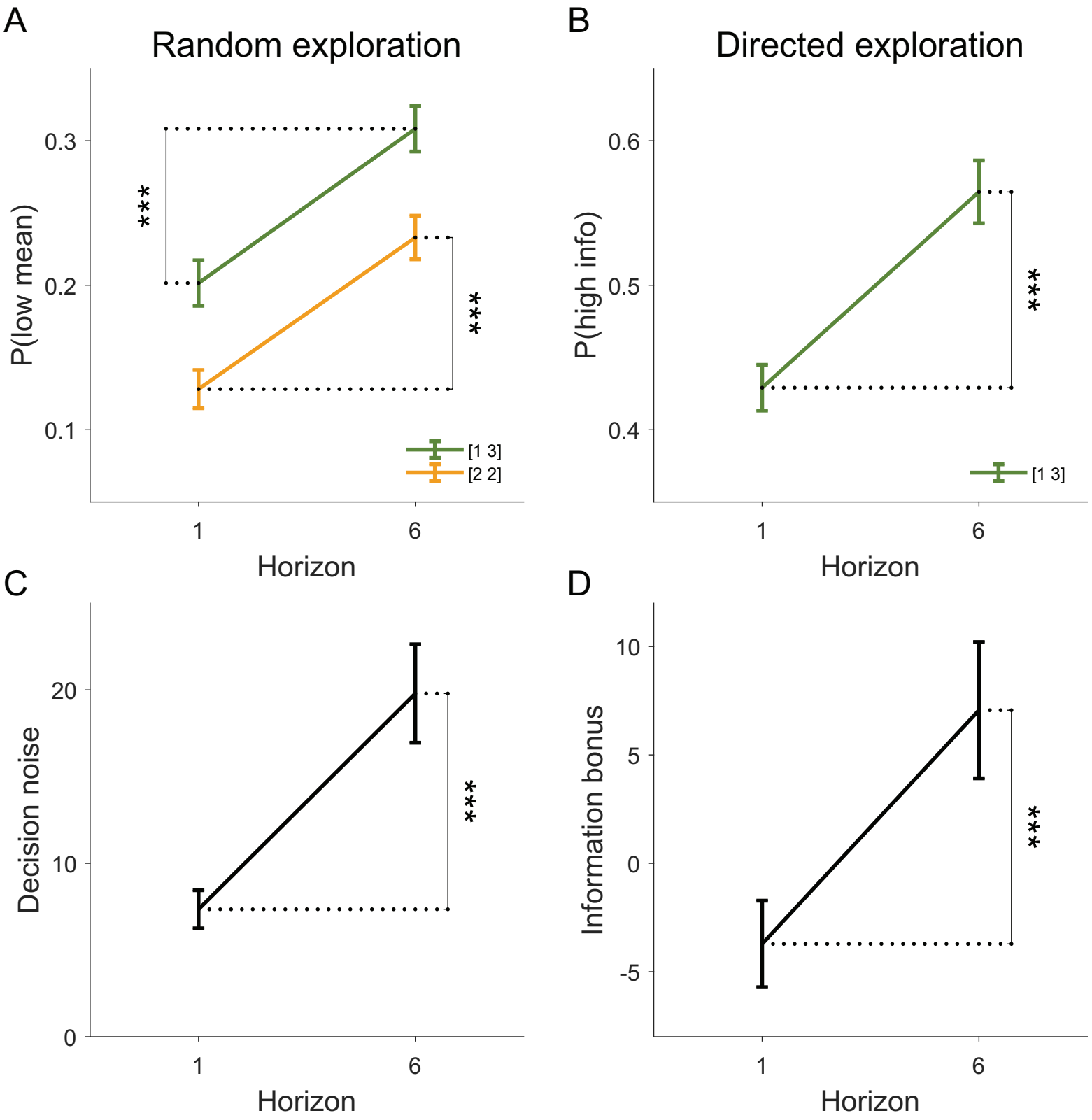
$n_{isgr}^{ran} \sim \text{Logistic}(0, \sigma_{is}^{ran})$

Observed choices

$\Delta Q_{isgr} \leftarrow \Delta R_{isg} + A_{is}\Delta I_{isg} + b_{is} + n_{isgr}^{ran} + n_{isg}^{det}$

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





Separating random and deterministic sources of computational noise in explore-exploit decisions

Siyu Wang[1] and Robert C. Wilson[1,2,3]

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

[2]Neuroscience and Physiological Sciences Graduate Interdisciplinary Program, University of Arizona, Tucson AZ, USA

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

April 22, 2025

Abstract

Human decision making is inherently variable. While this variability is often seen as a sign of suboptimal behavior, both theoretical work in machine learning and empirical human studies suggest 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 can encourage us to explore options we might otherwise ignore. In line with this idea, several studies have found evidence that people increase their behavioral variability when it is valuable to explore. A key question, however, is whether this variability in so-called '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. By modeling behavior in this task, we were able to estimate a lower bound on the amount of variability that is deterministically driven by the stimulus and an upper bound on the amount of variability that is random. Using this approach, we find evidence that at least 14% of the variability in random exploration in our studied task can be accounted for by deterministic processing of the stimulus. Conversely, this suggests that up to 86% of the variability is truly 'random', although it is still possible that this variability is driven by deterministic factors not related to the stimulus. Finally, our results suggest that both deterministic and random sources of variability change proportionally to each other as the value of exploration increases, suggesting that a common noise gating mechanism may be at play in random exploration.

Author Summary

Human decisions often seem random. Even simple decisions like what food to order at a restaurant can be difficult to predict ahead of time. This randomness in our decisions can be beneficial, effectively allowing us to explore new options. One outstanding question is where the randomness in our decisions comes from. Sometimes, our seemingly random decisions are driven by predictable external factors, like what the guest at the next table ordered could influence what we order. Other times, our decisions are not driven by external factors but are instead made by random thoughts within our brain. In this work, we developed a computational method that quantifies the extent to which the apparent randomness in our decisions can be explained by deterministic sources of variability in the external stimuli, or random variability unexplained by the stimuli. We found evidence that randomness in exploratory decisions can be explained by both random (up to 86%) and deterministic (more than 14%) sources of variability. Moreover, our results suggest that both sources of variability are adaptive, which enables humans to explore more when it is more beneficial to explore. The joint adaptation of random and deterministic noises also suggests a common noise-gating mechanism for exploration.

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 computational complexity, there is considerable interest in how humans and animals solve the explore-exploit dilemma in practice (Mehlhorn et al., 2015, Schulz and Gershman, 2019, Wilson et al., 2021).

One particularly effective strategy for solving the explore-exploit dilemma is choice randomization (Bridle, 1990, Thompson, 1933, Watkins, 1989), also known as random exploration. In this strategy, high value 'exploit' options are not always chosen and exploratory choices are sometimes made by chance. From a modeling perspective, random exploration works by adding 'decision noise' to the value of the options such that sub-optimal exploratory options can sometimes have a higher total score (i.e., value + noise) than the exploit option and get chosen. 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 decision noise when it is more beneficial to explore (Gershman, 2018, Wilson et al., 2014), as has also been suggested in computational models of animal behavior (Doya, 2002, Khamassi et al., 2013). In one of these tasks, known as the Horizon Task (Wilson et al., 2014), 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 will be more likely to explore the new restaurant. Using such a horizon manipulation it has been shown 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 research, however, is that it is difficult to tell whether what we have called 'decision noise' actually reflect a noise process. From a modeling perspective, decision noise as

defined in previous research essentially quantifies the extent to which behavior cannot be explained by a computational model. A missing deterministic component from the model could give rise to variability in behavior that might appear to be random noise. 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 the 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. To dissociate a missing deterministic component from a true random process, choice consistency between repeated decisions can be utilized to decompose behavioral variability into predictable deterministic components and unpredictable random components (Findling et al., 2019, Findling and Wyart, 2021, Wyart, 2018, Wyart and Koechlin, 2016).

In this paper, we investigate the extent to which the apparent randomness in random exploration can be explained by deterministic processing of the stimulus (which we refer to as 'deterministic noise') vs other processes, including deterministic processing that is unrelated to the stimuli as well as truly stochastic processes (which we refer to as 'random noise'). To distinguish between these two types of noise, we modify the Horizon Task (Wilson et al., 2014) to have people face the exact same explore-exploit choice twice. If the decision is a purely deterministic function of the stimulus (i.e., decision noise is purely deterministic noise), then people's choices should be identical for both decisions, since the stimulus is the same both times. Conversely, if the decision is a purely random function of the stimulus (i.e., decision noise is purely random noise), then people's choices will be different 50% of the time, since the random noise is different each time. In between these two extremes of purely deterministic and purely random drivers of behavioral variability, the extent to which people's decisions are consistent between the two decisions can be used to estimate the amount of deterministic and random noise.

In the following, we analyze behavior on the repeated decisions version of the Horizon Task in both a model-free and model-based manner. Our model-free analysis estimates the extent to which people's behavior is consistent across repeated versions of the same decision. By measuring how this choice consistency changes as a function of horizon, this model-free analysis offers qualitative insight into the extent to which behavioral variability is driven by deterministic vs random noise. Our model-based analysis uses a computational model of the explore-exploit decision in the Horizon Task that incorporates both noise processes. By fitting this model to the behavioral data, this model-based analysis allows us to quantify the relative size of the two sources of noise and how they change in the service of exploration.

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' in explore-exploit decisions (Figure 1). In this task, participants make a series of choices between two slot machines, or 'one-armed bandits', that pay out probabilistic rewards. They are asked to choose between the two bandits to maximize the total rewards. One bandit always has a higher mean payout than the other. Participants need to try each bandit a few times to learn about the distribution of payout from that bandit. 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.

The task is organized in games (Figure 1A). The mean payout of the two bandits are held fixed within a game and reset between games. Each game consists of either 5 or 10 trials. The first four trials of each game are 'forced-choice' trials. In the first four trials, participants are instructed about which bandit to choose, this allows us to manipulate what information from both bandits participants receive before they make their first free choice between the two bandits. From the 5th trial, participants make free choices between the two bandits. Participants have either 1 or 6 free choices to make.

The Horizon Task has two key features that together allow it to quantify explore-exploit behavior. The first of these features is the time horizon — the number of decisions participants will make in the future. By changing this horizon from short (1 free-choice trial) to long (6 free-choice trials), the Horizon Task allows us to control the relative value of exploration and exploitation. Just like the restaurant example in the introduction, when the horizon is short, participants should be more likely to exploit the option they believe to be best, because this leads to the highest payoff in the short term. Conversely, when the horizon is long, participants should be more likely to explore at first, because this allows them to gather information to make better choices later on. By contrasting behavior between short and long horizon conditions *on the very first free-choice trial*, when all else is equal, the Horizon Task allows us to quantify how behavior changes, when it is more valuable to explore.

The second key feature of the Horizon Task are the 4 forced-choice trials at the start of each game that allow us to control exactly what participants know about the two bandits before they make their choice. In these forced-choice trials, participants are instructed which of the bandits to play allowing us to control how much information they have about each of the options. The forced-choice trials are used to set up one

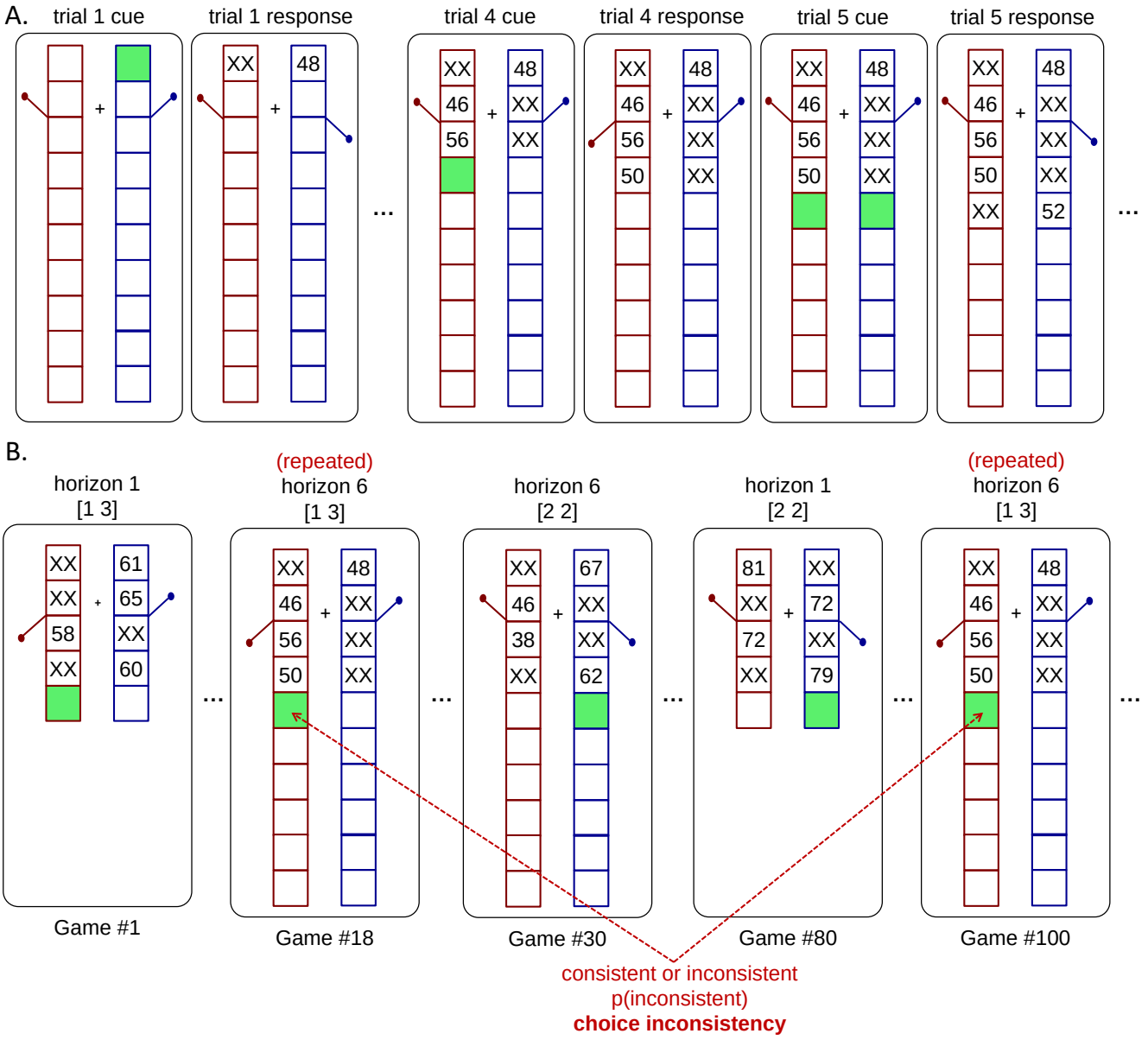


Figure 1: Schematic of the experiment. (A) Dynamics of an example horizon 6 game. Here the first four trials are forced trials in which participants are instructed which option to play. After the forced trials, participants are free to choose between the two options for the remainder of the game. (B) Example repeated games over the course of the experiment. On average, participants play more than 150 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 two information conditions: an 'unequal information' or [1 3] condition, in which participants play one bandit once and the other three times, and an 'equal information' or [2 2] condition, in which participants play both bandits twice.

Relative to the original Horizon Task, the key modification in this paper is to give people 'repeated games' (Figure 1B), in which they see the exact same set of forced-choice plays twice in two separate games separated by several minutes in time so as to avoid detection. By repeating the forced-choice 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, then they will make the same decision in both games and their choices will be consistent. Conversely, if their behavior is not driven by a deterministic function of the stimulus, then their choices on the repeated games will be inconsistent some fraction of the time. The extent to which participants' choices are consistent on the repeated versions of the games allow us to quantify the extent to which the variability in their behavior was driven by a deterministic process vs a random noise process.

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 using both a model-free and model-based approach (Wilson et al., 2014). In both analyses, we focus on just the first free-choice trial in each game, where the only thing that differs between the horizon conditions is the number of choices that participants will make in the future. Subsequent choices in Horizon 6 games were not analyzed.

Model-free analysis

In the model-free analysis, we quantify random and directed exploration using simple choice probabilities. Random exploration is quantified as the probability of choosing the option that has the lower average payout in the forced-choice plays in the equal, or [2 2], condition, $p(\text{low mean})$. The idea here is that, in the equal condition, the optimal strategy is to compute the mean payout for each bandit from the forced-choice plays and then always choose the option with the highest mean. When participants do not choose the option with the higher mean, the assumption is that this is due to some kind of 'decision noise', making the probability of choosing the low mean option a measure of behavioral variability. In this view, random

exploration corresponds to an increase in $p(\text{low mean})$ with horizon, which is exactly what we see in the data (Figure 2A; t(64) = 7.99, p $<$ 0.001).

Directed exploration is quantified as the probability of choosing the more informative option $p(\text{high info})$ in the unequal, or [1 3], condition. The more informative option is the option played once during the forced-choice plays as choosing this option gives relatively more information (doubling the number of samples from 1 to 2) than choosing the option played three times (only increasing the number of sample by a third, from 3 to 4). In this view, directed exporation corresponds to an increase in $p(\text{high info})$ with horizon, which is exactly what we see in the data (Figure 2B; t(64) = 6.92, p $<$ 0.001).

Model-based analysis

Another approach to understanding behavior in the Horizon Task is to use a computational model (Wilson et al., 2014). In this case, we model participants' choices on the first free-choice trial by assuming they make decisions by computing the difference in value (or utility) ΔQ between the right and left options, choosing right when $\Delta Q > 0$ and left otherwise. Specifically, we write

$$\Delta Q = \Delta R + A\Delta I + b + n \tag{1}$$

where, the experimentally controlled variables are $\Delta R = R_{right} - R_{left}$, the difference between the mean of rewards shown on the forced-choice trials, and ΔI, the difference in 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 in the [1 3] condition, and in [2 2] condition, ΔI is 0.

Here, n denotes decision noise, which, in this version of the model is a combination of deterministic and random noise. n is assumed to come from a logistic distribution with mean 0 and standard deviations σ.

The free parameters of this model are: the information bonus A, which controls the level of directed exploration; the noise standard deviation, σ, which controls the level of random exploration, and the spatial bias, b, which determines the extent to which participants prefer the option on the right. These free parameters are fit separately for each participant in each horizon condition, allowing us to test whether directed and random exploration increase with horizon. Consistent with previous research, we find that this is indeed the case (Figure 2C; t(64) = 5.35, p $<$ 0.001. Figure 2D; t(64) = 3.54, p $<$ 0.001).

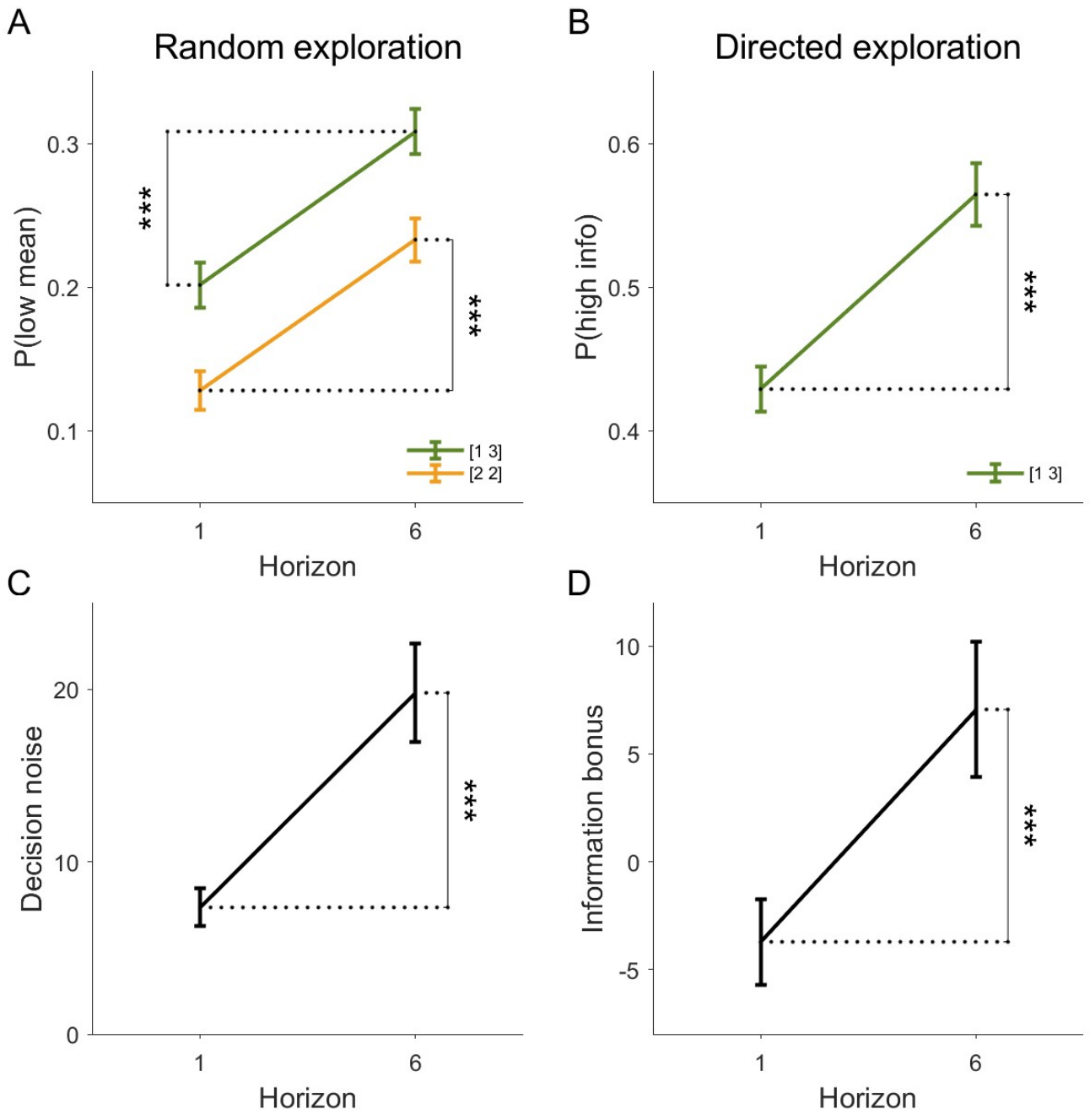


Figure 2: Replication of previous findings that people use both random and directed exploration in this task. (A) model-free measure of behavioral variability, $p(\text{low mean})$, increases with horizon. (B) model-free measure of information seeking, $p(\text{high info})$, increases with horizon. (C) model-based measure of behavioral variability, decision noise σ, increases with horizon. (D) model-based measure of information seeking, information bonus A, increases with horizon.

Taken together, our model-free and model-based analyses agree with previous findings showing in-

creased behavioral variability and increased information seeking in the long horizon condition, consistent with humans using random and directed exploration (Figure 2, Supplementary Figure S1). However, for random exploration, this previous analysis cannot distinguish between deterministic and random sources of noise. For this we analyze the extent to which people's choices are consistent on the repeated games.

Model-free analysis of repeated games suggests that random exploration involves 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 entirely driven by a deterministic process and is also driven by random noise, participants' choices should be more inconsistent across the repetitions of the game. Moreover, if decision noise is purely random noise, meaning there is no unobserved deterministic process, we will show that we can actually predict the expected level of choice inconsistencies across repetitions of games by accounting for the known deterministic processes and assuming that the random noise process is independent in repetitions of the game.

To quantify choice inconsistency we computed the frequency with which participants made different responses for pairs of repeated games (Figure 3, Supplementary Figure S2). Using this measure we found that participants made inconsistent choices in both the unequal ([1 3]) and equal ([2 2]) information conditions ($p(\text{inconsistent}) > 0$), 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 (For [1 3] condition, t(64) = 5.41, $p < 0.001$; for [2 2] condition, t(64) = 6.26, $p < 0.001$), suggesting that at least some of the horizon dependent noise is not a deterministic function of the stimulus, but rather random noise.

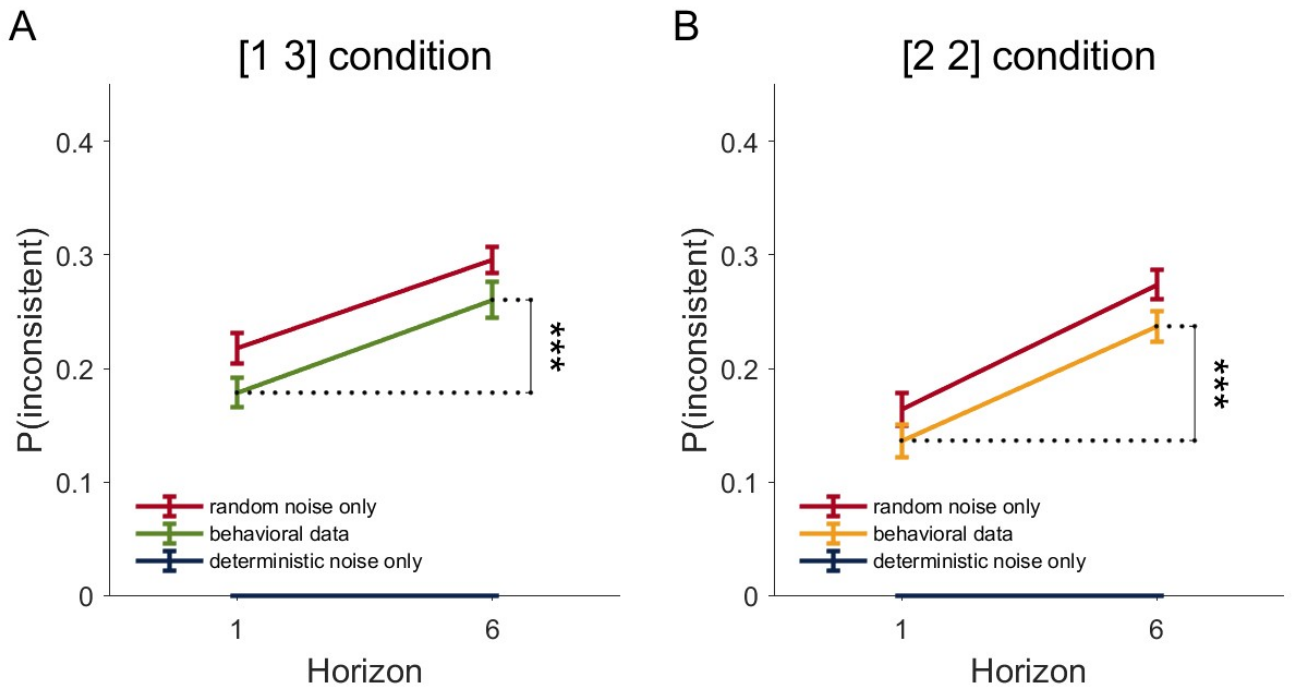


Figure 3: Model-free analysis suggests that both deterministic and random noise contribute to the choice variability in random exploration. For both the [1 3] (A) and [2 2] (B) condition, people show greater choice inconsistency in horizon 6 than horizon 1. However, the extent to which their choices 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. Since repeated decisions mean that participants either choose the low-mean option twice, or choose the high-mean option twice, we could predict the choice inconsistency, $p(\text{inconsistent})$, based on the probability of choosing the low mean option, $p(\text{low mean})$, as

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

Furthermore, to account for the fact that $p(\text{low mean})$ is a function of reward difference ΔR between

the two bandits and the information condition I, we estimated the conditional probability:

$$p(\text{inconsistent}|\Delta R, I) = 2p(\text{low mean}|\Delta R, I)(1 - p(\text{low mean}|\Delta R, I))$$

Then based on the likelihood that each condition (ΔR and I) occurs in the task $\rho(\Delta R, I)$, we have

$$p(\text{inconsistent}) = \sum_{\Delta R, I} \rho(\Delta R, I) p(\text{inconsistent}|\Delta R, I)$$

As shown in Figure 3, people's behavior falls in between the pure deterministic noise prediction and the pure random noise prediction. Specifically, behavior is different from the pure random noise prediction in both the [1 3] condition (t(64) = 4.83, $p < 0.001$ for horizon 1, t(64) = 3.12 $p = 0.003$ for horizon 6) and the [2 2] condition (t(64) = 3.92, $p < 0.001$ for horizon 1, t(64) = 3.71, $p < 0.001$ for horizon 6). Likewise, behavior is different from pure deterministic noise prediction in both the [1 3] condition (t(64) = 13.72, $p < 0.001$ for horizon 1, t(64) = 16.71, $p < 0.001$ for horizon 6) and the [2 2] condition (t(64) = 9.55, $p < 0.001$ for horizon 1, t(64) = 17.93, $p < 0.001$ for horizon 6). As a negative control of our method for estimating $p(\text{inconsistent})$ for purely random noise, we simulated choices using a decision model that only includes random noise (Equation 2), and found that $p(\text{inconsistent})$ in this simulated data is not different from our pure random noise prediction in all horizon and uncertainty conditions ($p > 0.05$, Supplementary Figure S3). Together, our results suggest that both random noise and deterministic noise contribute to the choice variability in random exploration. However, the relative contribution from each of these types of noise, as well as how each type of noise changes with horizon, are difficult to discern.

Model-based analysis provides a lower-bound estimate of deterministic noise and an upper-bound estimate of random noise

To more precisely quantify the contribution of deterministic noise and random 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 (Equation 1) 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'). 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

To model participants' choices on the first free-choice trial, we use a modified version of Equation 1.

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

where, as before ΔR, is the the difference in mean rewards shown on the forced-choice trials, ΔI, is the difference in information, A is the information bonus, and b is the spatial bias. New in Equation 2 are the terms n_{det} and n_{ran}. 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 repeated plays and changes every game. n_{det} and n_{ran} are assumed to come from logistic distributions with mean 0, and standard deviations σ_{det} and σ_{ran}.

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 $n^i_{ran} \neq n^j_{ran}$. While an increase in either random or deterministic noises could lead to higher p(low mean), an increase in random noise predicts higher p(inconsistent) while an increase in deterministic noise predicts lower p(inconsistent).

We used hierarchical Bayesian analysis to fit the parameters of the model (see Figure 8 for a 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.

We also fit a series of reduced and alternative models to the data. This includes reduced models that assume only deterministic or random noises. We also fit models in which the standard deviation of random and deterministic noises σ_{det} and σ_{ran} are estimated separately for [1 3] and [2 2] information conditions. Lastly, we fit a model with an alternative definition of ΔI that ΔI is defined to be difference between the variances of rewards shown on the forced-choice trials. Results on these model variants are presented in the Supplementary Materials.

Model validation

To be sure that our fit parameter values were meaningful and to understand the limits of our model, we evaluated our model extensively using simulated data. This allowed us to quantify whether deterministic

and random noise can be identified under ideal conditions where the behavior is generated by the model with known parameters. Full details are presented in the Supplementary Materials section 2.

In this section we focus on our results for parameter recovery (Wilson and Collins, 2019). In a parameter recover analysis, behavioral data is simulated by the model with known parameters and then this simulated behavioral data is fit with the model to quantify the extent to which fit parameters match the input simulated parameters — that is, whether the simulated parameters can be recovered.

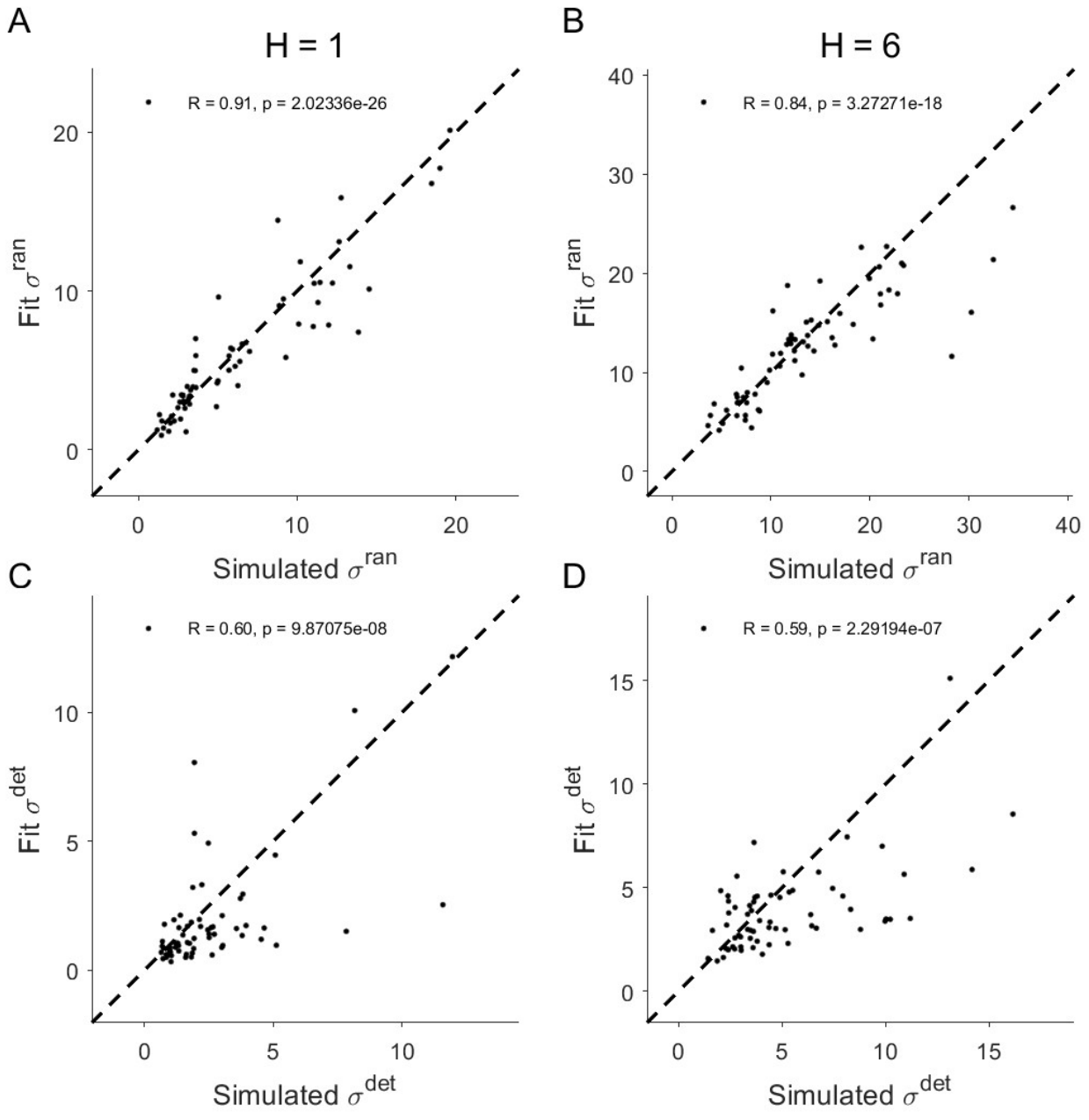


Figure 4: Parameter recovery analysis for random (A,B) and deterministic (C,D) noises in the two horizons.

Parameter recover in this task was good for this model (Figure 4, Supplementary Figure S7, S8), with fit values for σ_{ran} and σ_{det} showing strong correlations with their simulated values in both horizon (H) conditions (For σ_{ran}, $R = 0.91(\text{H} = 1)$ and $0.84(\text{H} = 6), p < 0.001$, For σ_{det}, $R = 0.60(\text{H} = 1)$ and $0.59(\text{H} = 6), p < 0.001$). However, while the relationship was near perfect for random noise ($\frac{\text{Recovered } \sigma_{ran}}{\text{Simulated } \sigma_{ran}} = 1.01$), there was a systematic bias to underestimate the level of deterministic noise by about 32% ($\frac{\text{Recovered } \sigma_{det}}{\text{Simulated } \sigma_{det}} = 0.68$). Despite this underestimation of deterministic noise in both horizon conditions, the difference in deterministic noise between horizons is much better captured (see Supplementary Materials section 2.2). This is because the underestimation of deterministic noise is partially canceled out when the difference is taken between horizon conditions. In addition, we see better parameter recovery for random noise than deterministic noise. This is likely because we effectively have half as many trials for deterministic noise. In particular, while we generate two samples of random noise for each repeated game pair, we only generate one sample of deterministic noise, which by definition is the same in both of the repeated games.

In addition to the conventional subject-level parameter recovery analysis presented here, we also performed parameter recovery analysis that examined how faithful the full posterior distribution of group-level parameters can be recovered in simulated data (Supplementary Figure S5, S6, S9, S10). Qualitatively, we also showed that our way of modeling deterministic noise is capable of capturing known deterministic processes intentionally omitted from the full model (Supplementary Figure S4). Full details of these additional analysis are presented in the Supplementary Materials.

Overall, we were able to detect both deterministic and random noises using our model. Because random noise is modeled as non-stimulus-driven noise, it can reflect both true stochastic random noise and possible deterministic noises which do not depend on the stimuli. Thus conceptually our random noise estimate provides an upper bound for the true 'random noise' induced by intrinsic stochastic processes in the brain. Thus, our model provides a lower bound for deterministic noise and an upper bound for random noise.

Model-based results

Posterior distributions over the group-level means of the deterministic and random noise standard deviation σ_{det} and σ_{ran} are shown in Figure 5 and Supplementary Figure S11. Consistent with our model-free results, we see that both random and deterministic noise are non-zero. Numerically, random noise is about 2-3 times larger than the deterministic noise. By computing the posterior distribution of $\sigma_{det}^2/(\sigma_{det}^2 + \sigma_{ran}^2)$, our data suggests that 14.25% of the variability in random exploration is accounted for by deterministic

noise ([4.90%, 28.81%], 95% CI). In addition, we find that both random and deterministic noise increase with horizon. This increase was larger for random noise (mean = 7.13, 100% of samples showed an increase in random noise with horizon) than deterministic noise (mean = 2.59, 98.64% of samples showed an increase in deterministic noise with horizon). But intriguingly, the relative increase in both types of noise was similar (Figure 6). That is, when we compute the relative increase in deterministic noise with horizon, $\sigma^{det}_{horizon6}/\sigma^{det}_{horizon1}$, it is very similar to the relative increase in random noise with horizon $\sigma^{ran}_{horizon6}/\sigma^{ran}_{horizon1}$.

Similar results are found in other variants of our model. Results for a model that estimates random and deterministic noises separately for [1 3] and [2 2] conditions, and a model with an alternative definition of information bonus dI, can be found in the Supplemental Materials (Supplemental Figure S12, S13).

To ensure that the joint increase of random and deterministic noises is genuine and not an artifact from the fitting procedure, we computed the correlation between ground-truth values of random noise, and best-fitting values of deterministic noise (and vice versa), and they do not correlate (Supplementary Figure S14). Furthermore, we simulated data from a series of reduced models with known random and deterministic noise values in which either random or deterministic noise does not change with Horizon, and fit our model to the simulated data. Our model detects and only detects a change in random/deterministic noise with horizon, when the change is present in the model that simulates the data (Supplementary Figure S15, S16).

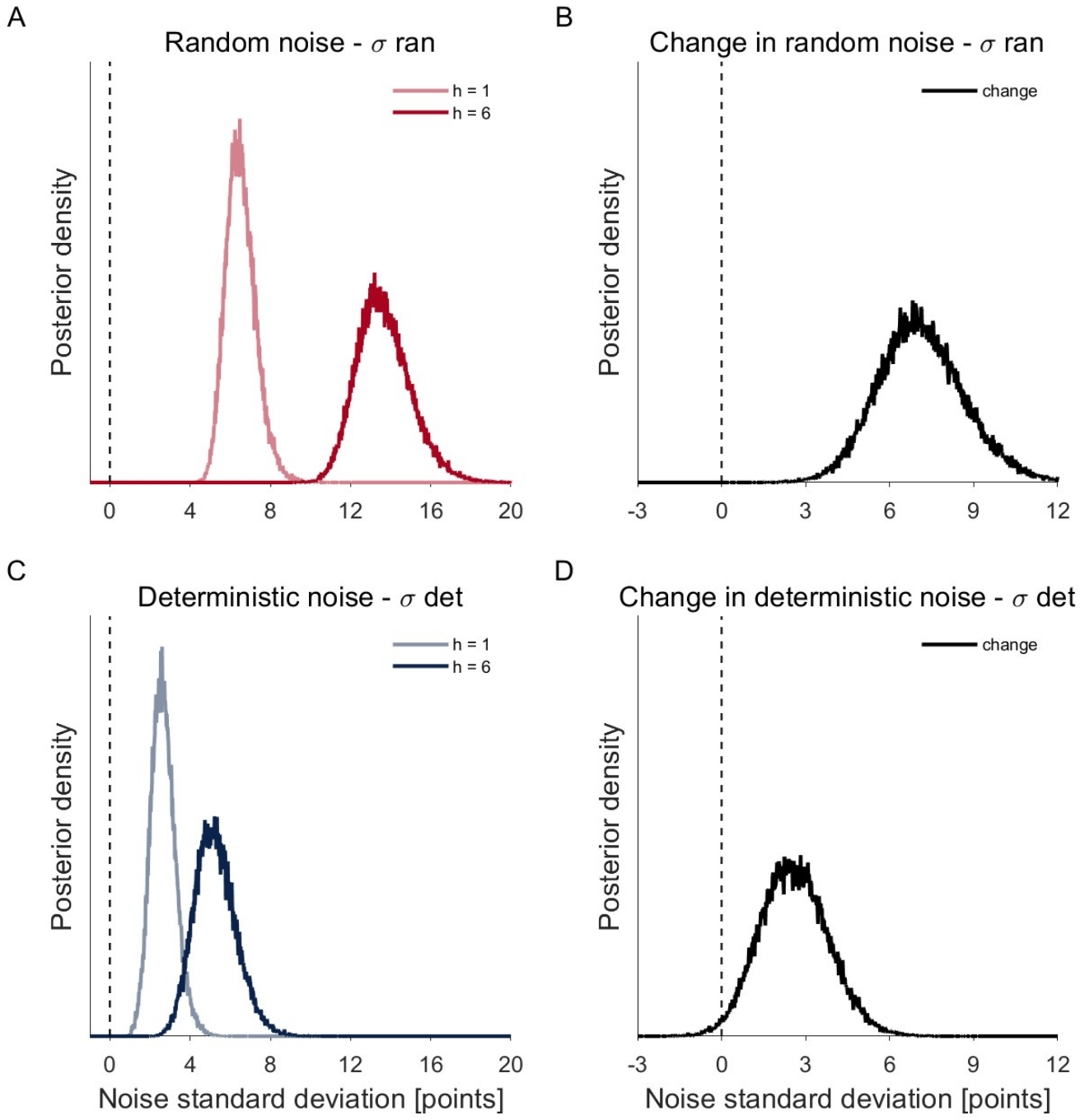


Figure 5: Model based analysis showing the posterior distributions over the group-level mean of the standard deviations of random and deterministic noise. Both random (A, B) and deterministic (C,D) noises are nonzero (A, C) and increase with horizon (B, D).

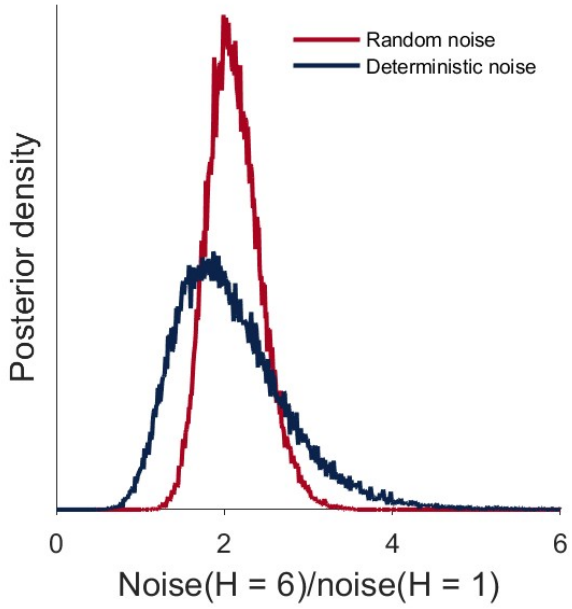


Figure 6: Model based analysis showing the posterior distributions over the ratio of the group-level mean of the standard deviations of random and deterministic noise between horizon 6 and horizon 1 respectivelly. The ratios in the standard deviations of noises between horizon 6 and horizon 1 are similar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

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 (Palminteri et al., 2017, 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) and for p(inconsistent), but the model slightly overestimate p(low mean) in [2 2] conditions. This has to do with the skewness of the subject-level posterior distribution (see Supplemental Materials).

As a control, we also applied posterior predictive checks on alternative models that consider only deterministic or only random noise, and these reduced models fail to capture all qualitative patterns (Supplementary Figure S17, S18). Full details of this analysis can be found in Supplementary Materials section 3.5.

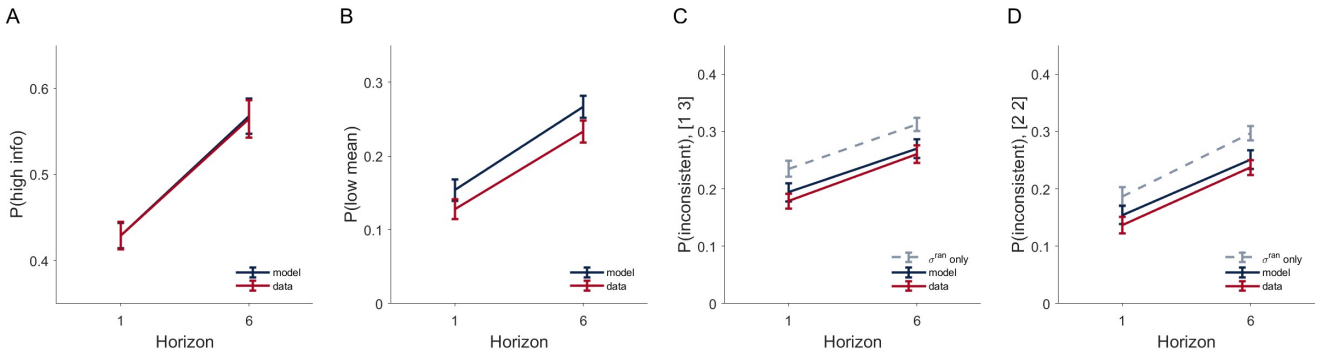


Figure 7: Our model accounts for all qualitative patterns of the data, namely, (A) p(high info) and (B) p(low mean) increase as a function of horizon, p(inconsistent) increases as a function of horizon for both [1 3] (C) and [2 2] (D) conditions and it lies between the pure random and pure deterministic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

Discussion

In this paper, we investigated whether random exploration is really random or whether it is 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 around 14% of the overall variability in people's behavior.

One interpretation for this low level of deterministic noise is that most of the variability in random exploration is truly random. Such a random noise interpretation, would be consistent with recent work showing that variability in perceptual decisions may be driven by imperfections in mental inference (Drugowitsch et al., 2016). In this view, apparently random behavior is not due to sensory processing or response selection, but to suboptimal computations in the brain. Although suboptimal inference is different from simply adding random noise to neural circuitry(Beck et al., 2012), as long as the suboptimality in neural computation is not a deterministic function of the stimuli, it is a form of random noise in our definition. Indeed, a strong interpretation of this hypothesis would suggest that randomness in explore-exploit behavior is

due to imperfect inference about the correct course of action. 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 (Wilson et al., 2020).

Although the random noise interpretation is theoretically appealing, our approach, while an improvement on previous methods, is not without limitations. Most important is that our measure of 'random' noise is only an upper bound on the true level of randomness and that, in principle, the random decision noise could be lower. Specifically, in our model, what we labeled random noise was really 'non-stimulus-driven variability'. While this non-stimulus-driven variability could be driven by truly random stochastic processes, it could also be driven by deterministic processing that is unrelated to the stimuli in the task. For example, such deterministic noise could be driven by differences in where people look, or for how long they look, or by whether they were fidgeting or scratching their nose (Musall et al., 2019). Another limitation is that deterministic noise is defined based on the stimulus within a game, between-game deterministic strategies and stimuli from previous games (e.g., memory of previously seen game) were also treated as random noises in our model, although our model could be potentially extended by considering deterministic noise over a sequence of stimuli across games (Wyart, 2018, Wyart and Koechlin, 2016). In addition to this conceptual limitation in measuring deterministic noise, parameter recovery simulations suggest that our estimation method also slightly underestimates deterministic noise (see Figure 4, Supplementary Figure S5, S6). As a result, from both a conceptual and methodological perspective, it is possible that the remaining 86% of the decision noise that is not stimulus-driven noise, could be deterministic.

Like the random noise account, the deterministic noise account is also in line with previous work in which neural variability can be accounted for by fluctuations in sensory inputs. For example, MT neurons were shown to have a reproducible temporal modulation in response to a fixed random motion stimulus (Bair and Koch, 1996). In other words, 'irrelevant' features in the stimuli are represented in a reliable way in the brain that could drive downstream choices in a predictable way.

Regardless of whether we interpret the noise as random or deterministic, a key finding in this paper is that both types of noise change with horizon. Such a horizon increase is a hallmark of an exploratory process and suggests that the modulation of deterministic and random processes may underlie random exploration. Moreover, the fact that the horizon change in the two types of noise are proportional to each other (Figure 6) suggests a possible mechanism for random exploration: a reduction in the strength with which reward drives the choice.

We show first that a change in noise is mathematically equivalent to a change in reward signal strength

in our decision model (see also Cinotti et al. (2019)).. To show how a change in reward processing could affect random and deterministic noise, consider the simple decision model we introduced in Equation 2. In this model, choice is determined by the sign of the difference in utility ΔQ between the two options, where

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

Now imagine a case where the reward signal is scaled by a factor β. In this case, ΔQ becomes

$$\Delta Q = \beta\Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{4}$$

Because the choice only depends on the sign of ΔQ, scaling ΔQ by a factor of $1/\beta$ will not change the behavior of the model. Thus, if we divide both sides of the above equation by β we get

$$\Delta Q/\beta = \Delta R + A\Delta I/\beta + b/\beta + n_{det}/\beta + n_{ran}/\beta \tag{5}$$

which is equivalent to a scaling of both deterministic and random noise by the same factor $1/\beta$. Thus, one interpretation of our result that both deterministic and random noise change across horizons with the same ratio, is that this reflects a change in reward processing. That is, the reward signal is reduced in the longer horizon condition (smaller β in horizon 6 than horizon 1).

Such a reduction in the strength of reward coding in exploration, is consistent with our recent work using a drift diffusion model (DDM) to model explore-exploit decisions (Feng et al., 2021). In the drift diffusion model, changes in behavioral variability can be driven by changes in the decision threshold (smaller threshold = more noise) or changes in the signal-to-noise ratio with which reward is encoded (lower SNR = more noise). By fitting both choices and response times, we were able to distinguish between these two accounts showing that the majority of the horizon-change in variability was driven by changes in SNR and not threshold. However, this model could not determine whether the changes in SNR were driven by signal or noise. By showing that the change in deterministic and random noise have approximately the same ratio, the present work suggests that this SNR change is driven by changes in reward-signal processing, not noise. Of course, to truly see whether changes in signal or noise are driving random exploration will require more direct measurements of neural processing such as with neuroimaging and electrophysiology (Costa et al., 2019, Ebitz et al., 2018, Hogeveen et al., 2022, Tomov et al., 2020)

Materials and Methods

Ethics statement

Human subject protocols were approved by the University of Arizona institutional review board (IRB # 1411567117). Written informed consent was given by all participants prior to participating in the study.

Participants

80 participants (ages 18-25, 37 male, 43 female) from the University of Arizona undergraduate subject pool participated in the experiment. 15 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 by making a completely random choice between the two options and excluded participants with a likelihood greater than 0.1%, in other words, participants who didn't show an accuracy significant above chance with $p < 0.001$ were excluded in the analysis. This left 65 for the main analysis. Note that including the 15 badly performing subjects did not change the main results (Supplementary Figures S1, S2, S11)

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 153.7 games (minimum = 90 games, maximum = 192 games) and the whole task lasted between 12.37 and 32.15 minutes (mean 22.78 minutes). Participants played an average of 65.3 repeated pairs of games (minimum = 30 repeated pairs, maximum = 79 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 'forced-choice' 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 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).

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 deterministic noise from 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

$$\Delta Q = \Delta R + A\Delta I + b + n_{det} + n_{ran} \tag{6}$$

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 of 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. 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} and n_{ran} are deterministic noise and random noise respectively. 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 (Table 1): 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}. 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 (Figure 8). 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.

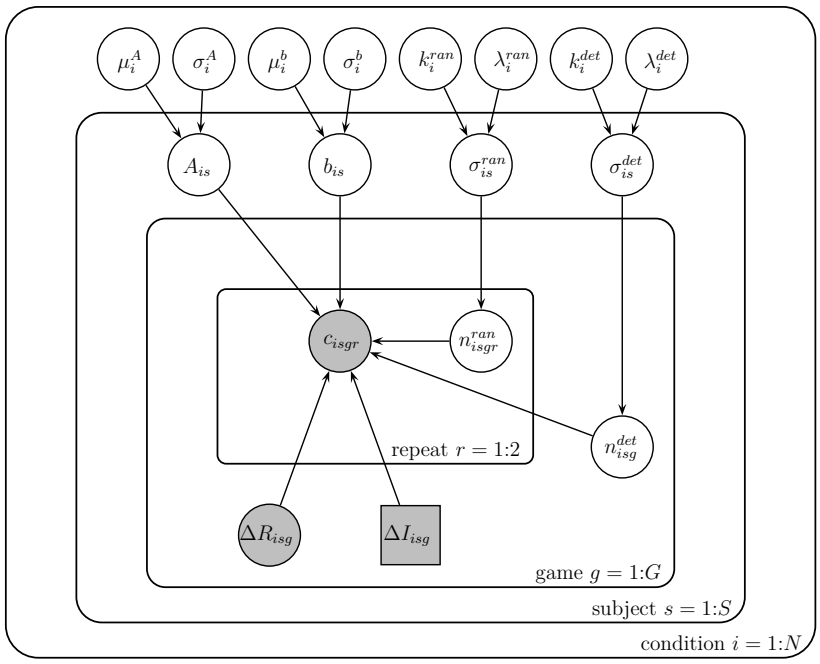
Parameter	Prior	Hyperparameters	Hyperpriors
information bonus, A_{is}	$A_{is} \sim$ Gaussian(μ_i^A, σ_i^A)	$\theta_i^A = (\mu_i^A, \sigma_i^A)$	$\mu_i^A \sim$ Gaussian(0, 100) $\sigma_i^A \sim$ Exponential(0.01)
spatial bias, b_{is}	$b_{is} \sim$ Gaussian(μ_i^b, σ_i^b)	$\theta_i^b = (\mu_i^b, \sigma_i^b)$	$\mu_i^b \sim$ Gaussian(0, 100) $\sigma_i^b \sim$ Exponential(0.01)
deviation of deterministic noise, σ_{isg}^{det}	$\sigma_{is}^{det} \sim$ 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}	$\sigma_{is}^{ran} \sim$ 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 10 independent Markov chains to generate 50000 samples from the posterior distribution over parameters (5000 samples per chain). Each chain had a burn in period of 5000 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

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

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

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

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

Subject specific parameters

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

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

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

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

Deterministic noise for repeated game

$n_{isg}^{det} \sim \text{Logistic}(0, \sigma_{is}^{det})$

Random noise for each game

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

Observed choices

$\Delta Q_{isgr} \leftarrow \Delta R_{isg} + A_{is}\Delta I_{isg} + b_{is} + n_{isgr}^{ran} + n_{isg}^{det}$

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

Figure 8: Schematic of the hierarchical Bayesian model using notation of Lee and Wagenmakers (2014b)

Data and code

Behavioral data as well as MATLAB codes to recreate the main figures from this paper will be made available upon publication at `https://github.com/wangxsiyu/RW_RandomDeterministicNoise.git`

Supporting information

S1 Fig. Replication of previous findings with data from all participants (i.e. no exclusions). (A) model-free measure of behavioral variability, $p(\text{low mean})$, increases with horizon. (B) model-free measure of information seeking, $p(\text{high info})$, increases with horizon. (C) model-based measure of behavioral variability, decision noise σ, increases with horizon. (D) model-based measure of information seeking, information bonus A, increases with horizon.

S2 Fig. Model-free analysis with data from all participants (i.e. no exclusions) 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..

S3 Fig. Model-free analysis with simulated choices from a model that has only random noise validates our prediction of p(inconsistent) for pure random noise. The extent to which simulated choices are inconsistent completely overlaps with our pure random noise prediction($p > 0.05$). This suggests that when choice inconsistency lies below the pure random noise prediction indeed provides evidence that deterministic noise exists in random exploration (Figure 3).

S4 Fig. Deterministic noise can recover known deterministic processes that's intentionally omitted by the model. In the reduced model where the deterministic effect of uncertainty condition is omitted from the model, deterministic noise is higher compared to the full model that accounts for the effect of uncertainty. Random noise remains unchanged between the two models.

S5 Fig. Hyperprior recovery. Parameter recovery over the posterior distribution of random and deterministic noise standard deviations σ_{det} and σ_{ran}. Solid lines are true posterior used to simulate choices. Lighter color shades represent the re-fitted posterior to the simulated choices. Our model fitting procedure faithfully recovers the non-stimulus-driven random noise (A, B), but systematically underestimates deterministic noise in both horizons (D, E). The horizon differences in random noise is also faithfully recovered (C). The horizon differences in deterministic noise is also underestimated but not significant (F).

S6 Fig. Frequentist coverage analysis. Parameter recovery over the mean estimates of random and deterministic noise standard deviations σ_{det} and σ_{ran}. Solid lines are true posterior used to simulate choices, dashed black line is the mean of the true posterior. Histograms represent the mean estimates of the respective parameters in the refitting to the simulated data. (A) and (B) are random noise at H = 1 and H = 6, respectively. (C) is the random noise differences between horizons. (D) and (E) are deterministic noise at H = 1 and H = 6, respectively. (F) is the deterministic noise differences between horizons.

S7 Fig. Parameter recovery. Parameter recovery over the subject-level means of information bonus, A, spatial bias, b, random noise standard deviation, σ_{ran}, and deterministic noise standard deviation, σ_{det}, for horizon 1 (left column) and horizon 6 (right column) games.

S8 Fig. Parameter recovery (200 repetitions). Same as Figure S7, except that the recovered parameters were averaged across 200 repetitions and then compared to the original parameters.

S9 Fig. Parameter recovery with 0 random noise or 0 deterministic noise. Parameter recovery over the posterior of random noise standard deviation, σ_{ran}, and deterministic noise standard deviation, σ_{det}, for purely random noise (top row) and purely deterministic noise (bottom row) games.

S10 Fig. Parameter recovery on arbitrary combinations of random and deterministic noises. A. Recovered posterior distributions of random noise. B. Recovered posterior distributions of deterministic noise. For both A and B, from the top row to the bottom row, the true noise standard deviation that is used in the simulations go from 0 to 10. The y limit of each panel is 4 (+/- 2 from the true value). Our model did a relatively good job in recovering all combinations of deterministic and random noises.

S11 Fig. Model based analysis with data from all participants (i.e. no exclusions) 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.

S12 Fig. Model based analysis from a model that estimates random and deterministic noises separately for [1 3] and [2 2] conditions. The posterior distributions over the group-level mean of the standard deviations of random and deterministic noise. Both random (A, E) and deterministic (C,G) noises are nonzero (A, C, E, G) and change with horizon (B, D, F, H). However, random noise has both a greater magnitude overall (A, E) and a greater change with horizon (B, F) than deterministic noise. Moreover, both random and deterministic noises have a greater magnitude in [1 3] compared to [2 2] conditions.

S13 Fig. Model based analysis from a model that uses variance differences as dI. 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.

S14 Fig. Parameter recovery for shuffled data. To show that the joint increase of random and deterministic sources of noise is not caused by a limitation of the fitting procedure, we calculated the correlation between ground-truth values of random noise, and best-fitting values of deterministic noise (and vice versa). Ground-truth values are shuffled best-fit parameters. As expected, ground-truth random values do not correlate with recovered deterministic noises, showing that the increase of deterministic noise with horizon is genuine and not a by-product of increase of random noise, and vice versa.

S15 Fig. Model based analysis with reduced models. Each row is one model. These models varied in whether deterministic σ^{det} and random noise σ^{ran} are present or not and whether either types of noise is dependent on horizon (subscript denotes the dependence on horizon).

S16 Fig. Hyperprior recovery of reduced models. Our model qualitatively captures whether deterministic and random noise are present or not and whether either types of noise is dependent on horizon. A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S17 Fig. Posterior checks for reduced models Model comparison. A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S18 Fig. Posterior checks for reduced models (maximal likelihood estimation) Model comparison (using maximal likelihood estimation). A-D. both deterministic and random noise are horizon dependent, E-H. only random noise is horizon dependent, I-L. only deterministic noise is horizon dependent, M-P. neither random nor deterministic noise is horizon dependent, Q-T. only deterministic noise is assumed to be present, U-X. only random noise is assumed to be present.

S1 Table. Variants of the model.

S1 File. Supplemental Materials containing additional analyses.

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.

Wyeth Bair and Christof Koch. Temporal Precision of Spike Trains in Extrastriate Cortex of the Behaving Macaque Monkey. *Neural Computation*, 8(6):1185–1202, 1996. ISSN 08997667. doi: 10.1162/neco.1996.8.6.1185.

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

J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–9, 2012. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2012.03.016. URL `https://www.ncbi.nlm.nih.gov/pubmed/22500627`. Beck, Jeffrey M Ma, Wei Ji Pitkow, Xaq Latham, Peter E Pouget, Alexandre eng R01 EY020958/EY/NEI NIH HHS/ R01EY020958/EY/NEI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2012/04/17 Neuron. 2012 Apr 12;74(1):30-9. doi: 10.1016/j.neuron.2012.03.016.

D. H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

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`.

François Cinotti, Virginie Fresno, Nassim Aklil, Etienne Coutureau, Benoît Girard, Alain R. Marchand, and Mehdi Khamassi. Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, 9(1):6770, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-43245-z. URL https://doi.org/10.1038/s41598-019-43245-z.

V. D. Costa, A. R. Mitz, and B. B. Averbeck. Subcortical substrates of explore-exploit decisions in primates. *Neuron*, 103(3):533–545 e5, 2019. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2019.05.017. URL https://www.ncbi.nlm.nih.gov/pubmed/31196672. Costa, Vincent D Mitz, Andrew R Averbeck, Bruno B eng Z99 MH999999/ImNIH/Intramural NIH HHS/ ZIA MH002928/ImNIH/Intramural NIH HHS/ ZIA MH002928-09/ImNIH/Intramural NIH HHS/ Research Support, N.I.H., Intramural 2019/06/15 Neuron. 2019 Aug 7;103(3):533-545.e5. doi: 10.1016/j.neuron.2019.05.017. Epub 2019 Jun 10.

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.

Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4):495–506, 2002. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(02)00044-8. URL https://www.sciencedirect.com/science/article/pii/S0893608002000448.

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.

R. B. Ebitz, E. Albarran, and T. Moore. Exploration disrupts choice-predictive signals and alters dynamics in prefrontal cortex. *Neuron*, 97(2):475, 2018. ISSN 1097-4199 (Electronic) 0896-6273 (Linking). doi: 10.1016/j.neuron.2018.01.011. URL https://www.ncbi.nlm.nih.gov/pubmed/29346756. Ebitz, R Becket Albarran, Eddy Moore, Tirin eng Published Erratum 2018/01/19 Neuron. 2018 Jan 17;97(2):475. doi: 10.1016/j.neuron.2018.01.011.

Samuel F. Feng, Siyu Wang, Sylvia Zarnescu, and Robert C. Wilson. The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, 11(1): 3077, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82530-8. URL https://doi.org/10.1038/s41598-021-82530-8.

C. Findling, V. Skvortsova, R. Dromnelle, S. Palminteri, and V. Wyart. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nat Neurosci*, 22(12):2066–2077, 2019. ISSN 1097-6256. doi: 10.1038/s41593-019-0518-9. 1546-1726 Findling, Charles Skvortsova, Vasilisa Dromnelle, Rémi Palminteri, Stefano Orcid: 0000-0001-5768-6646 Wyart, Valentin Orcid: 0000-0001-6522-7837 Journal Article Research Support, Non-U.S. Gov't United States 2019/10/30 Nat Neurosci. 2019 Dec;22(12):2066-2077. doi: 10.1038/s41593-019-0518-9. Epub 2019 Oct 28.

Charles Findling and Valentin Wyart. Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, 38:124–132, 2021. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2021.02.018. URL `https://www.sciencedirect.com/science/article/pii/S2352154621000401`. Computational cognitive neuroscience.

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

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

J. Hogeveen, T. S. Mullins, J. D. Romero, E. Eversole, K. Rogge-Obando, A. R. Mayer, and V. D. Costa. The neurocomputational bases of explore-exploit decision-making. *Neuron*, 110(11):1869–1879 e5, 2022. ISSN 1097-4199 (Electronic) 0896-6273 (Print) 0896-6273 (Linking). doi: 10.1016/j.neuron.2022.03.014. URL `https://www.ncbi.nlm.nih.gov/pubmed/35390278`. Hogeveen, Jeremy Mullins, Teagan S Romero, John D Eversole, Elizabeth Rogge-Obando, Kimberly Mayer, Andrew R Costa, Vincent D eng P51 OD011092/OD/NIH HHS/ P30 GM122734/GM/NIGMS NIH HHS/ ZIA MH002929/ImNIH/Intramural NIH HHS/ P20 GM109089/GM/NIGMS NIH HHS/ ZIA MH002928/ImNIH/Intramural NIH HHS/ R01 MH125824/MH/NIMH NIH HHS/ Research Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, U.S. Gov't, Non-P.H.S. 2022/04/08 Neuron. 2022 Jun 1;110(11):1869-1879.e5. doi: 10.1016/j.neuron.2022.03.014. Epub 2022 Apr 6.

Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Chapter 22 - medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In V.S. Chandrasekhar Pammi and Narayanan Srinivasan, editors, *Decision Making*, volume 202

of *Progress in Brain Research*, pages 441–464. Elsevier, 2013. doi: https://doi.org/10.1016/B978-0-444-62604-2.00022-8. URL https://www.sciencedirect.com/science/article/pii/B9780444626042000228.

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.

Katja Mehlhorn, Ben Newell, Peter Todd, Michael Lee, Kate Morgan, Victoria Braithwaite, Daniel Hausmann, Klaus Fiedler, and Cleotilde Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2, 07 2015. doi: 10.1037/dec0000033.

S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nat Neurosci*, 22(10):1677–1686, 2019. ISSN 1546-1726 (Electronic) 1097-6256 (Print) 1097-6256 (Linking). doi: 10.1038/s41593-019-0502-4. URL https://www.ncbi.nlm.nih.gov/pubmed/31551604. Musall, Simon Kaufman, Matthew T Juavinett, Ashley L Gluf, Steven Churchland, Anne K eng R01 EY022979/EY/NEI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2019/09/26 Nat Neurosci. 2019 Oct;22(10):1677-1686. doi: 10.1038/s41593-019-0502-4. Epub 2019 Sep 24.

Stefano Palminteri, Valentin Wyart, and Etienne Koechlin. The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433, 2017. ISSN 1364-6613. doi: 10.1016/j.tics.2017.03.011. URL https://doi.org/10.1016/j.tics.2017.03.011. doi: 10.1016/j.tics.2017.03.011.

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

Eric Schulz and Samuel J. Gershman. The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55:7–14, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2018.11.003. URL https://www.sciencedirect.com/science/article/pii/S0959438818300904. Machine Learning, Big Data, and Neuroscience.

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

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.

M. S. Tomov, V. Q. Truong, R. A. Hundia, and S. J. Gershman. Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nat Commun*, 11(1):2371, 2020. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-020-15766-z. URL https://www.ncbi.nlm.nih.gov/pubmed/32398675. Tomov, Momchil S Truong, Van Q Hundia, Rohan A Gershman, Samuel J eng R01 MH109177/MH/NIMH NIH HHS/ S10 OD020039/OD/NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2020/05/14 Nat Commun. 2020 May 12;11(1):2371. doi: 10.1038/s41467-020-15766-z.

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.

Robert C Wilson, Siyu Wang, Hashem Sadeghiyeh, and Jonathan D Cohen. Deep exploration as a unifying account of explore-exploit behavior. Feb 2020. doi: 10.31234/osf.io/uj85c. URL https://doi.org/10.31234/osf.io/uj85c.

Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38: 49–56, 2021. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2020.10.001. URL https://www.sciencedirect.com/science/article/pii/S2352154620301467. Computational cognitive neuroscience.

Valentin Wyart. Leveraging decision consistency to decompose suboptimality in terms of its ultimate predictability. *Behavioral and Brain Sciences*, 41:e248, 2018. doi: 10.1017/S0140525X18001504.

Valentin Wyart and Etienne Koechlin. Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, 11:109–115, 2016. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2016.07.003. URL https://www.sciencedirect.com/science/article/pii/S235215461630136X. Computational modeling.