Behavioral Neuroscience

The effects of time horizon and guided choices on explore-exploit decisions in rodents

--Manuscript Draft--

Manuscript Number:	BNE-2022-0363R1
Full Title:	The effects of time horizon and guided choices on explore-exploit decisions in rodents
Abstract:	Humans and animals have to balance the need for exploring new options with exploiting known options that yield good outcomes. This tradeoff is known as the explore-exploit dilemma. To better understand the neural mechanisms underlying how humans and animals address the explore-exploit dilemma, a good animal behavioral model is critical. Most previous rodents explore-exploit studies used ethologically unrealistic operant boxes and reversal learning paradigms in which the decision to abandon a bad option is confounded by the need for exploring a novel option for information collection, making it difficult to separate different drives and heuristics for exploration. In this study, we investigated how rodents make explore-exploit decisions using a spatial navigation Horizon Task (Wilson, Geana, White, Ludvig, & Cohen, 2014) adapted to rats to address the above limitations. We compared the rats' performance to that of humans using identical measures. We showed that rats use prior information to effectively guide exploration. In addition, rats use information-driven directed exploration like humans, but the extent to which they explore has the opposite dependence on time horizon than humans. Moreover, we found that free choices and guided choices have fundamentally different influences on exploration in rodents, a finding that has not yet been tested in humans. This study reveals that the explore-exploit spatial behavior of rats is more complex than previously thought.
Article Type:	Research Article
Corresponding Author:	Jean-Marc Fellous, Ph.D. UA: The University of Arizona UNITED STATES
Corresponding Author E-Mail:	fellousjm@gmail.com
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	UA: The University of Arizona
Other Authors:	Siyu Wang
	Blake Gerken
	Julia Wieland
	Robert Wilson
Corresponding Author's Secondary Institution:	
First Author:	Siyu Wang
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED STATES
Order of Authors:	Siyu Wang
	Blake Gerken
	Julia Wieland
	Robert Wilson
	Jean-Marc Fellous, Ph.D.
Manuscript Classifications:	10.10.030: decision making; 10.50.060: reward and punishment; 20.70.040: motivated laboratory behavior (appetitive/aversive); 30.20.020: Rats

Cover Letter

Jean-Marc Fellous
Professor
Departments of Psychology, Biomedical Engineering and Applied Mathematics
1503 E. University Blvd.
Tucson, AZ 85721

THE UNIVERSITY OF
ARIZONA
TUCSON ARIZONA

Tel (office): (520) 626-2617
Tel (department): (520) 621-7447
Fax: (520) 621-9306
Email: fellous@email.arizona.edu
Web: http://www.u.arizona.edu/~fellous

May 18, 2022

Dear editorial team,

Please find attached our revision entitled:
"The effects of time horizon and guided choices on explore-exploit decisions in rodents".
This manuscript present a novel rodent behavioral paradigm involving spatial navigation and decision making that closely match that traditionally used in humans. We compare and contrast the rat and human behavioral data and found interesting and unexpected differences.
We have responded to both reviewer's comments point by point and added 6 new figures to the submission.

Yours Truly,

Jean-Marc Fellous

The effects of time horizon and guided choices on explore-exploit decisions in rodents

Responses to the Reviewers

We thank both reviewers for their comments. We have addressed them point-by-point below, modified the manuscript as indicated, and added 6 Supplemental figures to the manuscript to document our responses. The original comments are in black, our responses are in blue.

Reviewer 1 comments:

This manuscript by Wang et al. describes a novel rodent task for exploring explore-exploit decision making across varying "time horizons". Their approach is based on a similar task characterized humans in an earlier paper, and the current study includes further examination of human explore-exploit decisions on this task. The influence of relevant variables on task performance are described (though inferential statistics are scant), and this is supplemented with Bayesian model fitting to characterize parameters associated with two established heuristics for exploration: directed exploration (exploration threshold parameter) and random exploration (decision noise parameter). Like humans, rats performed the task well and were able to guide their choices by comparing recent information from guided trials with prior knowledge of possible outcomes. A major species difference was that humans increased their initial exploration of the unguided (unknown) option with time horizon (i.e., number of future opportunities to exploit feedback from guided and early choice trials), which represents a rational, directed exploration heuristic. Rats, instead, showed the reverse relationship, and additional data indicated that their exploratory behavior may have been driven in part by volatility in reward contingencies. Rats also appeared to show improved performance and less initial exploration when allowed to self-guided vs. being forced to sample one of the two response options. This is a really interesting study and the authors do a good job of highlighting the potential advantages and novel elements of this task versus other rodent tasks for measuring explore-exploit decision making. Despite some current issues (listed below), I believe the manuscript would be of interest to many in the field.

1) The conceptual framework for the current study is not very well developed, particularly in the Introduction, which assumes quite a bit of readers. The main question at hand - i.e., how time horizon relates to explore/exploit decisions is not really discussed until late in the paper (in Discussion), which would make it difficult for readers to understand the purpose of experimental design parameters, as well as predictions about results. Also, much of the introduction lists limitations of existing reversal learning tasks but it is not very clear how the current approach will improve on this. For instance, without providing a clearer idea of the structure of the current task and how it relates to relevant decision-making variables, it is not obvious why the current task does not have the same limitations. The current task also involves a 2-option choice with differential reward, so why is it better than reversal learning?

- We thank the reviewer for the critique on the clarity of the conceptual framework. We have expanded the description of time horizon in the introduction and elaborated on how time horizon relates to exploration. The description of time horizon is now written in a separate paragraph in the modified manuscript.
- We also modified the paragraph about the limitations of the reversal learning paradigm, and explicitly stated how our design addressed these limitations in the last paragraph of the introduction. Our task is better than the reversal learning approach mainly in the following ways:
    - Firstly, both good and bad outcomes should occur in exploration. However, in reversal learning, after the reversal point, "exploring" the previously suboptimal option will always lead to a better outcome. In reversal learning paradigms, exploration is confounded with simply abandoning a bad option, while in our design, exploring the unguided option can lead to either better or worse outcomes.
    - Secondly, it is notoriously difficult to separate different drives and heuristics for exploration in reversal learning paradigms. To study directed exploration (i.e. uncertainty driven exploration) for example, there needs to be a difference in uncertainty between the two options. For reversal learning, this uncertainty difference is implicit in that the less chosen option has more uncertainty. Since the less chosen option in reversal learning usually also has a lower estimated value, value and uncertainty are confounded in reversal learning. However, in our design both value and uncertainty are manipulated independently from each other, allowing us to dissociate uncertainty from value and properly measure directed exploration.

2) There are very few statistical results provided. Instead, the Results consists of general descriptions of group level performance without evidence of the significance of findings. This is generally problematic but especially for subtle effects (e.g., constant vs. random reward differences in Fig 10; horizon effects on model parameter estimates in Figure 8; reference to horizon effect on switch in Fig 4D on p. 14).

We conducted additional statistical analyses and included more details. We added missing p-values throughout the results section of the manuscript. We re-ran all the analyses after incorporating suggestions from other comments of the reviewer (e.g. controlling for feeder preference). Here we highlight the main findings of the paper with statement of statistical significance.

1. (Similar to humans) Rats were able to use prior information to guide exploratory choices. In figure 4C, rats were able to choose the high reward option at the first free choice (without knowing the reward of the unguided option) significantly above chance ($p < 0.001$ for $H = 1$ and 6, $p = 0.01$ for $H = 15$). Rats can only perform above chance if they have access to prior information.
2. (Similar to humans) Rats adapted the extent to which they explore based on the guided reward size. In figure 7C, Two-way ANOVA (Horizon x Guided reward) revealed a significant main effect of guided reward ($p < 0.001$).

3. (Different than humans) Rats used less directed exploration (i.e. have lower thresholds) in long horizons.
    a. In Experiment 1(Figure 8), there is a significant main effect of horizon on the threshold parameter for humans ($p < 0.001$), but there is no significant effect of horizon on the threshold parameter for rats ($p > 0.05$). In other words, when different horizon conditions are tested between sessions, we didn't find a significant horizon effect on directed exploration in rats.
    b. In Experiment 2(Figure 9), there is a significant main effect of horizon on the threshold parameter for rats ($p < 0.001$). There is also a significant main effect of horizon on the model-free p(unguided) measure ($p = 0.003$). Together, these results showed a significant horizon effect on directed exploration if rats experienced both horizons within the same session. (Note that the horizon conditions are always within-session for human experiments.)
4. (Rats alone) Rats explored differently in self-guided exploration compared to cue-guided exploration. In Figure 11D, we showed that there was a significant main effect of nGuided (0 for self-guided vs 1 for cue-guided) on p(unguided), $p < 0.001$.

Addressing the particular figures that the Reviewer mentioned:

5. (Figure 10D, F) For random reward condition, we showed that there was a significant increase of both threshold ($p < 0.01$) and noise ($p < 0.01$) parameters compared to the constant reward condition.
6. (Figure 8) See 3a above. We have added to figure 8 that the effect was non-significant.
7. (Figure 4D) P(switch) at later trials (trial# 2-6 in H = 6 vs trial# 11-15 in H = 15) is significantly lower for H = 15 than for H = 6 ($p < 0.001$).

3) Some additional methodological details should be provided or clarified. What volume were the sugar water drops? Were the rats food/water restricted? How many trials/games per session? Were rats trained through different phases at different rates based on performance? Were rewards ever symmetrical across options within games and if so how was this dealt with for analysis?

- The volume of sugar water drop was 150 microliter per drop.
- The rats were food restricted to 85% of their ad libitum weight and were not water restricted.
- For Experiment 1, rats on average do 31.3 games and 125.3 trials per session for H = 1, they do 13.5 games and 121.5 trials per session for H = 6, and an average of 6.5 games and 123.4 trials for H = 15. On average, each rat completed 348.7 H = 1 games, 375.8 H = 6 games, and 170 H = 15 games.
  For Experiment 2, rats completed an average of 25.62 games and 130.44 trials per session.
- Rats were pretrained through different phases at different rates during pretraining. We first trained rats to associate light with reward, then they were trained to go to the homebase to trigger the lights at the reward feeders (first with reward at homebase, then without the homebase reward), then they were trained to learn that two feeders give

different amounts of rewards (first 0 vs 1 drops, then 1 vs 5 drops, then the full reward schedule).Rats went through these phases of pretraining at different rates based on their individual performance. After pretraining, all rats performed three experiments in the order: Experiment 1, Experiment 2 and then Experiment 3. We have added a paragraph describing this pre-training protocol in the manuscript.

- Rewards of the two options were generated independently using a custom written MATLAB program for each game. So indeed, rewards for two options can be identical (6.6% of all trials). These trials were included in all analysis that focused on the 1st free choice (since rats only know the guided reward before making the first free choice, whether the unguided option has an identical value did not matter). These trials were excluded when doing analysis of later trials, as in Figure 4, 5, 10AB, 11AB.

4) For Experiment 2, it was unclear exactly how H conditions were organized within session, though it seems to be the case that they were strictly alternated when confounded with home base. This should be in Methods. Also for this experiment, how were nG 0, 1, and 3 conditions organized (e.g., blocks of sessions, randomly across sessions).

- Horizons were strictly alternated in Experiment 2, for a given session, one of the homebase was always used for H = 1 and the other always for H = 6. For each session, a MATLAB program pseudo-randomly made the homebase/horizon condition pairing. Consequently, Homebase A could be H = 1 for one session and H = 6 for another session.
- nG = 0,1,3 were run in blocks of sessions.

5) For the human task, does the schematic in Figure 3 represent actual task stimuli and procedures? For example, were subjects given a tally of past reward histories for all trials within a game? This should be indicated in Methods.

- The schematic in Figure 3 represents the actual task stimuli. Subjects were indeed given a tally of past reward histories for all trials. We have added this information to the Methods.

6) Do the data presented in figures represent all the data from all relevant sessions or were they restricted to sessions after rats had time to learn about the new task contingencies? For example, rats presumably took some time to learn about the change in time horizon across blocks of sessions in Experiment 1. And the same goes for when they switched to the within-session analysis of horizon in Experiment 2, and the random task in Experiment 3. As noted by the authors, performance in the random task shows some clear carryover from the earlier phases of testing. This should be specified as good practice but also raises questions about species differences.

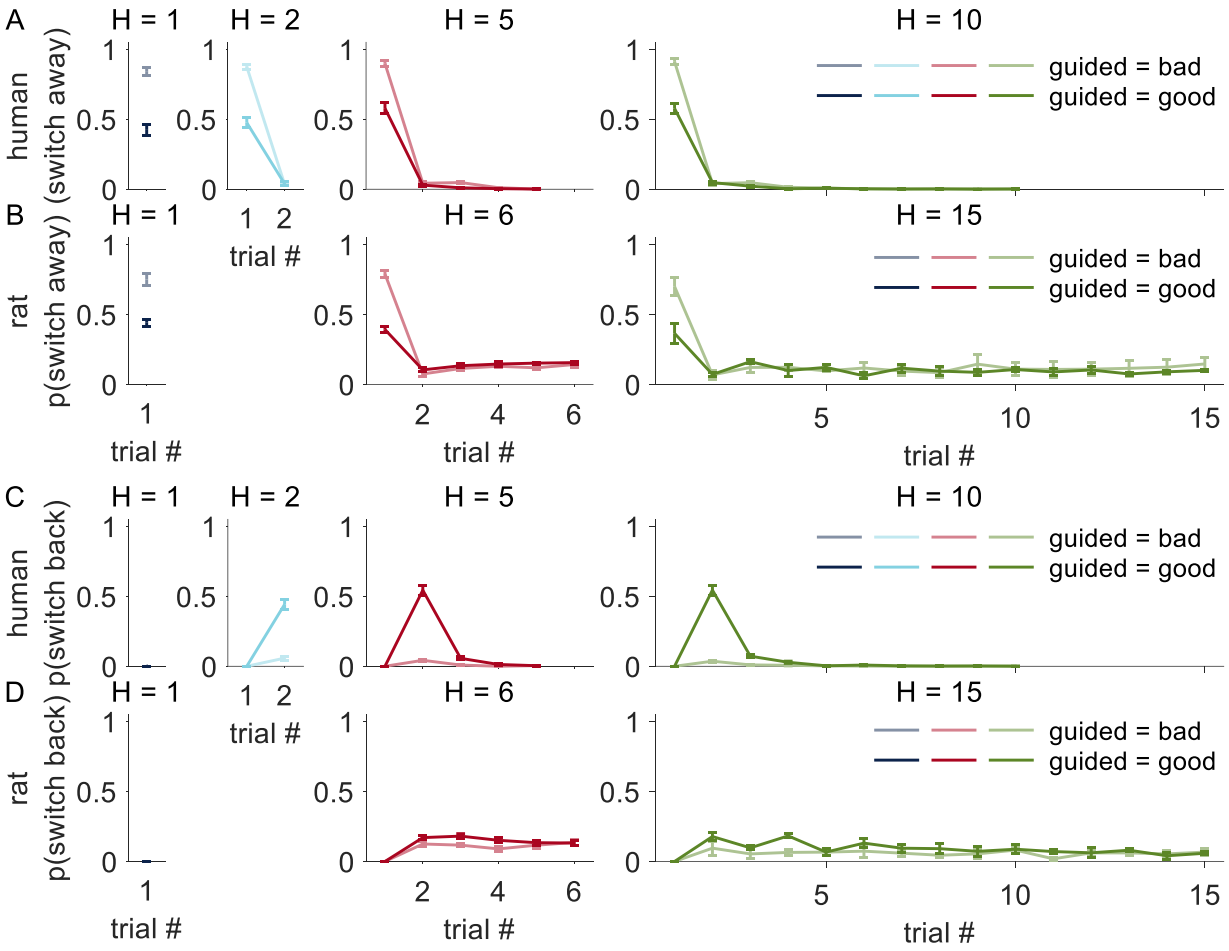
- We thank the reviewer for this useful suggestion. The data presented in the initial submission used all the data from all relevant sessions. In the current submission, we adopted the reviewer's suggestion and analyzed the data after excluding the transitioning

sessions. In Experiment 1, we excluded 1 whole session after each transition of horizon conditions. In Experiment 2, we excluded the first session for each rat, and we excluded the first 2 games in each session (Although we have high/low pitch sound cues played at the homebases to help signal H = 6/H = 1 horizon condition respectively, in practice, it may take rats 1 full game to learn the associated horizon condition with each homebase). These exclusions did not change our results or conclusions.

- Experiment 3 was carried out at the end of Experiment 2 to test how volatility might account for horizon-dependent changes in exploration in rats. Despite the possible carry-over effects (we expect carry-over effects for threshold but not for decision noise), the fact that volatility changes both the decision noise parameter and the threshold parameter, and that only threshold changes across horizons in actual behavior, suggest that volatility may not be what is driving the threshold specific horizon-dependent changes in rats.

7) Humans seem to have little trouble deploying a directed exploration strategy based on time horizon and guided choice feedback (reward size). They seem to explore the unguided option during the first free choice if there is any question about what the best option is, particularly when there is a long horizon to exploit that information. The authors state (p. 10) while showing similar early exploration, "it took longer for rats to switch back," referring to their persistent switching behavior in Fig 4 and 5. But these data are really able to get at this question precisely because they don't describe whether the switch is moving away or toward the guided choice. Given the rats' generally poorer performance and persistent tendency to switch within games, even with long horizons, suggests that they were switching back and forth from the best option. This is later discussed in the context of the random reward task, but the authors should avoid giving the wrong impression when discussion Fig 4 and 5.

This is a very good point. To address the reviewer's comment, we split p(switch) into p(switch away) vs p(switch back). We found that the difference between the "guided = good" option and the "guided = back" option occurred in the direction of switching back to the guided choice. On top of the persistent baseline switching in rats, rats do switch back from the unguided option to the guided option more (when the guided option is objectively better) up to trial #4 in H= 6, p = 0.01. A new figure was added.



8) The cross species comparisons are a bit strained despite the general similarities across tasks. For instance, humans appear to receive a continuous tally of past reward within each game, which explains their lack of later exploration once they understand the basic task. But rats could forget this information and decide to re-explore the options, particularly in games with long horizons though reward volatility on short horizon games may provide a separate reason to explore.

We thank the reviewer for bringing up this critical point. We want to point out the following:

1. We agree that despite the matching underlying structure, there are some differences between the rat and human version of the task (reward history vs no history, points for humans and juice for rats, effortful spatial runs in rats vs effortless key presses in humans). However, our main interest is in how humans and rats change their exploration behavior across horizon conditions. This horizon comparison is done within species (We know of no evidence that differences in the physical implementation of the task would contribute to the horizon difference within species), and then we showed a qualitative difference between species that humans increase whereas rats decrease the threshold parameter in long horizon condition compared to short horizon condition. Given that the horizon difference is calculated in a similar manner within species, the results should be valid within each species. This being said, our results are indeed prompting further work and new tasks in humans that would

match that of the rats, for example a physical spatial task where human subjects walk from home base to the bandits. These new tasks are left for future work.

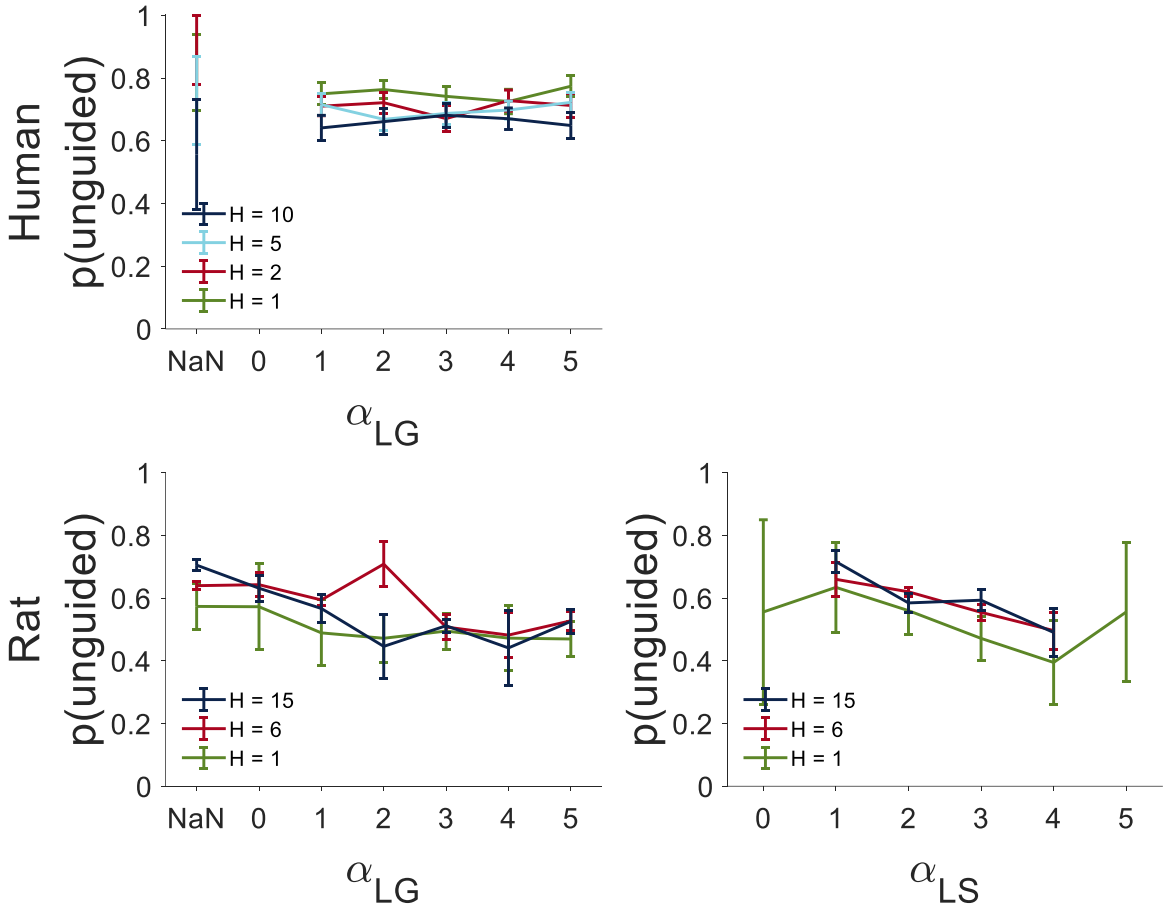
2. The reviewer is correct that rats might re-explore because of forgetting. For the same number of guided trials, rats are equally likely to forget in H = 1 and H = 6 however. We agree that reexploring is an important form of exploration (uncertainty increases as animals forget, and reexploring can be considered a form of directed exploration), but forgetting does not predict a difference in behavior across horizon conditions.

9) The authors suggest that rats may be performing the task to satisfice instead of optimize, like their human counterpart. This is related (p. 17) to the higher tendency for humans to explore than rats. This conclusion is partly based on the idea that rats must exert more effort and may be less willing to explore (a point also made earlier in the manuscript). But of course this specific data refers to first trial exploration. Rats were not less willing to explore on long horizon games. As indicated above, they tended to switch back and forth throughout the game (i.e., even when they had sampled both reward contingencies).

This is a really good point. We agree that the difference in effort can not account for the fact that rats were willing to switch in later trials in long horizon games. This constant switching in later trials may indeed reflect random exploration. We agree it was not completely accurate to say that "more effort" leads to "being less willing to explore". In the modified manuscript, we emphasize that "more effort" relates more to "not optimize" rather than "not explore". It changes the relative utility/effort balance of the options, so that it is not worth risking getting a 1 drop (to optimize the check if the unguided reward is a 5) if the guided reward is 4 drops for rats. This deliberate way of exploration is directed exploration rather than random exploration. "Effort" might influence directed exploration more than random exploration. To properly test the effect of efforts on exploration, future experiments could potentially run identical tasks in long/short distance (or maze vs boxes). We would predict that rats in an effortful setup will be less willing to engage in directed exploration than rats operating in an effortless setup (e.g. a lever pressing task). We re-wrote the relevant sections.

10) The greater preference for high reward option and lower level of switching in the free choice vs. guided choice is interesting but one account not considered is that rats some feeder/spatial preferences that bias them on these free choice first trials that continuous to bias their performance to the same degree within each game. This make sense generally and also explains why the effect of free vs. guided choice is so stable across variables like guided reward size and horizon. This could be a long-term feeder bias or perhaps something more dynamic (e.g., like a preference for whichever feeder has been paying off better in recent sessions).

We thank the reviewer for raising this critical and insightful point. In an attempt to measure the influence of short-term/long-term feeder bias, we calculated two measures. We first computed the reward from last game R_{LG}(most recent reward of the current feeder from a previous game within session) to quantify short-term feeder bias, and computed the average reward from last session R_{LS} to quantify long-term feeder bias.



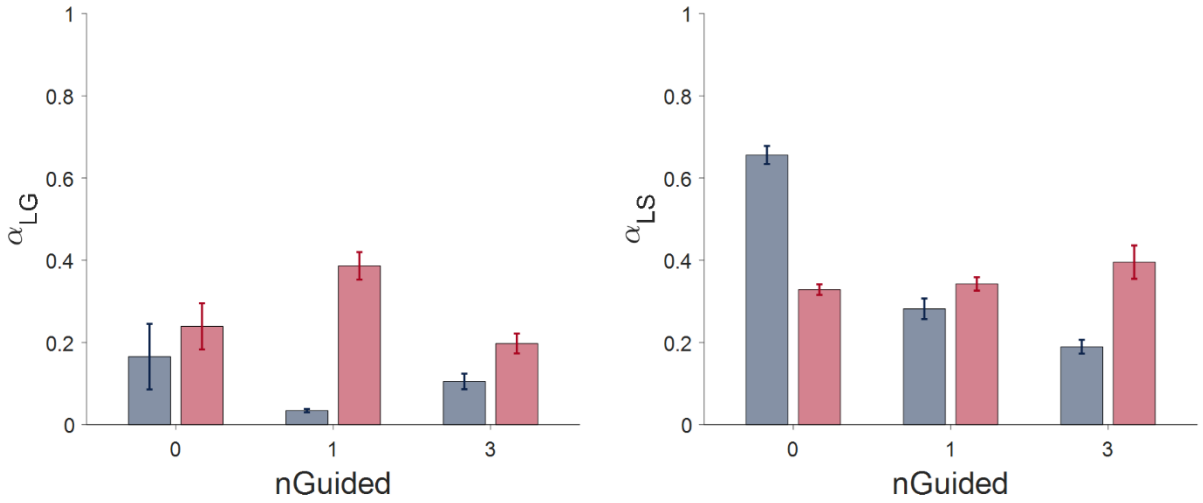
We first plot p(explore) as a function of both R_{LG} (NaN means that the rat didn't choose the guided feeder in the previous game) and R_{LS} (LS values are binned). ANOVA analysis showed that both LG ($p < 0.001$) and LS ($p = 0.02$) have significant influence on p(explore). These showed that rats do have feeder biases. (Humans do not.)

As a result, we repeated all the model-based analysis by including these two additional parameters in the model to account for feeder biases. Specifically, we now have

$$\Delta Q = R_{guided} - \theta + \alpha_{LG}(R^{LG}_{guided} - R^{LG}_{unguided}) + \alpha_{LS}(R^{LS}_{guided} - R^{LS}_{unguided}) - b * s_{guided}$$

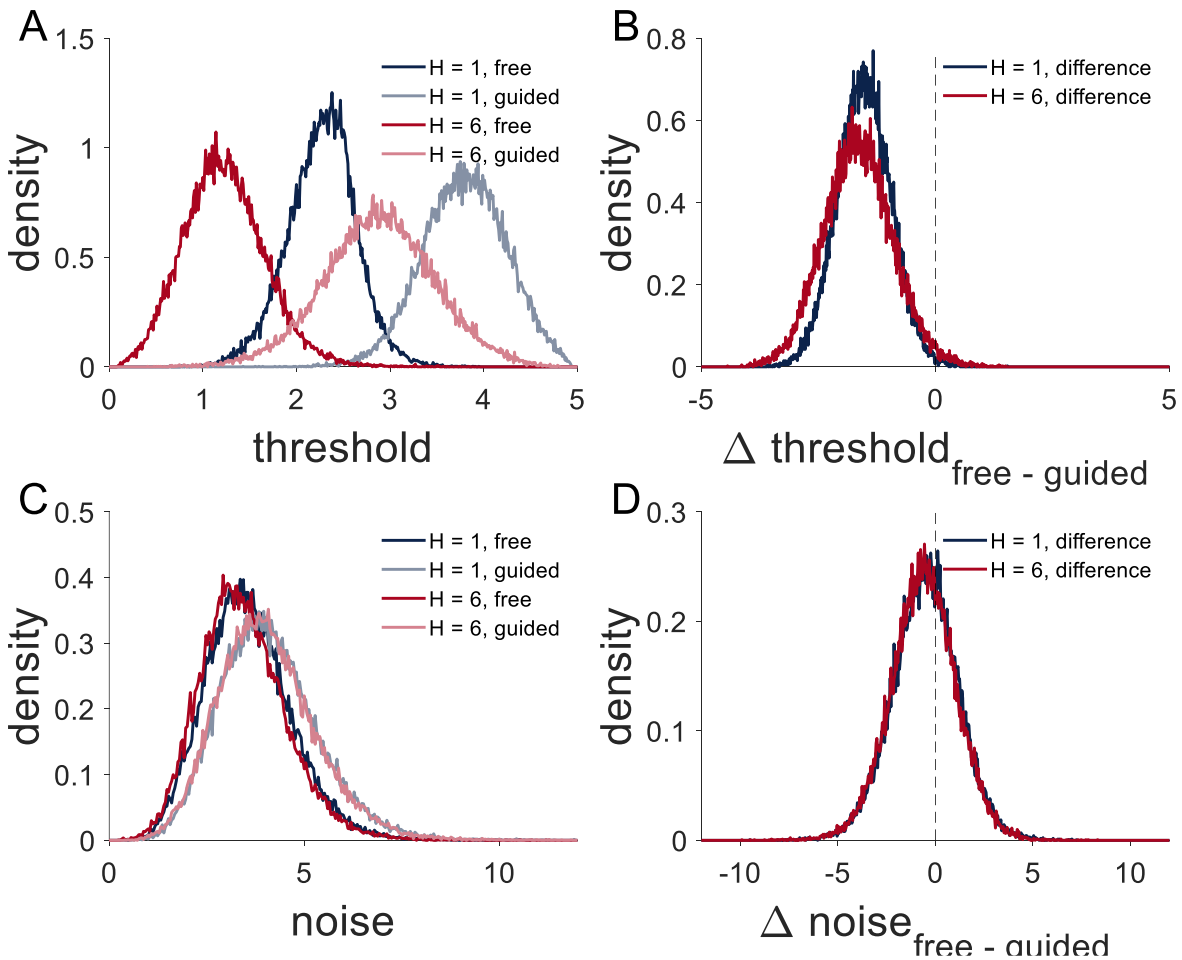
Here, in addition to the current reward, decision threshold, and spatial bias, we added two new terms, the first LG term quantifying the influence of feeder bias from last game, and the second LS term quantifying the influence of feeder bias from last session. Parameter recovery shows that our model estimates the upper bound of the feeder bias coefficients (Our model tends to overestimate α_{LS}, Fig S1). Through model fitting, we confirm and acknowledge that rats are influenced by both short-term and long-term feeder bias. LG coefficient is significantly larger in H = 6 than H = 1 condition, showing that short term feeder bias (from last game) has a significantly bigger influence on H = 6 games ($p < 0.001$). This is likely due to the fact that rats

spend more trials at H = 6 feeders within a session. There are no differences in long term feeder bias (from last session) between horizon conditions (p = 0.48).



(We note that the coefficients are relative. α_{LS} appears to be larger than α_{LG} only because $\Delta R^{LS} < \Delta R^{LG}$, in fact, $\alpha_{LS} = 0.6$ has roughly the same impact on choices as $\alpha_{LG} = 0.2$.)

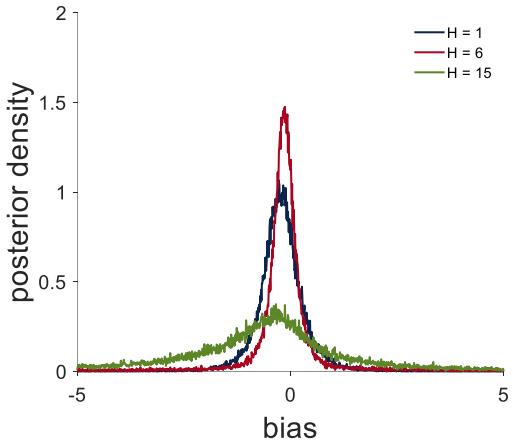
Next, we confirmed that after accounting for feeder biases, the model-based results in the paper still holds, e.g., we still see reliable differences in threshold between horizon conditions (All the figures have been re-built using this extended model) in Experiment 2. In particular, we still observe the difference in the threshold parameter between self-guided vs cue-guided trials.



Despite these feeder biases, our model suggests that the large difference in thresholds between self-guided vs cue-guided conditions is still present. These figures and results were added in supplemental materials.

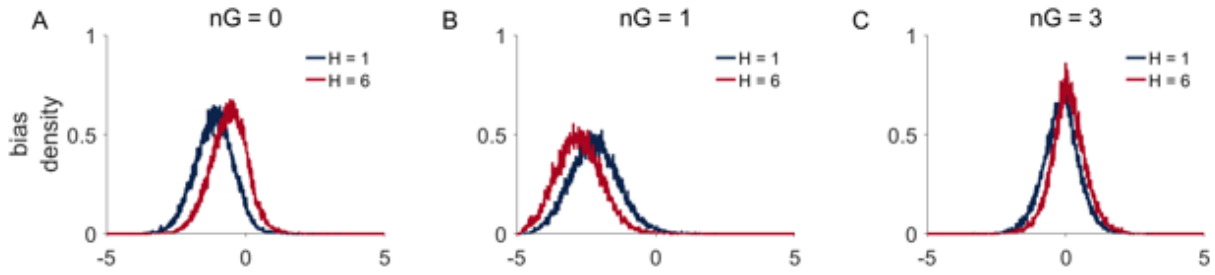
11) On a related note, a spatial bias parameter was computed in the model fitting but was not discussed.

The results of spatial bias parameter are now added to the supplementary materials. The spatial bias is centered at 0 in all conditions in Experiment 1.



But we did observe a left side bias when rats were guided only once (nG = 0 and 1) in Experiment 2. Interestingly, this left side bias is compatible with previous work from our laboratory in a different spatial task relevant to spatial navigation optimization, and may be related to rat right-handedness (Watkins de Jong, 2011).

The Traveling Salesrat: Insights into the Dynamics of Efficient Spatial Navigation in the Rodent. Watkins L., Gereke B. G. M. Martin, JM Fellous . J Neural Engineering, 8(6), 2011.



The bias does not change significantly with horizon ($p > 0.05$).

Minor issues:

12) P. 10 - statement about boredom or motor error seems to refer to residual responding at last trial but worded as if about the reason for the decrease in switching.

We thank the reviewer for pointing this out. We have rephrased this on P.10.

13) P. 11 - it says "3 or 4 drops for humans"

We have rephrased this on P.11

Reviewer 2 comments:

(Please distinguish between essential and non-essential revisions, as indicated in instructions at the top of this page, and please do not indicate whether or not paper should be published.)

The effects of time horizon and guided choices on explore-exploit decisions in rodents
Wang / Gerken / Wieland / Wilson / Fellous

In this manuscript, the authors translate an elegant exploration/exploitation task originally developed for humans to rats. They find that rats show different effects of time horizon than humans do. While I applaud the goal of this project and both versions of the task (human and rodent) are elegant, there are too many differences between the tasks and the subjects to draw the conclusions they want to.

Overview: I think this could be an absolutely excellent and groundbreaking study. But doing so requires additional experimentation and additional analyses. In the manuscript's current form the data is inadequate to draw conclusions from.

Comments:

- [ESSENTIAL REVISION] Sex of the subjects. The experiment was done with 6 (very small n!) of only male rats. The human studies were done with 45 psychology undergraduates (14m, 31f). Given that we know there are sex differences in both human and non-animals on exploration/exploitation choices (for example, the cited study by Chen…Grissom), it is inadequate to compare 6 male rats to a human distribution that is 70% female. (Note that I don't have a problem with using undergraduates as the human model species. It's probably an appropriate comparison to the laboratory rat.) [However, isn't it NIH policy now that all non-human animal experiments *must* include both male and female animals unless a scientific reason can be given for limiting it? Nevertheless, whatever the NIH policy is, the rat cohort is simply too small and limited for the conclusions.]

We thank the reviewer for raising the question about gender differences.

1. We agree that the issue of sex differences is important, but we do not propose or formulate any hypothesis about sex differences in this task. Recent published work using similar task designs to ours reported no sex differences in either directed or random exploration measures (Smith et al 2021). Also, in our own human data for Experiment 4, a Two-way ANOVA (horizon by gender) analysis showed a significant main effect for horizon ($p < 0.001$) and a non-significant effect of sex ($p = 0.59$).
2. To properly study how female rats might behave differently using our design would require an assessment of their estrous cycles, and possibly the inclusion of ovarectomized animals. This work, which is not funded by our current grants, would add 2 years of additional work. Again, acknowledging the need for more work on female animals, we leave such work for future studies.
3. The rodent portion of our study is not a NIH funded study and is not subject to any requirement regarding sex-balance. Our emphasis is on developing and characterizing a

new rodent model of the explore-exploit task compatible with a human task. Study of how populations characteristics (e.g. sex, age), while certainly important, are left for future work.

We thank the reviewer for raising the concern about our sample size.

1. Despite having small number of rats (6), we have a very large number of games from each rat. Excluding pretraining, we collected a total of 530 sessions, 67781 trials for our rat experiments. Our human study has a total of 70400 trials which is similar and comparable to that of our rats. The number of trials are matched. Through hierarchical Bayesian modeling (which pools all trials from all subjects in model fitting), we believe our human/rat comparison was carried out in a fair way.
2. For Experiment 1 and 2, we have over 30000 trials in each experiment. This number is comparable to the trial sizes used in the literature. Here we list of the number of animals/trials used in previous explore-exploit rodent studies.

	Maze or Box	Number of rodents	Number of sessions	Total running time	Number of trials
Our study	Maze	6	530 (200+ per experiment)	~530 hours (1-2 hours/session)	67781 total (30000+ per experiment)
Beeler et al., 2010	Box	10 (2 groups)	130	Unknown	68096
Laskowski et al., 2016	Maze	22 (2 groups)	352	~120 hours (20 min/session)	~29300
Chen et al., 2021	Box	32 (2 groups)	256	512 hours (up to 2 hours/session)	70656
Cinotti et al., 2019	Box	23	184	Unknown	52992
Verharen et al., 2020	Box	60	196	Unknown	~49000

Although our number of subjects is 6, our number of trials is comparable to that of other rodent explore-exploit studies (especially the Laskowski paper that also uses a spatial maze design). Also note that we reported another 70000 trials in humans in our study. Since our design compares H = 1 and H = 6 within subjects, we only needed half as many rats as in a between-subject design study.

3. We are minimizing the use of animals as per IACUC rules. We did not report claims in our results based on a trend, all our main results are statistically significant.
4. We acknowledge that our Experiment 3 should be considered preliminary. However, Experiment 3 is not essential to the conclusions of this study.

- [ESSENTIAL REVISION] The rat experiment starts with a subset of cued responses. It is not clear whether the rats treated the cued responses and the uncued as part of the same "condition".

I don't know how to fix this, but it needs to be considered. The authors do acknowledge that the rats may have been making some decisions based on overall volatility rather than the actual time-horizon experiment desired, but they do not disprove this hypothesis. I think it is necessary for the authors to directly test these alternative theories (using Bayesian model fits?) and show that these alternate explanations are not as good explanations as their intended experimental logic.

We do not believe the "cued" and "uncued" conditions affect the validity of our results for the following reasons.

1. There are no "uncued" conditions in our design. In the guided trials, light at one of the two feeders is on and the rat is guided to visit that feeder. In the unguided trials, both lights are on at the two feeders. Rats are pretrained to understand that they are allowed to travel to either feeder locations with lights on to obtain rewards. Blinking the light is a necessary step to draw the attention of the rat to those two feeders that are active (there are 6 other inactive feeders at the periphery). We have revised and expanded the method section to clarify this point.
2. The fact that rats adapted the percentage of switching in the first "uncued" choice based on the rewards gained in the "cued" responses, shows that rats seemingly transferred the learned value from the "cued" phase to the "uncued" phase. This shows that rats did not treat them as different conditions.

We thank the reviewer for raising about the issue of the effect of volatility on exploratory behavior.

1. Our results actually suggested that the horizon difference can not be accounted for by volatility alone. In Experiment 3, our results showed that increase volatility in the task (the random condition) increased both directed and random exploration. Since only directed exploration and not random exploration selectively changed between horizons in rats, volatility does not account for the horizon difference we observed in Experiment 2.
2. We have modified both the results and the discussion section of the paper to emphasize the above point.

- [ESSENTIAL REVISION] The rat experiments used the same 6 rats for all experiments. This is a major problem as the rats are almost certainly going to come to the subsequent experiments with learned expectations from the previous ones. The authors should, instead, identify the number of rats needed for each cohort and do a separate experiment with each group of rats naïve to the experimental paradigm, so that they all have the same training experience.

We thank the reviewer for raising the question about the carry-over effect across experiments.

1. From a training perspective, in order to teach a rat to learn Experiment 2, it is actually necessary that they perform Experiment 1 first. For all our rats, they first learned how to perform the task in a fixed horizon condition (Experiment 1), then we introduced both

horizon conditions within each session (Experiment 2). Naïve rats are in our hands incapable of doing Experiment 2 without going through Experiment 1 first.

2. Any carry-over effects from Experiment 1 to Experiment 2 should affect both horizon conditions equally in Experiment 2. The horizon difference we observed in Experiment 2 cannot arise from previous Experiment 1 exposures, because we randomly assigned the 2 horizon conditions to the 2 homebases each day. This has now been indicated in the method section.
3. Experiment 3 may have some carry-over effects from Experiment 1 and 2. We do not think it is necessary to run Experiment 3 in naïve rats, however. First, Experiment 3 is not essential to the main conclusions of the paper. We acknowledged that exposure to Experiment 1 and 2 may contribute to the high threshold (more switching) in Experiment 3, but it is unlikely that the high decision noise parameter arises from any carry-over effects from Experiment 1 and 2. Since increasing volatility increases decision noise, it still holds from Experiment 3 that volatility itself does not account for the horizon differences we observed in rats. We made this clearer in the text.

- There remain some real differences between the tasks. In particular, the human rewards are not consumed in the present, and thus form an amortized goal that can only be used (can they? were the points used for anything? That was unclear) in total. In contrast, the rats are receiving a reward with direct intrinsic value (sugar water). This means that the rewards are both immediate and biologically necessary for the rats, but neither for the humans. It would be better to try to match these if possible. That being said, I am not as concerned about the physical differences between tasks. I agree that rats treat space and levers differently, as do humans. The real question is whether the authors can show somehow that the computational algorithms *necessary* to solve these tasks are equivalent or whether alternative computational algorithms are possible. If there are alternative computational algorithms possible, then the authors need to show that they do not match with the observed behavior.

We thank the reviewer for the question regarding the differences across tasks.

8. We agree with the reviewer that there are differences between the rat and human version of the task (e.g. juice for rats & points for humans). Despite the physical differences, we want to highlight that the underlying structure of the two tasks are identical. So computationally, any algorithm that can solve the human version of the task, can identically solve the rodent version of the task. Many researchers have worked on comparing various computational models in various explore-exploit tasks (for example, see Gershman et. al, 2018). In the human version of our task, the computational model that separates directed vs random exploration has been well validated (Wilson et al., 2014). The interest of our paper is not to compare different explore-exploit algorithms, instead, we applied the established model in Wilson et al. 2014 to estimate directed and random exploration in both humans and rats and compare how directed vs random exploration parameters differ across horizon conditions and between species. We do find the suggestion of the reviewer that different algorithms might be worthy of study. In this work, we do not have any specific hypotheses on alternative models however, so we did not explore those possibilities. They will be left for further work. We make a mention of the reviewer's suggestion in the text.

References:

Gershman SJ. Deconstructing the human algorithms for exploration. Cognition. 2018 Apr;173:34-42. doi: 10.1016/j.cognition.2017.12.014. Epub 2017 Dec 29. PMID: 29289795; PMCID: PMC5801139.

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore-exploit dilemma. J Exp Psychol Gen. 2014 Dec;143(6):2074-81. doi: 10.1037/a0038199. Epub 2014 Oct 27. PMID: 25347535; PMCID: PMC5635655.

9. We agree that there is room to make the two tasks more identical in future studies. We edited the discussion to raise this issue. Since our research interest is in the change of exploration strategy with time horizon, within each species, the differences in physical implementation of the task should affect both horizon condition equally and hence should have minimal impact on the differences in horizon adaptive exploration.

10. Human participants, receive research credits after completing the task. However, the points in the task do not convert to monetary or other forms of physical reward.

While the authors do acknowledge many of these caveats (not all) in the discussion, I think that the experiment itself is (in its current form) too damaged by these caveats for us to take their conclusions.

We have substantially revised the interpretations of the results and added new figures. Our main contributions are:

a. We developed a rodent task that addressed the limitation of reversal learning paradigms (See introduction of the manuscript) and allows for a separation between directed and random exploration.
b. We showed that rats were able to use prior information to guide exploration (Figure 4C trial #1, Figure 7C).
c. We assessed how rats explored under different time horizon. We found that unlike humans, rats decrease their decision threshold (which quantifies directed exploration) in longer horizon condition (Figure 9).
d. We reported difference in decision thresholds between self-guided vs cue-guided exploration (Figure 11, 12).

We hope that the new figures, and re-write of the results clarify the conclusions, and make the caveats and limitations of the study more explicit.

- The description of the tasks is very poorly written. I did not understand the task as described in the experimental methods on my initial readthru and was only able to make sense of it as I read through the paper and kept coming to "wait, that doesn't work unless they did…" "oh, I see, yes, OK, they did." All of the factors do seem to be in the methods, but it was very hard to understand on a first read through.

We apologize for the confusions, which likely to have subtended some of the comments above. We have added a paragraph on a high-level description of the task before diving into the details in the methods section. We also have edited multiple sections of the manuscript to clarify. we hope this has improved the readability of our methods section.

- The Bayesian analysis of the tasks is very elegant, as is the separation of directed and random exploration.

Thank you.

- Figure 4 seems to show that there is no effect of horizon on the choices. I'm confused how this supports their conclusions. Figure 5 suggests that the choice is made less based on exploration than on whether the guided cue is near the boundary. (As the authors acknowledge, if the cue is near the boundary, then expected value of the other option is known to be higher [if guided is low] or lower [if guided is high].) Figure 5 seems to show that the entire effect is due to these boundary effects. Figure 7 argues that the humans are showing a subtle effect, but the rats are not [the decrease isn't significant, is it?]. Figures 8 and 9 seem to argue that there is an effect, particularly in the directed exploration parameter. How do all of these results fit together? How much of the difference is due to n(rats)=6 and n(humans)=45?

We thank the reviewer for bringing up these questions. We answer them in order below:

1. Figure 4 shows that both humans and rats have learned the task and performed well above chance, and that both humans and rats switched more on the 1st free choice compared to later trials. Figure 4 does not show measures of exploration. Despite having similar accuracy, exploration strategies can be very different. In the literature for example, Chen et al. showed that male and female rats have different learning rates and decision noises despite achieving a similar level of accuracy (Chen et al., 2021). In our task, exploration strategies were quantified using two parameters: “decision threshold” and “decision noise” using hierarchical Bayesian analysis. The differences in exploration parameters were shown in Figure 8 and 9.

2. We have clarified Figure 5. The good vs bad option in Figure 5 is relative, for example, if rats were guided to 1 drop and the unguided option offered 0 drop, then we consider the 1 drop option to be the good option, similarly if the guided option has 4 drops but the unguided option has 5 drops, then the 4 drops option is considered a bad option. So, Figure 5 does not indicate boundary effects. We have modified the manuscript accordingly to avoid the confusion.

   In fact, if an agent (rat or human) acts completely based on boundary effects, then we would predict that the agent will choose identically in both horizon conditions, since a boundary effect does not change with horizon condition. The fact that we observed significant changes in decision thresholds across horizon conditions indicated that these choices cannot arise entirely from boundary effects.

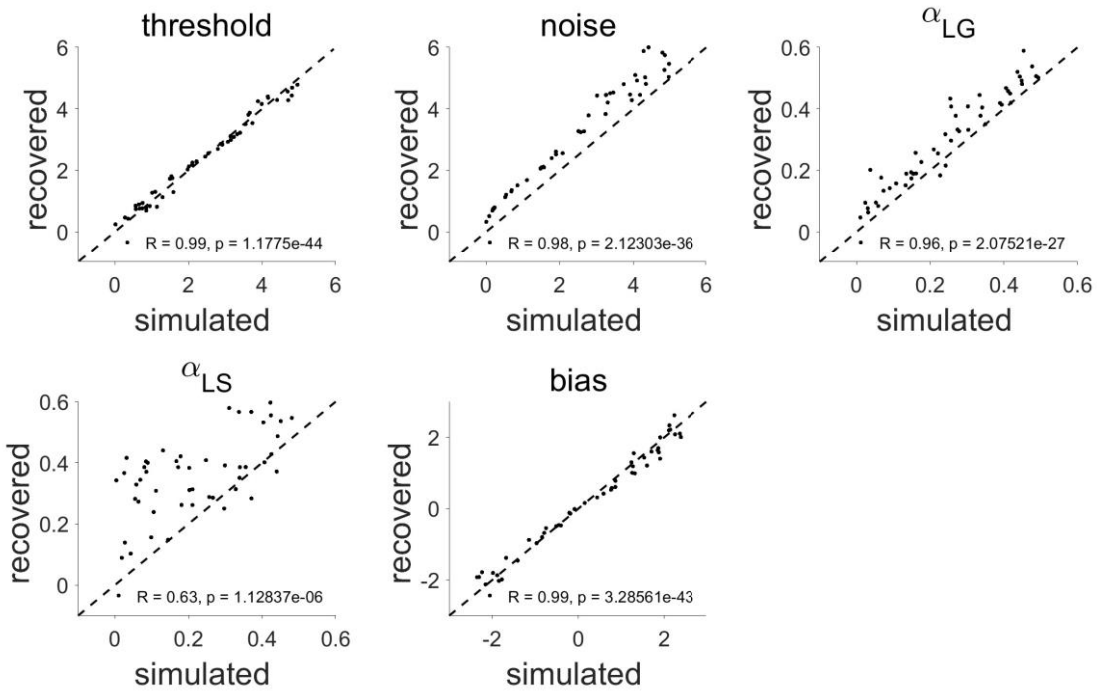
Figure 7C shows how our rats use a more general form of such "boundary effects" in solving this task (It is referred to as the 'win-stay lose-shift strategy' in our manuscript). Rats switch more when the guided reward is low (so do humans). But this does not contradict exploration for the reasons stated in the previous paragraph. We also would like to point out that at 5 drops, rats still switch more than 20% of the times, this also can not arise from pure boundary effect.

3. We found that humans increased their decision threshold with horizon (consistent with the existing human published literature), and that rats decreased their decision threshold with horizon (novel results from our paper).
    a. In Experiment 1 (Figure 7 and 8), there is a significant main effect of horizon on the threshold parameter for humans ($p < 0.001$), but there is no significant effect of horizon on the threshold parameter for rats ($p > 0.05$). In other words, when different horizon conditions were run between sessions (there is only one horizon condition for each session), despite the numerical trend, we didn't find a statistically significant horizon effect on directed exploration in rats. One critical difference between Experiment 4 and Experiment 1 was that horizon conditions were run within-session for humans and between sessions for rats. Third variables like training effects make it difficult to detect horizon dependent changes in exploration in Experiment 1. As a result, we conducted Experiment 2 (within design) in which rats also have both horizon conditions in the same session.
    b. In Experiment 2 (Figure 9), we found a significant main effect of horizon on the threshold parameter for rats ($p < 0.001$). There was also a significant main effect of horizon on the model-free p(explore) measure ($p = 0.003$). Bayesian analysis also shows clear separation of exploration thresholds between horizons (but not decision noise). Together, we showed a significant horizon effect on directed exploration in rats.
    c. The lack of significance in Experiment 1 could occur for many reasons. Firstly, in Experiment 1, different horizons were run in different blocks of days and sessions. Consequently, hidden variables such as the amount of exposure to training before each horizon condition (depending on the order of blocks), time of the day, weight of the rat, etc. could not be controlled and therefore could influence the horizon difference. The problem of latent variables as mentioned above was completely resolved in Experiment 2, since both horizons occurred in the same session. Secondly, experiencing both horizon conditions at once, magnifies the difference between horizon conditions. (Pilot human data in the lab also suggests that the horizon effect is weaker if different horizon conditions were blocked rather than interleaved)
4. The difference in the number of subjects between rats and humans should not affect our conclusions because we matched the number of trials and because the number of games are commensurate with that published by others (see above). Furthermore:
    a. Our main result was that rats decrease decision threshold with horizon, while humans increase decision threshold with horizon. The horizon difference was assessed separately within each species, using the same quantitative analyses.
    b. Given the difference in the nature of data acquisition for human participants and rats, we unavoidably have to deal with the fact that we have more subjects and fewer trials

per person in humans, and fewer subjects but more trials per subject in rats. By using Hierarchical Bayesian analysis we consider both variances across trials (within each subject) and variances between subjects.

- I also notice that the rats are showing much more random exploration (broader sigma density 8G/H) than humans (8C/D). Given that exploration is explained by a combination of directed [theta] and random [sigma] exploration, what is the effect of these sigma differences on the theta distributions?

This is an important control, thank you. We answer the reviewer question using parameter recovery. In Supplementary Figure 1, we simulated behavior using 50 combinations of independent samples of all the parameters used in the model, for each set of simulated behavior, we fit the simulated behavior and compare the recovered model parameters with the ground truth. Our model does a near perfect job in recovering both threshold and noise parameters in the experiment. This result suggests that the variation of sigma does not affect the model estimates of theta (threshold) in our model.



The effects of time horizon and guided choices on explore-exploit decisions in rodents

Siyu Wang[1], Blake Gerken[2], Julia R. Wieland[2], Robert C. Wilson[1,3], and Jean-Marc Fellous[1,4,5]

[1]Department of Psychology, University of Arizona
[2]Neuroscience and Cognitive Science Program, University of Arizona
[3]Cognitive Science Program, University of Arizona
[4]Program in Applied Mathematics, University of Arizona
[5]Department of Biomedical Engineering, University of Arizona

Corresponding author:
Jean-Marc Fellous
Department of Psychology
1503 E University Blvd, Room 312
Tucson, AZ 85721

Tel: 520-626-2617
Fax: 520-621-9306
Email: fellous@arizona.edu

Abstract: 219 words
Number of figures: 12 + 8 supplemental

Abstract

Humans and animals have to balance the need for exploring new options with exploiting known options that yield good outcomes. This tradeoff is known as the explore-exploit dilemma. To better understand the neural mechanisms underlying how humans and animals address the explore-exploit dilemma, a good animal behavioral model is critical. Most previous rodents explore-exploit studies used ethologically unrealistic operant boxes and reversal learning paradigms in which the decision to abandon a bad option is confounded by the need for exploring a novel option for information collection, making it difficult to separate different drives and heuristics for exploration. In this study, we investigated how rodents make explore-exploit decisions using a spatial navigation Horizon Task (Wilson, Geana, White, Ludvig, & Cohen, 2014) adapted to rats to address the above limitations. We compared the rats' performance to that of humans using identical measures. We showed that rats use prior information to effectively guide exploration. In addition, rats use information-driven directed exploration like humans, but the extent to which they explore has the opposite dependence on time horizon than humans. Moreover, we found that free choices and guided choices have fundamentally different influences on exploration in rodents, a finding that has not yet been tested in humans. This study reveals that the explore-exploit spatial behavior of rats is more complex than previously thought.

Keywords: explore-exploit dilemma, directed and random exploration, prior information, guided vs free exploration.

Acknowledgments: We thank Blaine Harper for giving insightful feedback on the task design, Maddie Souder for giving extensive help with rat training, and Kristine Gradisher and Blaine Harper for help with the rat experiments. The work was funded in parts by the Undergraduate Biology Research Program at the University of Arizona, NIH R01AG061888 (RCW) and NSF IIS-1703340 (JMF).

Introduction

Humans and animals constantly face a choice between exploiting options that are known to be good and exploring unknown options in the hope of discovering better future outcomes. Humans face this dilemma in many scenarios, from simple choices such as deciding whether to try a new restaurant for dinner, to important life decisions such as deciding whether to explore a new career. Animals face the explore-exploit dilemma when deciding whether to explore and forage for food, territory, or mates. The cognitive ability to balance exploration and exploitation is vital to animal and human survival and success. In recent years, the study of explore-exploit decisions in humans and animals has become an active field of investigation (Mehlhorn et al., 2015; Schulz & Gershman, 2019; Wilson, Bonawitz, Costa, & Ebitz, 2021).

An optimal solution to explore-exploit decisions is, in general, computationally intractable (Bellman, 1954), leading humans and animals to use approximations or heuristics. Previous research revealed that subjects were likely to use one or both of two main heuristics. The first is an information-driven heuristic known as *directed exploration* in which action is biased towards the more uncertain option (Banks, Olson, & Porter, 1997; Frank, Doll, Oas-Terpstra, & Moreno, 2009; Krebs, Kacelnik, & Taylor, 1978; Lee, Zhang, Munro, & Steyvers, 2011; Meyer & Shi, 1995; Payzan-LeNestour & Bossaerts, 2012; Steyvers, Lee, & Wagenmakers, 2009; Wilson et al., 2014; Zhang & Yu, 2013). The second is a noise-driven heuristic known as *random exploration*, in which exploratory actions with suboptimal estimates of reward value are chosen by chance (Badre, Doll, Long, & Frank, 2012; Feng, Wang, Zarnescu, & Wilson, 2021; Gershman, 2018, 2019; Kao, Doupe, & Brainard, 2005; Wang & Wilson, 2018; Wilson et al., 2014).

One key factor in explore-exploit decisions is the time horizon, i.e., the number of known future choices remaining which can be influenced by the current decision. Horizon adaptation is thought to be a hallmark of effective exploration. In a long horizon context in which you will make a lot of similar decisions later on, it's more beneficial to explore. For example, exploring a new restaurant in your area of living can benefit you in the long run, since you might enjoy it for the rest of your life. However, in a short horizon context, it is optimal to choose what is known to be best. For example, going to a highly rated restaurant is probably better than trying a newly opened restaurant if you are going for a one-time meal in a new city on a trip. Recent studies showed that humans were able to adapt the extent of their directed and random exploration with the time horizon (Wilson et al., 2014).Yet apart from one early study in birds (Kacelnik, 1979), very little work has investigated how animals explore under different time horizons.

More generally, relatively few studies have investigated how animals, in particular rodents, make explore-exploit decisions. To study such behavior, most rodent explore-exploit studies use a reversal learning paradigm (RLP). In the reversal learning design, animals choose between two options where one is better than the other. These can be options with high vs low physical costs (Beeler, Daw, Frazier, & Zhuang, 2010), options with large reward and short delay vs small reward and long delay (Laskowski et al., 2016), or binary reward options with high vs low probabilities (Chen, Knep, Han, Ebitz, & Grissom, 2021; Cinotti et al., 2019; Parker et al., 2016; Verharen, den Ouden, Adan,

& Vanderschuren, 2020). As animals explore the two options, they will eventually converge to the better one and keep exploiting it, until the outcomes of the two options are swapped. Deviating from the previously exploited option after the switch of their outcome is considered exploration in these tasks. Such reversal learning paradigms however have several limitations. Firstly, both good and bad outcomes should occur in exploration. However, in reversal learning, after the reversal point, “exploring” the previously suboptimal option will always lead to a better outcome. Thus, exploration is confounded by simply abandoning a currently bad option. Secondly, it is in general not possible to separate different drives and heuristics for exploration in reversal paradigms. For example, to study directed exploration, we need to measure how choices are biased towards the more uncertain option. However, uncertainty is implicit in RLP in that the less chosen option is more uncertain. Since the less chosen option is usually also the option with a lower estimated value, value and uncertainty are confounded in RLP. Thirdly, most of the tasks mentioned above are implemented in operant boxes that are not natural environments for a rat and hence may not engage the decision circuitry fully. As pointed out recently, head-fixed monkeys exhibit a risk preference opposite to that of freely moving monkeys using the same task, suggesting that decision making may be directly influenced by the physical constraints of the experimental paradigms (Vodicka et al., 2019). One of the most fundamental and natural behaviors of rats is spatial navigation. It is unknown how rats would behave in a setting in which the explore-exploit dilemma taps into their spatial navigation abilities. There is also a gap between the human and rodent literature in our understanding of the explore-exploit decision processes. The complexity of the tasks and their quantifications are different across species, and whether similar heuristics are in play in humans and rodents remain an open question.

In this paper, we designed a rodent version of an exploration task similar to the human “Horizon Task” used in Wilson et al, 2014, in which rats explore under different time horizon conditions. The objective of our experiment is two-fold. The first objective was to address the limitations in reversal learning paradigms as mentioned above. Specifically, in our design, the three parameters: *the exploit value*, *the explore value* and *the uncertainty* are manipulated independently to resolve the confounds among choice, uncertainty, and value in the RLP design. Consequently, using Bayesian modeling, we are able to separate directed exploration from random exploration in rodents. Furthermore, our rat version was designed in an open field maze where rats can make explore-exploit decisions by navigating, which is more ethologically naturalistic than using operant boxes. The second objective of the current study was to examine how rodents explore in different time horizon conditions. A similar version of the rodent exploration task was run in human subjects to directly compare, using the same behavioral measures, the similarities, and differences in horizon adaptive exploration between humans and rats.

Methods

Animals

Six Brown Norway rats were used in these experiments. All rats were male between 6 and 7 months old at the start of the experiment. All rats were housed under reverse 12:12 light cycles. Rats were food restricted to 85% of their ad libitum body weight but were not water restricted. All animal procedures were approved by the IACUC of the University of Arizona and followed NIH guidelines.

Human participants

Forty-seven undergraduates from the University of Arizona participated in this study. Two were excluded for being under 18 (in line with the IRB agreement for using the Psychology Department subject pool), leaving 45 participants (14 males, 31 females). In addition, participants who did not perform significantly above chance were excluded. Five were excluded for human experiment 1 (leaving 40 for analysis), and 3 were excluded for human experiment 2 (leaving 42 for analysis). All participants were from the undergraduate psychology subject pool and earned academic credits for their participation in the study. The human experiments were approved by the University of Arizona Institutional Review Board.

Experiments - rats

In this paper, we used a close variant of the human Horizon Task (Wilson et al., 2014). In the rodent version of the Horizon Task, rats were asked to choose between two options that give out different number of drops of sugar water. The reward size from one of the two options is known to the rat, whereas the reward size of the other option is unknown. We assess how rats “explore” the unknown option as a function of time horizon, i.e., the number of choices they have in a game.

Apparatus: The rodent experiments were run on an open field maze that consisted of a circular area (1.5 m diameter) with 8 equidistant feeders at its periphery (B. Jones, Bukoski, Nadel, & Fellous, 2012; B. J. Jones, Pest, Vargas, Glisky, & Fellous, 2015). Each feeder delivered sugar water (150 ul/drop, 0.15g sugar/ml) in the form of computer-controlled drops. A blinking LED was attached to each feeder and if active, indicated that the feeder could deliver a reward is the rat decided to visit it (note the LED was active even if the reward was 0 drop, see below).

Pre-training: Rats were first pre-trained to associate light with reward. Then we pre-train rats on fully guided games, in which they run back and forth between the home base and either one of the target feeders. We then removed the reward at the homebase, so rats are pre-trained to go to the home base to trigger the light at the feeder without getting a direct reward at homebase. Then we introduce free choice trials to rats, after guided choices. 2 lights at the two feeders blinked simultaneously and rats were pre-trained to make a choice between the two feeders. We first pre-train rats to choose between 0 and 1 drops. Once the rat could reliably choose the 1 drop feeder, we introduce different amounts of rewards. We then engaged in pretraining with 1 vs 5 drops, and once the rat showed a preference for choosing the 5 drops feeder, we introduce the full reward schedule (Experiments section below, randomly sample a

reward size from 0 to 5 drops). Rats were pretrained through these phases at different rates based on performance.

Experiments: The experimental sessions were divided into ‘games.’ In each game, only 3/8 feeders were activated in an isosceles pattern (Fig 1A, yellow light bulbs). One feeder was the home base; the two others, equidistant from the home base, were the rewarded feeders**.** The home base was never rewarded, but animals had to reach it to trigger/activate the 2 rewarded feeders. The home base was flanked by two Lego blocks, forcing the animal to start its navigation to the 2 rewarded feeders without directional bias (Fig 1, blue rectangles). At the start of each game, depending on the conditions, the two rewarded feeders were associated with a fixed number of sugar water drops drawn uniformly from 0 to 5 and always gave the same number of drops during that game (Fig 1B). Before making their free choices, rats were guided to one of the rewarded feeders in the first nG trials (i.e., only one LED was blinking, nG=3, ‘Trial 1 cue’ to ‘Trial 3 cue’, Fig 1B). Critically, only one of the two rewarded feeders was cued during the guided trials, leaving the value of the other rewarded feeder unknown to the rat before making free choices. Rats performed versions where nG = 0, 1, or 3 (In cases of nG = 0, rats were not guided to any target feeder and started with a free choice between the 2 rewarded feeders instead.). Fig 1B illustrates the version with nG = 3. From the nG+1st trial, they were cued to make free choices (e.g., the LED of the 2 rewarded feeders blinked simultaneously, ‘Trial 4 cue’ Fig 1B). The guided trials were followed by H free choices between the 2 rewarded feeders. Rats performed versions where H = 1, 6, or 15. Fig 1B illustrates the version with H = 1. After the first game was completed, an 8s increasing sweep tone was played to indicate the start of a new game. The layout was then switched, and the feeder directly opposite to the initial home base was now activated as the new home base in this new game (Game 2 start, Fig 1B). The new rewarded feeders became the feeders opposite to the new home base (Game 2, Fig 1A, 2A). The number of free choices H is also referred to as the ‘horizon’.

Experiment 1: between-session version

In this version, rats are always guided 3 times before a free choice can be made (i.e., nG=3). There are 3 different horizon conditions, the short condition H = 1 in which only 1 free choice is allowed after the guided trials, the long condition H = 6 in which 6 free choices are allowed and the extra-long condition H = 15 in which 15 free choices are allowed (Fig 1C). In the same session, both home bases are associated with the same horizon condition (Fig 1A). Rats performed games of different horizons in blocks of consecutive days before switching to the next horizon condition. H = 1 and H = 6 sessions were run in counterbalanced orders between rats, and H = 15 conditions were run after the H = 1 and H = 6 sessions were completed. For clarity, the transition sessions (the first session after each horizon condition change) were excluded from the analyses. Inclusion of these games in the analysis do not change our conclusions. Six rats participated in this experiment and completed a total of 292 sessions and 4802 games (36664 trials).

Experiment 2: within-session version

In this version, rats performed H = 1 and H = 6 games within the same session. A sound cue was played at each home base visit during each game indicating the corresponding horizon for that game, a low pitch sound was paired with short horizon games (H = 1) and a high pitch sound was paired with long horizon games (H = 6). For 192 out of a total of 218 sessions, one home base was always associated with the short horizon game and sound cue (H = 1), whereas the other home base was always associated with the long horizon game and sound cue (H = 6) (Fig 2A). For the other 26 sessions, long and short horizon games could occur at either home bases signaled by the sound cue. Results from these 26 sessions were analyzed separately in the Supplementary Fig S6. In each session, rats were guided nG times before a free choice could be made, nG = 0, 1, or 3 for different sessions (Fig 2B). For nG = 0, the rat started each game by making free choices without guided trials. In order to compare nG = 0 with nG = 1 games, rats were trained to make H+1 free choices in nG = 0 games. For instance, for a long horizon game with no guided trials (nG = 0, H = 6), the rat would make 7 free choices. In the analysis, we treated the first free choice as if it was guided in nG = 0 games (in other words, the rats guide themselves in the first trial) to contrast it with nG = 1 games in which in the first trial the rat was actually guided (by the single blinking light). Since rats experienced the horizon condition at the two homebases for the first time during the first 2 games of each session, these games were excluded from the analyses. Inclusion of these games in the analyses do not change our conclusions. Four rats participated in this experiment and completed a total of 218 sessions and 5587 games (28436 trials).

Experiment 3: randomized reward

In this control experiment, we always used the long horizon condition (H = 6). However, instead of having a fixed reward for each rewarded feeder within a game, all rewarded feeders gave a uniformly random number of drops between 0 to 5 each time. In this case, there was nothing to learn. The reward contingency was completely random. Rats were guided 3 times before a free choice could be made. Four rats participated in this experiment and completed a total of 20 sessions and 309 games (2781 trials).

Experiments – humans

Experiment 4 – small reward version

In this experiment, participants were sitting in a booth, in front of a computer screen. They were asked to choose between two slot machines (also referred to as bandits, Fig 3A) that gave out a fixed number of reward points uniformly drawn from 1 to 5. The schematic in Figure 3 represents the actual task stimuli. Participants were instructed to maximize the total number of points. The height of the boxes indicated the number of choices allowed in the current game (i.e., the horizon condition, H=2 in Fig 3A) and each row represented a trial. Before participants made their own choices, in the very first trial, they were guided to pick one of the bandits (Trial 1 cue, nG=1, Fig 3A). The option available was cued with a green background color. Participants indicated their choices by pressing an arrow key on the keyboard. Their response was followed by an indication of how many rewards they obtained, the reward of the unchosen option was

not shown and showed up as 'XX' (Trial 1 response, Fig 3A). From the 2nd trial, both bandits were available and participants were free to make their own choices. Rewards from each trial of a game remained on the screen until the end of the game. There were four horizon conditions (H=1, 2, 5, 10 free choices), and games with different horizons were pseudo-randomly interleaved (Fig 3B). Fourty human participants completed a total of 6080 games (33440 trials).

Experiment 5 – large reward version

This experiment was the same as experiment 4 except that the reward points were drawn uniformly from 1 to 100. Results from this version is shown in the Supplementary Fig S3. Forty two human participants completed a total of 6720 games (36960 trials).

Model-free analysis

We computed the following model-free measures of exploration. P(high reward) is the probability of choosing objectively the option with a higher deterministic reward. This measure quantifies 'exploitation'. P(switch) is the probability of switching from the last chosen option, this quantifies 'exploration'. P(unguided) is the probability of choosing the unguided option on the first free choice, i.e., P(switch) on the first free choice. P(unguided) is akin to p(high info) in previous human studies (Wilson et al., 2014). On later free choices, P(switch) could have both a directed and random component. We computed and compared the above measures between humans and rats (Experiment 1, 4), between different horizon conditions (Experiment 2), and between guided and free choices (nG = 0 vs nG = 1 in rats, Experiment 2).

Hierarchical Bayesian analysis

We used hierarchical Bayesian analysis to quantify directed exploration and random exploration for both humans and rats. We focused on humans' and rats' first free choices to be able to compare across horizon conditions.

To model choices on the first free-choice trial, we assumed that subjects made decisions by computing the difference ΔQ between the reward value of the guided option, and an exploration threshold θ. Subjects were more likely to choose the unknown option when $\Delta Q < 0$, and more likely to exploit the guided option when $\Delta Q > 0$. The level of randomness in choices were controlled by a decision noise parameter σ. Both a higher exploration threshold θ and a higher decision noise σ could lead to more exploration. θ is a model-based measure of directed exploration and σ is a model-based measure of random exploration. Specifically, we write

$$\Delta Q = R_{guided} - \theta - b * s_{guided} + \alpha_{LG}\Delta R_{LG} + \alpha_{LS}\Delta R_{LS} \quad (1)$$

$$p(unguided) = \frac{1}{1+e^{\frac{\Delta Q}{\sigma}}} \quad (2)$$

where, R_{guided} is the reward value of the guided option, θ is the exploration threshold, b is the spatial bias, s_{guided} is 1 when the guided side is left and is -1 when the guided side is right, σ is the decision noise, $\Delta R_{LG} = R^{LG}_{guided} - R^{LG}_{unguided}$ is the difference in experienced rewards from the previous game, α_{LG} is the short-term feeder bias coefficient for the last game, $\Delta R_{LS} = R^{LS}_{guided} - R^{LS}_{unguided}$ is the difference in average

rewards from the previous session, α_{LS} is the long-term feeder bias coefficient for the last session.

Each subject's behavior in each horizon (H = 1,6 or 15 for rats and H = 1, 2, 5, 10 for humans) and in each guided condition (nG = 0, 1, or 3 for rats and nG = 1 for humans) was controlled by 5 free parameters, namely the exploration threshold θ, spatial bias b, short-term feeder bias α_{LG}, long-term feeder bias α_{LS} and decision noise σ. Model fitting was done separately for the rat and human experiments. Each of the free parameters was fit to the behavior of each subject using a hierarchical Bayesian approach (Allenby, Rossi, & McCulloch, 2005). The parameters for each subject were assumed to be sampled from group-level prior distributions whose parameters, the so-called 'hyperparameters', were estimated using a Markov Chain Monte Carlo sampling procedure. The hyperparameters themselves were assumed to be sampled from 'hyperprior' distributions whose parameters were set so that these hyperpriors were broad. The specific priors and hyperpriors for each parameter are shown in table 1. Here, the group-level mean of threshold $\theta_{hg} = \frac{a_{hg}^{\theta}}{a_{hg}^{\theta}+b_{hg}^{\theta}}$ and the group-level mean of decision noise $\sigma_{hg} = \frac{a_{hg}^{\sigma}}{a_{hg}^{\sigma}+b_{hg}^{\sigma}}$. Posterior distributions over the exploration threshold θ_{hg} and the decision noise σ_{hg} are shown for each experiment (Fig 8, 9, 10, 12).

Table 1 Model parameters, priors and hyperpriors

Parameter	Prior	Hyperpriors
Exploration threshold θ_{hgs} $\theta_{hgs} = \theta'_{hgs} R_{max}$	${\theta'}_{hgs} \sim \text{Beta}(a_{hg}^{\theta}, b_{hg}^{\theta})$	$a_{hg}^{\theta} \sim U(0.1, 10)$ $b_{hg}^{\theta} \sim U(0.1, 10)$
Decision noise σ_{hgs} $\sigma_{hgs} = \sigma'_{hgs} \Sigma_{max}$	$\sigma'_{hgs} \sim \text{Beta}(a_{hg}^{\sigma}, b_{hg}^{\sigma})$	$a_{hg}^{\sigma} \sim U(0.1, 10)$ $b_{hg}^{\sigma} \sim U(0.1, 10)$
Spatial bias b_{hgs} $b_{hgs} = 2 \cdot b'_{hgs} b_{max} - b_{max}$	$b'_{hgs} \sim \text{Beta}(a_{hg}^{b}, b_{hg}^{b})$	$a_{hg}^{b} \sim U(0.1, 10)$ $b_{hg}^{b} \sim U(0.1, 10)$
Short-term feeder bias α_{hgs}^{LG}	$\alpha_{hgs}^{LG} \sim \text{Beta}(a_{hg}^{LG}, b_{hg}^{LG})$	$a_{hg}^{LG} \sim U(0.1, 10)$ $b_{hg}^{LG} \sim U(0.1, 10)$
Long-term feeder bias α_{hgs}^{LS}	$\alpha_{hgs}^{LS} \sim \text{Beta}(a_{hg}^{LS}, b_{hg}^{LS})$	$a_{hg}^{LS} \sim U(0.1, 10)$ $b_{hg}^{LS} \sim U(0.1, 10)$

* R_{max} is the maximal reward in the experiment. $R_{max} = 5$ for all experiments except for human experiment 5, in which $R_{max} = 100$.
** Σ_{max} is set to be large enough that any reasonable σ falls between 0 and Σ_{max}. We set $\Sigma_{max} = 10$ for all experiments except for human experiment 5, in which $\Sigma_{max} = 100$.
*** b_1 and b_2 are set such that any reasonable b falls between -b_{max} and b_{max}. We set $b_{max} = 5$ for all experiments except for human experiment 5, in which $b_{max} = 50$.
**** h = horizon, g = nG, s = subject (each rat or each human participant)

The model fitting was implemented using the JAGS package (Depaoli et al., 2016, Steyvers, 2011) 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. We used 4 independent Markov chains to generate 80000 samples from the

posterior distribution over parameters (20000 samples per chain). Each chain had a burn in period of 10000 samples, which were discarded to reduce the effects of initial conditions, and posterior samples were acquired at a thin rate of 1. We simulated behavior using 50 uncorrelated random sets of five parameter values and assessed whether our model could successfully recover these simulated parameters. Parameter recovery results are reported in Supplementary Fig S1. In our model, we were able to near perfectly recover the exploration threshold θ (R = 0.99), decision noise σ (R = 0.98), spatial bias b (R = 0.99), and short-term feeder bias α_{LG} (R = 0.96). We have lower performance in recovering long-term feeder bias α_{LS} (R = 0.63). We note that despite these good correlations, our model systematically overestimated α_{LS}, providing therefore upper bound estimates for these parameters.

Results

As with humans, rats transition from exploration to exploitation in the course of a single game.

Both humans (Experiment 4) and rats (Experiment 1) were able to choose the objectively best option (P(high reward), the option with a higher reward magnitude between the two available sugar water feeders for rats, or the slot machine with a higher reward point payout for humans) significantly above chance (50%) for all trials and all horizon conditions (Fig 4A, C). Both humans and rats improved their performances with the number of trials given (Fig 4A, C). Their performances were significantly higher during the last trial of longer horizons compared to shorter horizons ($p < 0.001$ for human, $p = 0.002$ for rats, Fig 4A, C, see also Fig 6A, C). At the last trial of the longest horizon condition (H = 10), humans could achieve an accuracy of 98.4% (Fig 4A) whereas at the last trial of the longest horizon condition (H = 15), rats could achieve an average accuracy of 83.0% (Fig 4C).

Rats switched from the last chosen option at a significantly higher rate on trial 1 (58.7% for H = 6 and 51.8% for H = 15, $p < 0.001$ when compared with trial 2, Fig 4D) and then adopted a more constant and lower rate of switching for later trials (averaged 26.0% for H = 6 and 20.7% for H = 15, Fig 4D). Humans switched more at trial 1 (70.2%, 71.7% and 74.2% for H = 2, 5, 10, Fig 4B) and trial 2 (27.9%, 33.9% and 37.4% for H = 2,5,10), and eventually stopped switching (4.8% and 4.3% at the last trial of H = 5 and H = 10).These results may be partly explained by the deterministic nature of the reward delivery in the experimental design because it only takes a single switch after the guided trials to learn the value of the unguided option. When humans were guided to a good choice (when the unguided reward is objectively lower than the guided reward) and switched on the 1st free choice to find out that the alternative was worse, they immediately switched back on the 2nd choice (Fig 5C, S2). It took several trials for rats to switch back (Fig 5D, S2). The percentage of switching remained higher when guided to a good choice than to a bad choice until the 4th trial (for H = 6, $p = 0.01$). Nonetheless, the fact that P(switch) at trial 1 is higher than later trials clearly separated the 1st free choice from the later trials. Unlike in reversal learning where exploratory behavior manifests over a series of trials, we are able to analyze exploratory behavior by focusing on the 1st free choice in our design. Interestingly, when guided to a good option at first, both rats ($p < 0.01$ for H = 6, $p = 0.07$ for H = 15) and humans ($p < 0.01$

for H = 2 and 5, p = 0.008 for H = 10) showed a better accuracy in later trials compared to when guided to a bad option (Fig 5A, B).

As with humans, rats were able to use prior information to guide exploratory choices.

On the first free choice of each game, participants have only sampled one of the options and thus have no information *from this game* about the payoff of the other option. Thus, if participants were to perform above chance on this first free choice, they *must* have been making use of information from past trials, for example about the prior distribution of possible rewards.

Intriguingly, both humans (Experiment 4) and rats (Experiment 1) performed above chance on the first free-choice trials ($p < 0.001$), both achieving a similar average (66.6% for rats and 69.0% for humans). The fact that the average accuracy was significantly above chance in the first non-guided trial showed that humans and rats used prior information to guide subsequent exploration. In this particular experiment with repeated games, humans and animals were able to assess the relative 'goodness' of the guided target in the current game based on the reward they obtained in previous games.

Their performances in the first free-choice trial were not uniform and displayed a U shape (Fig 6A, C). The accuracy was the highest when they were guided to 0 or 5 drops (or 1 and 5 points for humans), and the lowest when they were guided to more ambiguous reward amounts such as 3 drops. With prior information alone, it is theoretically not possible for humans and rats to choose correctly on the first free-choice trial when guided to intermediate rewards, but through learning in long-horizon games, their performance curves in the last trial were higher ($p < 0.05$ for all drops except 4 in rats, $p < 0.01$ for all drops in humans) and became more uniform across reward sizes (Fig 6B, D).

As with humans, rats can adapt the extent to which they explore based on the reward of the guided choice.

We computed P(unguided), the probability of choosing the option that was not guided when the first free-choice trial occurred (i.e., p(switch) at the first free choice) as a function of the reward size during the guided trials (Fig 7A, C). Two-way ANOVA (Horizon x Guided Reward) showed a significant main effect of guided reward on p(unguided), $p < 0.001$. Like humans (Fig 7A), we found that rats were likely to explore the unguided option if they obtained a low reward during the guided trials (e.g., 0 drops, mean = 95.2% Fig 7C), and were unlikely to explore the unguided option if they obtained a large reward (e.g., 5 drops, mean = 27.5%, Fig 7C). Overall, when guided to the option with an objectively lower reward, rats chose the unguided feeder significantly more (at 74.6%, $p < 0.001$ for H = 1 and 6, $p = 0.01$ for H = 15) on their first free choices, whereas when guided to the option with an objectively higher reward, rats only chose the unguided feeder at 39.9% on their first free choices (Fig 5D, trial 1). Humans chose the unguided option 89.5% on an objectively lower guided reward, and 54.4% on an objectively higher guided reward (Fig 5C, trial 1). Unlike the "win-stay lose-shift"

strategy in probabilistic exploration tasks, both “stay” and “shift” in our task were outcomes of a comparison between the current reward and estimated prior distribution of rewards and were not directly associated with a gain of reward or an absence of reward. Unlike with the reversal learning paradigms in which animals update values gradually and switch to the alternative option after experiencing a stream of bad outcomes, rats in our experiments can make exploratory decisions based on guided reward in a single trial (Fig 2B, nG = 0 or 1, Experiment 2) or after a small number of guided trials (nG = 3, Experiment 1).

As with humans, rats use directed exploration. However, time horizon has opposite modulation on directed exploration in rats and humans.

P(unguided) is akin to the p(high info) measure in previous human research and a model-free way of measuring directed exploration is to contrast P(unguided) across horizon conditions (Wilson et al., 2014). In line with previous research, humans explored the unguided option significantly more in long horizons than in shorter ones ($p < 0.001$, Fig 7B, Experiment 4, $p < 0.001$, Fig S3G, H, Experiment 5). However, for rats, we did not observe a significant difference in P(unguided) for different horizons in Experiment 1 ($p > 0.05$). To properly quantify directed exploration and random exploration, we turned to modeling. Posterior distributions over the group-level means of exploration threshold θ and decision noise σ for both humans and rats are shown in Figure 8A and E, the subject-level estimates of the parameters θ and σ are shown in Figure 8B and F. The posterior distributions of other parameters on spatial bias and feeder bias from previous games are shown in Figure S4. For humans, in line with the model-free result, we observed a significant increase of threshold as horizon increases ($p < 0.001$, Fig 8A, B, $p < 0.001$, Fig 3J), compatible with previous findings in the human horizon task (Wilson et al., 2014). In other words, in longer horizons, humans use more directed exploration in their first free choices than in shorter horizons. Again, we did not observe a significant effect of horizon on threshold in rats at the subject level ($p > 0.05$). However, both the model-free measure p(unguided) and the model-based measure θ showed the opposite trend (Fig 7D, Fig 8F) compared to humans. We will reexamine this in a more controlled Experiment 2 below.

While exploration threshold θ is theoretically tied to directed exploration, decision noise σ is tied to random exploration. Our task has some limitations when studying the horizon modulation of random exploration. Decision noise is consistently small in all horizon conditions for humans (Fig 8C, D). This may arise from the fact that rewards only take 5 different values (1 – 5) and are deterministic, in contrast to the stochastic rewards ranging from 1-100 in the human Horizon Task (Wilson et al., 2014). In human Experiment 5, the rewards are deterministic but range from 1 to 100. Decision noise in longer horizons (H = 5, 10) are significantly higher than decision noise (H = 1, 2) in shorter horizons ($p < 0.01$, Fig S1K, L), which is in line with the horizon adaptive random exploration reported in human studies (Wilson et al., 2014). The deterministic nature of the task seemed to limit the use of random exploration by humans and animals in the 1st free choices. In the 0-5 reward sizes version of the task, we were not able to detect significant horizon differences in random exploration in either humans ($p > 0.05$, Fig 8C, D) or rats ($p > 0.05$, Fig 8G, H).

One critical difference between the human Experiment 4 and the rat Experiment 1 is that human performed all horizon conditions within a single session, whereas rats had to perform the different horizon conditions on different days. Third variables such as the amount of training a particular rat was exposed to before each condition, weight of the rat, etc. could not be controlled and therefore could influence how they explore across different horizon conditions, and make it difficult to detect horizon dependent changes in exploration in Experiment 1. Furthermore, the within-session version may make the difference between horizon conditions more salient to the rats. To make a fairer comparison, as a result, in Experiment 2, we trained rats to run two horizon conditions H = 1 and H = 6 within the same session, where one home base was always associated with short-horizon games (H = 1) and the other home base was always associated with long-horizon games (H = 6). In this alternate design, there is therefore no confound of learning/training effect or weight-related motivation effect.

In Experiment 2, we showed that regardless of the number of guided trials, the model-free measure P(unguided) was significantly lower for Horizon 6 compared to Horizon 1 (Fig 9A, B). Through a two-way ANOVA analysis (Horizon x nG, i.e., the number of guided choices), we found a significant main effect of horizon on P(unguided) with $p < 0.01$. Using the model, we confirmed that regardless of the number of guided trials, exploration threshold θ for H = 6 was significantly lower than H = 1 ($p < 0.001$, Fig 9C, D). By computing the posterior distribution over the differences in exploration threshold between horizons $\Delta\theta = \theta(H = 6) - \theta(H = 1)$, we found that the percentage of samples that $\Delta\theta < 0$ was 96.2%, 89.3% and 75.8% for nG = 0, 1, and 3 respectively (Fig 9G). On the other hand, decision noise σ remained unchanged for H = 1 vs H = 6 ($p > 0.05$), regardless of the number of guided trials (Fig 9E, F). By computing the posterior distribution over the differences in decision noise between horizons $\Delta\sigma = \sigma(H = 6) - \sigma(H = 1)$, we found that the percentage of samples that $\Delta\theta > 0$ was 53.8%, 51.1% and 44.2% for nG = 0, 1 and 3 respectively (Fig 9H).

Moreover, we performed a variant of Experiment 2 in which we used low-pitch vs high-pitch sound cues to signal the horizon condition. The sound was played before the start of each game and during the guided trials to cue the rat to the horizon condition of the current game. The motivation for doing this was that all horizons were interleaved in the human version whereas they were alternated in Experiment 2 when each home base was tied to a specific horizon condition. With the sound cue, we could interleave the horizon conditions pseudo-randomly in rats as in the human version. Within a session, each home base could be associated with different horizon conditions. Again, we found that exploration threshold decreased as a function of horizon whereas decision noise remained unchanged (Fig S6). The fact that there was still a behavioral difference between games of different horizon conditions using only sound cues shows that rats can associate sounds with different time horizon conditions, which can be useful for future task developments.

Rats explore more in more volatile environments, but volatility alone does not account for the horizon adaptive exploration in rats.

The Horizon Task (Wilson et al., 2014) was originally designed in humans to assess exploratory behavior in terms of planning: In longer time horizons it is more beneficial to explore because there is a longer time (i.e. more trials) to benefit from the information gained from exploration. In longer time horizons, the environment is also more stable and less volatile, meaning the rewards from the two options will remain predictable for a longer time before changes occur. As a result, instead of planning rationally, rats may simply adapt the extent to which they explore based on the volatility of the environment and explore more when the environment is changing more frequently (shorter horizon). This may account for the opposite dependence of directed exploration on the horizon in rats compared to humans.

In order to test this hypothesis, instead of giving deterministic rewards that were fixed and learnable for the two rewarded feeders, in Experiment 3, each feeder gave an independently random reward that was sampled uniformly between 0 and 5 drops each time. In other words, the rewards of the two feeders were not learnable and changed independently from trial to trial, from game to game. The time horizon was always set to H = 6. In this version, since there was no information that could be learned and the rewards were random, the rat's accuracy was at chance at 54.3%. Possibly due to overtraining in Experiment 1 and 2, after the guided choices, rats still explored the unguided option significantly more on the first free choices than on subsequent ones ($p = 0.01$), suggesting that the novelty of the unknown feeder itself in addition to the potential better reward may drive exploration (Fig 10A). Critically, the percentage of exploring the unguided feeder was higher compared to P(unguided) in the constant reward scenario in Experiment 1, especially when the guided reward size was high (Fig 10B). For later choices, the overall level of switching was also slightly higher compared to that of the constant reward condition in Experiment 1 (Fig 10A). In a more volatile environment, rats increased their switching rate. This could account for the horizon difference in P(switch) in Figure 4D where there was a significantly lower rate of switching in H = 15 compared to H = 6 ($p < 0.001$), possibly due to the fact that the environment was less volatile in the H = 15 case. This difference in P(switch) could not be attributed to directed exploration and could arise from random exploration.

Despite that volatility could potentially account for random exploration in later trials, importantly, volatility alone cannot account for the opposite dependence of exploration threshold on horizon in rats. In Experiment 3, we observed an increase in both the exploration threshold (Fig 10C, D, $p < 0.01$) and decision noise (Fig 10E, F, $p < 0.01$) in the random reward condition compared to the constant reward condition. Since only exploration threshold (not decision noise) changes with horizon in Experiment 2. This suggests that the horizon difference we observed in Experiment 2 cannot be attributed to volatility. We note that exposure to Experiment 1 and 2 may contribute to the high threshold (more switching) in Experiment 3, but it is unlikely that the high decision noise parameter arises from any carry-over effects from Experiment 1 and 2. Since increasing volatility increases decision noise, it still holds from Experiment 3 that volatility itself does not account for the horizon differences we observed in rats.

Self-guided exploration is treated intrinsically differently than cue-guided choices in rats.

We investigated whether self-driven exploration was any different from cue-guided exploration. Did rats behave differently if they were guided by light cues on the first trials, or if they were instead invited to choose freely? Specifically, in separate weeks and between sessions, rats performed both a version in which they were guided to one feeder once before freely choosing between the 2 options (Guided condition, nG = 1 in Experiment 2), and a version in which they started off with 2 options to choose from (Free choice condition, nG = 0 in Experiment 2). In the analysis, we treated the first choice in the Free choice condition as if it were guided (i.e., self-guided by the rat itself, instead of by the blinking LED), and treated the second choice as choice number 1 (Fig 11).

Perhaps counter-intuitively, we found that overall, rats performed significantly better ($p < 0.01$) if the first trial was a free self-guided choice than when they were guided by a light cue (Fig 11A). Moreover, rats explore differently in the Free condition compared to the Guided condition. When rats were cue-guided, they switched significantly more on the first free choice than in subsequent choices as in other variants of the task (Fig 11B, 4B, D). However, when they chose freely, the 2nd choice did not differ from subsequent choices anymore ($p > 0.05$), and rats seemed to have kept a steady rate of switching throughout the game, at a rate higher than the Guided condition ($p < 0.001$, Fig 11B). Rats switched significantly more on the first free choice in the Guided condition compared to the Free choice condition ($p < 0.001$, Fig 11D), and they switched more regardless of the guided reward and the horizon condition (Fig 11C).

We have shown earlier that the exploration threshold was lower in H = 6 than with H = 1, regardless of whether the first trial was guided or not (Fig 9C, D) and decision noise remained unchanged (Fig 9E, F). Now we ask whether exploration threshold and decision noise differ in the Guided vs Free choice condition. For both horizon H = 1 and H = 6, exploration threshold in Free-choice condition was lower than in the Guided condition (Fig 12A). By computing the posterior distribution over the differences in exploration threshold between conditions $\Delta\theta = \theta(Free) - \theta(Guided)$, we found that the percentage of samples that $\Delta\theta < 0$ is 99.5%, and 98.4% for H = 1 and H = 6 respectively (Fig 12B). In other words, when rats were cue-guided, they explored more in the first free choice. Decision noise did not change significantly in the Guided condition vs Free choice condition (Fig 12C), by computing the posterior distribution over the differences in decision noise between conditions $\Delta\sigma = \sigma(Free) - \sigma(Guided)$, we found that the percentage of samples that $\Delta\sigma < 0$ is 62.0%, and 64.4% for H = 1 and H = 6 respectively (Fig 12D).

To our knowledge, such comparisons between self-guided and cue-guided 1st choice in explore-exploit tasks have done been done on humans yet, and as such, these results therefore predict human performance.

Finally, we note that there are no sex differences on horizon adaptation of directed exploration in humans ($p > 0.05$, Experiment 4). Recent work has also found no evidence for sex differences in either directed or random exploration in the original

Horizon Task (Smith et al., 2021). For these reasons, we did not include a population of female rats at this stage, although we recognize that further work on sex differences in the explore-exploit tasks will be needed.

Discussion

In this study, we investigated the exploratory behaviors of rats using a new model of the Horizon task. We addressed the limitations of previous rodent studies by designing a novel open-field task in which rodents choose between two locations that offered different amounts of rewards. To dissociate the uncertainty in the estimation of value from the ambiguity of an unknown novel option, we manipulated the magnitudes of rewards rather than the probabilities of their delivery. Rather than reversing (or drifting) the reward conditions at the same set of locations/feeders as in traditional reversal learning paradigms, we were able to use two sets of different locations alternatively as new games start and use independent rewards between games. As a result, we were able to dissociate exploration for information from abandoning a currently bad option (which are confounded in reversal learning paradigms). In our design, rats were guided to one of two feeder locations first, and the extent to which they explored the other unguided feeder location in their first free choice was compared across horizons. This measure is an equivalent of the model-free measure that is related to directed exploration in previous human studies (Wilson et al., 2014). In addition, rats performed the task in both a short and a long horizon condition to assess whether they explored differently in different time horizon contexts. Finally, we recruited human subjects to perform a version that was comparable to the rat task, and we compared the performance between humans and rats.

We showed that similarly to humans, rats were able to use prior information about the distribution of rewards to guide future exploration. Rats explored the unguided option more in their first free choice when the guided reward size was low compared to when the guided reward size was high. This is very similar to the win-stay/lose-shift strategies in reversal learning where animals choose to switch more when the exploit value is low and less when the exploit value is high. However, unlike in reversal learning where a “win” or a “loss” is computed by comparing the current reward with the expected value, in our design, a “win” or a “loss” is computed by comparing the current reward (or estimated value of the current option) with the estimated distribution of rewards using prior information. In order to assess whether the exploit value was low or high, instead of using short-term memory to recall the value of the exploit option before reversal within the same game, rats had to use their long-term memory from previous games and sessions in previous days to estimate the distribution of possible rewards. We showed that rats were indeed able to incorporate prior information in guiding their exploration.

In this study, we were able to separate directed exploration from random exploration. Both rats and humans switched significantly more at the first free choice than on subsequent choices. We further quantified directed and random exploration using hierarchical Bayesian modeling in both the rat and the human datasets. In line with previous human studies, humans have an increased exploration threshold (explore more) in longer horizons. Unlike humans however, rats showed an opposite adaptation of exploration threshold to the time horizon. For random exploration, with deterministic reward size in a small range (0 – 5), we did not observe adaptations of random exploration in either humans or rats in this task. With a deterministic larger reward range (1 – 100), in human Experiment 5, we did observe some level of random exploration

(Fig S1K, L), but not as strong as in the probabilistic version of the human Horizon Task (Wilson et al., 2014). This can be considered a limitation of the current design and may also be a limitation of using rats and their limited ability to assess and discriminate between large number of levels of rewards. exploration, our task is better suited to study directed exploration rather than random exploration. Variation in the probability of reward deliveries, and smaller (or different types) of rewards might be possible way to improve our task in the future, to better assess random exploration.

As with optimal agents, human have a higher level of directed exploration in longer time horizons since the value of the information gained through exploration is high if the remaining time horizon is long. Interestingly, rats have instead a lower level of directed exploration in longer horizon. Our results do not fully explain this phenomenon. We first speculate why rats do not increase threshold as humans do. There may be an 'optimize vs satisfice' discrepancy in humans vs rats due to the nature of the rewards received. The utility of 1 to 5 drops and the cost of actions are different for humans and rats. Humans receive hypothetical points with relatively effortless keypresses on a computer keyboard, whereas rats had to physically travel on a meter-long maze to earn sugar water. The efforts humans spent in making the decision was small. As a result, they over-explored to find out the best possible action. It costs little for humans to optimize by testing if the alternative reward is 5 when the guided reward is 3, however rats may risk running for 0 rewards by visiting the unguided feeder when they are guaranteed to have 3 drops of sugar water in the guided feeder. Rats therefore likely under-explored (directed exploration) to secure a satisfiable amount of return for each visit. In our data, rats had lower exploration thresholds compared to humans (Fig 8A, E). The drive to explore may therefore not be to optimize for rats, but to satisfice. Exploring more in longer horizon may be an optimal way to explore, but optimization may only offer marginal gain in total rewards, and it may not be worth for rats to achieve optimality in this task. To properly test this hypothesis would require future experiments that could involve effortful decision making in humans (e.g., physically walk from one building to another on campus) or potentially run identical tasks in long/short distance (or maze vs boxes) in rats. We would predict that rats in an effortful setup will be less willing to engage in directed exploration than rats in an effortless setup.

Why do rats show a decrease in exploration threshold with horizon? In short horizon, without fully understanding the structure of the task, rats may perceive the time horizon in terms of the volatility of the environment, and thus explore more in a more volatile condition (the short horizon condition). Experiment 3 supported this view, in that, by having random rewards, rats explored more compared to the constant reward case in Experiment 1 (Fig 10). However, volatility does not selectively increase threshold, but also increases decision noise. This is compatible with the theory that relative uncertainty correlates with directed exploration whereas total uncertainty correlates with random exploration (Gershman, 2019). In a more volatile environment, the uncertainty of both options increase, thus both total uncertainty and relative uncertainty increase, resulting in the increase in the threshold as well as the decision noise. Since rats selectively increased threshold without increasing decision noise in longer horizon condition, volatility alone cannot account for the behavior of the rats. Another possibility is that rats may be biased towards feeders that were more rewarding in past games. Indeed, we

observed that rats are slightly biased towards the feeder that had high rewards in the past (Fig S5). In Experiment 2, rats spent more trials on long horizon games (H = 6) compared to short horizon games (H = 1). As a result, rats might develop a stronger bias towards rewarding feeders in long horizon games. We did see a significantly larger feeder bias in long horizon games in Experiment 2 (Fig S7). However, feeder bias should in principle affect the guided feeder and unguided feeder equally and it is not clear how it might bias exploration. Parameter recovery results suggested that our model estimates the upper bound of these feeder biases (Fig S1). It is therefore unlikely that the horizon difference in exploration threshold arises from feeder biases. Lastly, a longer horizon means that there were many opportunities to explore the unguided option, making it less urgent to explore on the first trial compared to a shorter horizon. Future study is needed to compare these alternative algorithms that rats may use in explore-exploit tasks and explain why rats decrease their decision threshold with time horizon.

Nevertheless, we note that rats can adapt the level of directed exploration to the time horizon. The use of horizon context to explore requires (possibly irrational) planning and model-based reasoning (a mental model of the environment that reflects the time horizon). Win-stay/lose-shift strategies which are effective in solving reversal learning problems do not work in dealing with horizon changes. The win-stay/lose-shift strategy is solely dependent on experienced and estimated rewards and does not by itself adapt to time horizon changes. To the authors' knowledge, horizon adaption of exploration has only been examined in very limited species (humans, Wilson et al, 2014; great tits, Kacelnik, 1979). It remains an open question as to whether other species can adapt exploration to time horizons.

In addition, we believe our design has advantages in serving as a potential behavioral model in studying the neurophysiological mechanisms underlying real-time explore-exploit decisions and its neural substrate. In the reversal learning paradigm, the level of exploration had to be evaluated on the course of several trials, therefore the exact timing of "exploration" decision is difficult to estimate. In our design, however, exploration can be seen in a single trial (visiting the unknown option), which is advantageous.

Finally, we observed an interesting difference in the exploration strategy between when the first choice was self-driven vs cue-guided (a condition that was not studied in humans in this task). This suggests a different neural mechanism underlying voluntary vs guided learning. Rats explored the unvisited feeder more when they were guided first, but this was not observed when the first choice was made freely by themselves. A similar phenomenon was recently reported in a human explore-exploit study (Sadeghiyeh, Wang, & Wilson, 2018). More generally, learning differences in active and passive version of the same tasks have been shown in a number of tasks (Gureckis & Markant, 2012; Markant & Gureckis, 2014; Markant, Settles, & Gureckis, 2016).Our rat model has therefore the potential of probing the differential neural mechanism underlying active vs passive learning. Overall, our novel design provides a fruitful behavioral paradigm to investigate explore-exploit tradeoffs in future electrophysiological studies and suggest new avenues for further comparisons between rats and humans.

Figure Captions

Figure 1: A: In rat experiments, the 2 sets of home bases, lights and feeders were used alternatively between games. B. Timeline of the rat experiments. Rats were trained to start each trial by reaching the home base (HB, no reward). They were then given a small number (here nG = 3) of guided trial (e.g., Trial 1-3, one blinking light, here 1 drop). Subsequent trials consisted in 2 simultaneously blinking lights (here Horizon = 1). The end of a game was signaled by a sweeping tone and a change of home base. C. Horizon conditions (the number of free trials) in Experiment 1, the number of guided trials are always 3 in Experiment 1.

Figure 2: A. In rat Experiment 2 (except for the sound cue variant), horizon conditions are alternated between games. B. Task conditions (nG x Horizon) in Experiment 2. The number of guided trials is 0, 1 or 3 trials, the number of free trials (horizons) are either 1 or 6 trials. Note that when nG = 0, there are H + 1 free trials and the first of these are treated as a (self-guided) guided trial.

Figure 3: A. Timeline of the human experiments (Experiment 4 and 5): Human subjects were presented with a 2-armed bandit display of explicit time horizon (here Horizon = 2). They were guided to the first bandit and obtained a visible reward (here 3 points). Subsequent trials consisted in simultaneously colored squares indicating free choices between the two bandits. B. Task conditions in Experiment 4 and 5. There are four horizon conditions H = 1, 2, 5 and 10.

Figure 4: A and C. Probability of choosing the option with the highest reward for humans (A) and rats (C). B and D. Probability of switching from the last chosen option in free choices for humans (B) and rats (D). The human data is from Experiment 4 and the rat data is from Experiment 1.

Figure 5: Probability of choosing the option with the highest reward, i.e., p(high reward) and probability of switching from the last chosen option in free choices, i.e. p(switch), split up by whether the guided option is the objectively better option, for humans (A, C) and rats (B, D). Data from Experiments 1 (rats) and 4 (humans). High (low) contrast colors indicate games where the guided choices where in fact the best (worst) one of the two available choices.

Figure 6: Probability of choosing the option with the highest reward in the 1st and last free choice as a function of guided reward size. A and C. Probability of choosing the high reward option in the 1st choice of each horizon as a function of guided reward size for humans(A) and for rats (C). B and D. Probability of choosing the high reward option in the last free choice of each horizon as a function of guided reward size for humans (B) and for rats(D). Experiment 1 (rats) and 4 (humans).

Figure 7: A and C. Probability of exploring the unguided option (i.e., P(switch) at trial number 1) in the 1st free choice as a function of guided reward size for humans (A) and

for rats (C). B and D. Probability of exploring the unguided option as a function of horizon for humans (B) and for rats (D). Experiment 1 (rats) and 4 (humans).

Figure 8: Model-based estimates of exploration threshold and decision noise for humans (A-D) and rats (E-H). A and E: Posterior distributions over the group-level means of exploration threshold θ. B and F: Means of the subject-level estimates of exploration threshold θ as a function of horizon. C and G: Posterior distributions over the group-level means of decision noise σ. D and H: Means of the subject-level estimates of decision noise σ as a function of horizon. Experiment 1 (rats) and 4 (humans)

Figure 9: Differences in directed and random exploration in H = 1 vs H = 6 in rats. A. Probability of exploring the unguided option vs guided reward size separated by horizon condition, for nG = 0, 1 and 3 respectively. B. Average P(unguided) by horizon (blue is H = 1, red is H = 6) and nG. C. Posterior distributions over the group-level means of exploration threshold $\theta(H = 1)$ and $\theta(H = 6)$ for nG = 0, 1 and 3. D. Means of the subject-level estimates of exploration threshold θ as a function of horizon. E. Posterior distributions over the group-level means of decision noise $\sigma(H = 1)$ and $\sigma(H = 6)$ for nG = 0, 1 and 3. F. Means of the subject-level estimates of decision noise σ as a function of horizon. G. Posterior distribution over the group-level means of $\theta(H = 6) - \theta(H = 1)$. H. Posterior distribution over the group-level means of $\sigma(H = 6) - \sigma(H = 1)$. (Experiment 2). nG= number of guided trials.

Figure 10: Effects of volatility on exploration by comparing random vs constant reward conditions (Experiment 3). A. Probability of switching from the last chosen option as a function of trial number. B. Probability of exploring the unguided option in the 1st free choice as a function of guided reward size. C. Posterior distributions over the group-level means of exploration threshold θ. D. Means of the subject-level estimates of exploration threshold θ. E. Posterior distributions over the group-level means of decision noise σ. F. Means of the subject-level estimates of decision noise σ.

Figure 11: Differences in exploration in Guided vs Free choice condition (Experiment 2). At the start of a game, rats were given one guided trial (1 light blinking, Guided condition) or a free choice instead (2 lights blinking, self-guided condition). A: Probability of choosing the option with the highest reward in free choices after the guided trial vs after the first free choice for H = 1 and H = 6. B: Probability of switching from the last chosen option in Guided vs Free condition for H = 1 and H = 6. C: Influence of reward size during the first trials (Guided or Free choice) on exploration. D: Average percentage of exploring the unchosen option in Guided vs Free choice condition by horizon, blue is H = 1, red is H = 6, lighter color is Free choice condition and darker color is Guided condition.

Figure 12: Model estimates of exploration threshold and decision noise in Free choice condition vs Guided condition. A and C. Posterior distributions over the group-level means of exploration threshold θ (A) and decision noise σ (C). B. Posterior distribution

over the group-level means of $\theta(Free) - \theta(Guided)$. D. Posterior distribution over the group-level means of $\sigma(Free) - \sigma(Guided)$.

Figure S1: Parameter recovery of exploration threshold θ, decision noise σ, short-term feeder bias α_{LG}, long-term feeder bias α_{LS} and spatial bias b.

Figure S2: Probability of switching away (from guided to unguided option) and probability of switching back (from unguided option to guided option) in free choices, i.e. p(switch away) and p(switch back), split up by whether the guided option is the objectively better option, for humans (A, C) and rats (B, D). Data from Experiments 1 (rats) and 4 (humans). High (low) contrast colors indicate games where the guided choices where in fact the best (worst) one of the two available choices.

Figure S3: Human Experiment 5 (Rewards range from 1 to 100). A: Probability of choosing the option with the highest reward as a function of trial number. B: Probability of switching from the last chosen option as a function of trial number. C: p(high reward) in the 1st free choice as a function of guided reward size by horizon. D: average p(high reward, 1st choice) by horizon. E: p(high reward) in the last free choice as a function of guided reward size by horizon. F: average p(high reward, last choice) by horizon. G: P(unguided) as a function of guided reward size by horizon. H: average P(unguided) by horizon. I: Model estimates of group-level exploration thresholds. J: Average of subject-level estimates of exploration thresholds by horizon. K: Model estimates of group-level decision noise. L: Average of subject-level estimates of decision noise by horizon.

Figure S4: Posterior distribution over the group-level means of spatial bias b, short-term feeder bias α_{LG}, long-term feeder bias α_{LS} for both humans and rats in Experiment 1.

Figure S5: Rats are influenced by both short-term and long-term feeder bias. Left, Percentage of choosing the unguided feeder in 1st free choice as a function of the experienced reward of the guided feeder from last game in humans (Top) and rats (Bottom). Right, Percentage of choosing the unguided option in 1st free choice as a function of the average reward of the guided feeder from last session in rats. Humans do not have a LS panel since all humans only participated in a single session. NaN refers to cases when the guided feeder was not chosen in the last game that involved it.

Figure S6: Sound cue variant of Experiment 2. In this experiment, the different horizon conditions are cued by either a low-pitch sound (H = 1) or a high-pitch sound (H = 6). Games of different horizons are interleaved. A: P(unguided) as a function of guided reward size. B. Model estimates of exploration threshold. C. Model estimates of decision noise.

Figure S7: Short term feeder bias is larger in long horizon condition. LG (left) and LS (right) coefficients as a function of Horizon (1:blue, 6:red) and nG (number of guided choices, nG = 0, 1, or 3). LG coefficient is significantly larger in H = 6 than H = 1 condition, showing that short term feeder bias (from last game) has a significantly bigger

influence on H = 6 games ($p < 0.001$). This is likely due to that rats spend more trials at H = 6 feeders within a session. There are no differences in long term feeder bias (from last session) between horizon conditions ($p = 0.48$).

Figure S8: Posterior distribution over the group-level means of spatial bias b, short-term feeder bias α_{LG}, long-term feeder bias α_{LS} for rats in Experiment 2. Each row corresponds to one of the parameters, each column corresponds to one nG condition (nG = 0, 1 or 3).

References

Allenby, G. M., Rossi, P. E., & McCulloch, R. E. (2005). Hierarchical bayes models: A practitioners guide. ssrn scholarly paper id 655541. *Social Science Research Network, Rochester, NY*.

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*. doi:10.1016/j.neuron.2011.12.025

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory*. doi:10.1007/s001990050146

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience, 4*, 1-14. doi:10.3389/fnbeh.2010.00170

Bellman, R. (1954). The Theory of Dynamic Programming. *Bulletin of the American Mathematical Society*. doi:10.1090/S0002-9904-1954-09848-8

Chen, C. S., Knep, E., Han, A., Ebitz, R. B., & Grissom, N. (2021). Sex differences in learning from exploration. *Elife, 10*. doi:10.7554/eLife.69748

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific reports, 9*, 1-14. doi:10.1038/s41598-019-43245-z

Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific reports, 11*(1), 1-15.

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

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition, 173*, 34-42. doi:10.1016/j.cognition.2017.12.014

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*. doi:10.1037/dec0000101

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspect Psychol Sci, 7*(5), 464-481. doi:10.1177/1745691612454304

Jones, B., Bukoski, E., Nadel, L., & Fellous, J. M. (2012). Remaking memories: reconsolidation updates positively motivated spatial memory in rats. *Learn Mem, 19*(3), 91-98. doi:10.1101/lm.023408.111

Jones, B. J., Pest, S. M., Vargas, I. M., Glisky, E. L., & Fellous, J. M. (2015). Contextual reminders fail to trigger memory reconsolidation in aged rats and aged humans. *Neurobiol Learn Mem, 120*, 7-15. doi:10.1016/j.nlm.2015.02.003

Kacelnik, A. (1979). *Studies of foraging behaviour and time budgeting in great tits (parus major)* ([PhD thesis].). University of Oxford.,

Kao, M. H., Doupe, A. J., & Brainard, M. S. (2005). {C}ontributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature, 433*, 638-643.

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

Laskowski, C. S., Williams, R. J., Martens, K. M., Gruber, A. J., Fisher, K. G., & Euston, D. R. (2016). The role of the medial prefrontal cortex in updating reward value and avoiding perseveration. *Behavioural Brain Research, 306*, 52-63. doi:10.1016/j.bbr.2016.03.007

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*. doi:10.1016/j.cogsys.2010.07.007

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *J Exp Psychol Gen, 143*(1), 94-122. doi:10.1037/a0032108

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-Directed Learning Favors Local, Rather Than Global, Uncertainty. *Cogn Sci, 40*(1), 100-120. doi:10.1111/cogs.12220

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., . . . Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision*. doi:10.1037/dec0000033

Meyer, R. J., & Shi, Y. (1995). Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. *Management Science*. doi:10.1287/mnsc.41.5.817

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., . . . Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience, 19*, 845-854. doi:10.1038/nn.4287

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

Sadeghiyeh, H., Wang, S., & Wilson, R. C. (2018). Lessons from a “failed” replication: The importance of taking action in exploration. *PsyArXiv. doi, 10*. doi:10.31234/osf.io/ue7dx

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Curr Opin Neurobiol, 55*, 7-14. doi:10.1016/j.conb.2018.11.003

Smith, R., Taylor, S., Wilson, R. C., Chuning, A. E., Persich, M. R., Wang, S., & Killgore, W. D. S. (2021). Lower Levels of Directed Exploration and Reflective Thinking Are Associated With Greater Anxiety and Depression. *Front Psychiatry, 12*, 782136. doi:10.3389/fpsyt.2021.782136

Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2008.11.002

Verharen, J. P. H., den Ouden, H. E. M., Adan, R. A. H., & Vanderschuren, L. J. M. J. (2020). Modulation of value-based decision making behavior by subregions of the rat prefrontal cortex. *Psychopharmacology, 237*, 1267-1280. doi:10.1007/s00213-020-05454-7

Wang, S., & Wilson, R. (2018). Any way the brain blows? The nature of decision noise in random exploration. doi:10.31234/osf.io/rxmqn

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Curr Opin Behav Sci, 38*, 49-56. doi:10.1016/j.cobeha.2020.10.001

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen, 143*(6), 2074-2081. doi:10.1037/a0038199

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

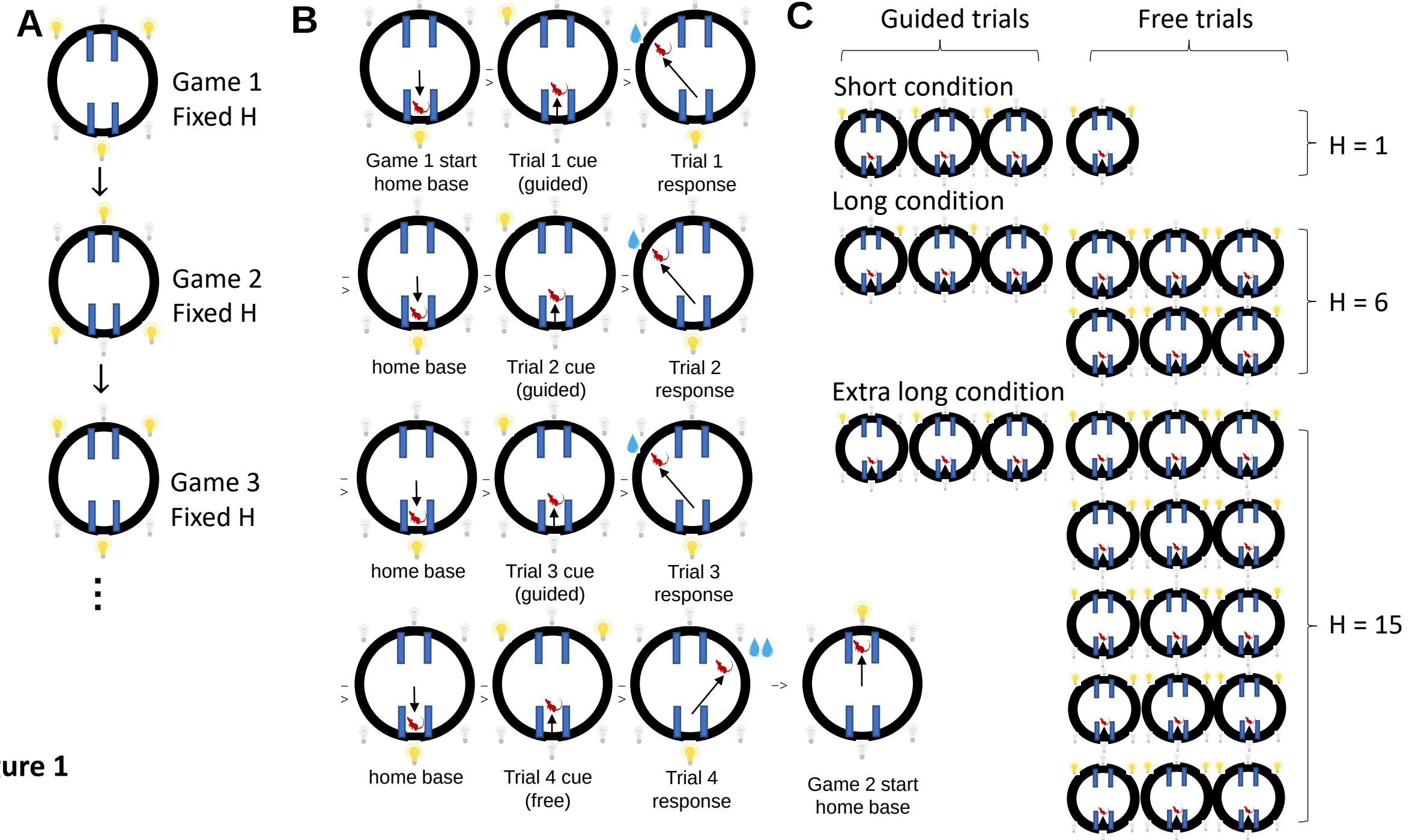

A
Game 1
Fixed H
Game 2
Fixed H
Game 3
Fixed H
B
Game 1 start
home base
Trial 1 cue
(guided)
Trial 1
response
home base
Trial 2 cue
(guided)
Trial 2
response
home base
Trial 3 cue
(guided)
Trial 3
response
home base
Trial 4 cue
(free)
Trial 4
response
Game 2 start
home base
C
Guided trials
Free trials
Short condition
H = 1
Long condition
H = 6
Extra long condition
H = 15

Figure 1

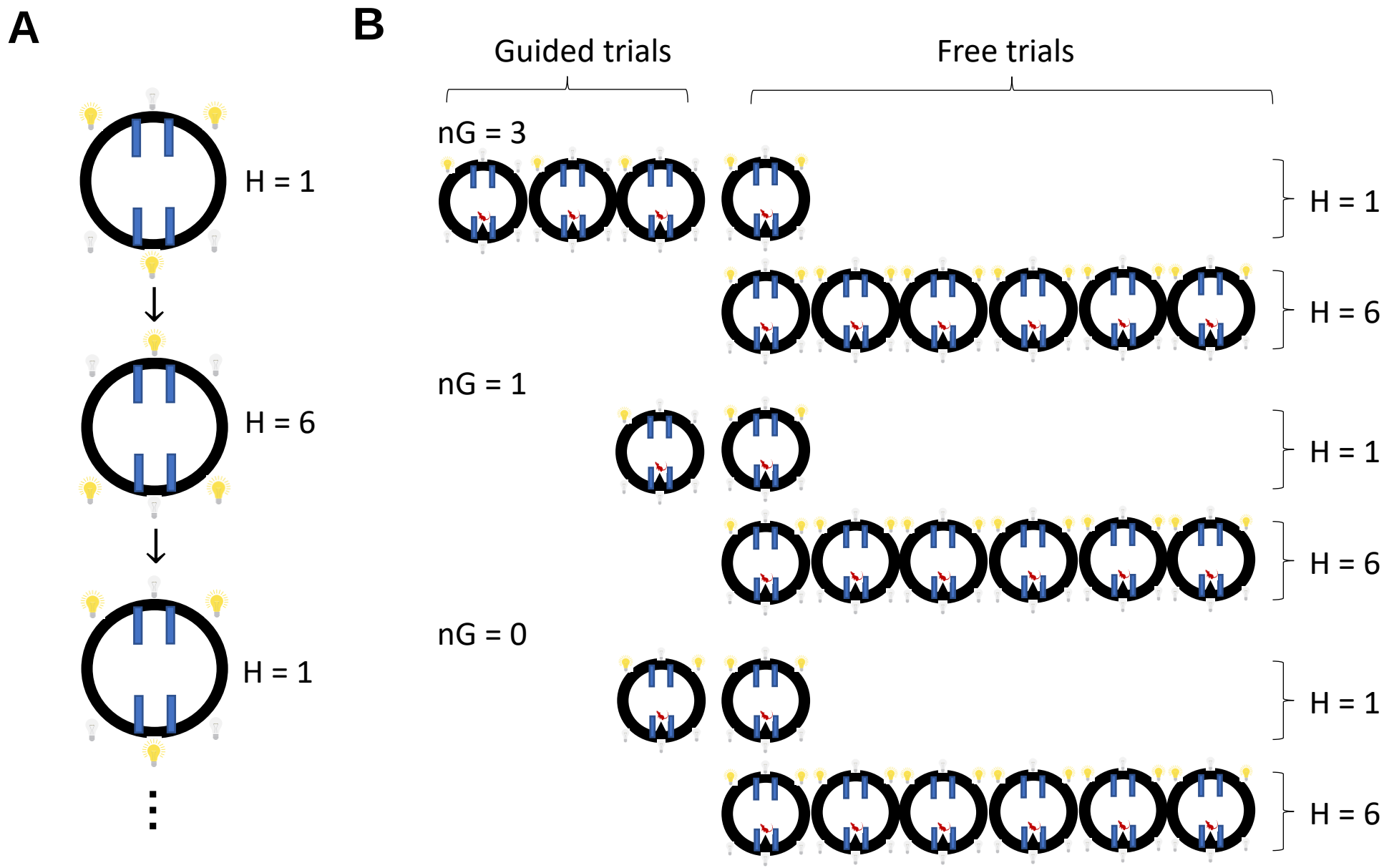

A
H = 1
H = 6
H = 1
B
Guided trials
Free trials
nG = 3
H = 1
H = 6
nG = 1
H = 1
H = 6
nG = 0
H = 1
H = 6

Figure 2

Figure 3

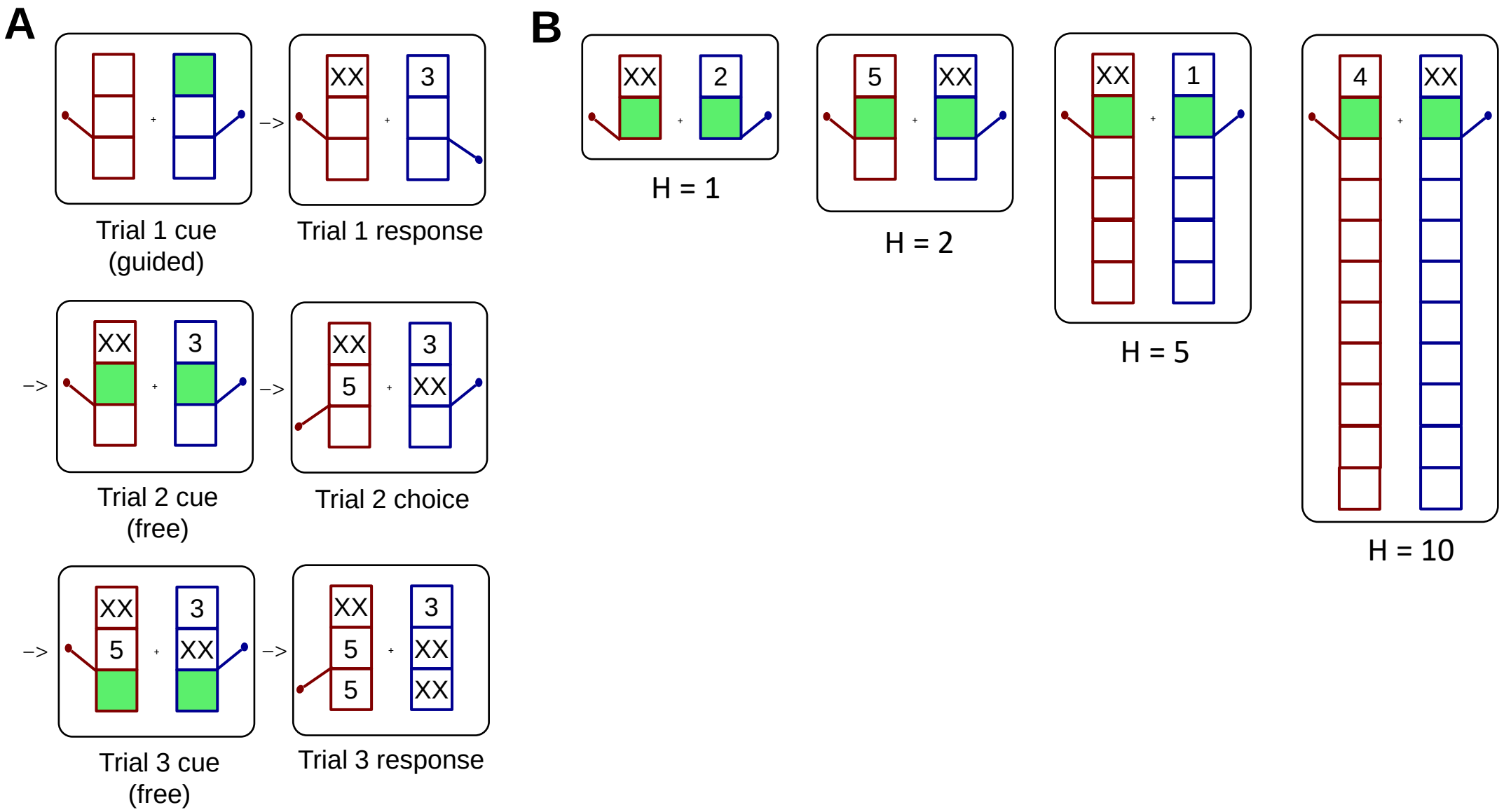

A
XX
3
Trial 1 cue
(guided)
Trial 1 response
Trial 2 cue
(free)
5
Trial 2 choice
Trial 3 cue
(free)
Trial 3 response
B
2
H = 1
H = 2
1
H = 5
4
H = 10

Figure 4

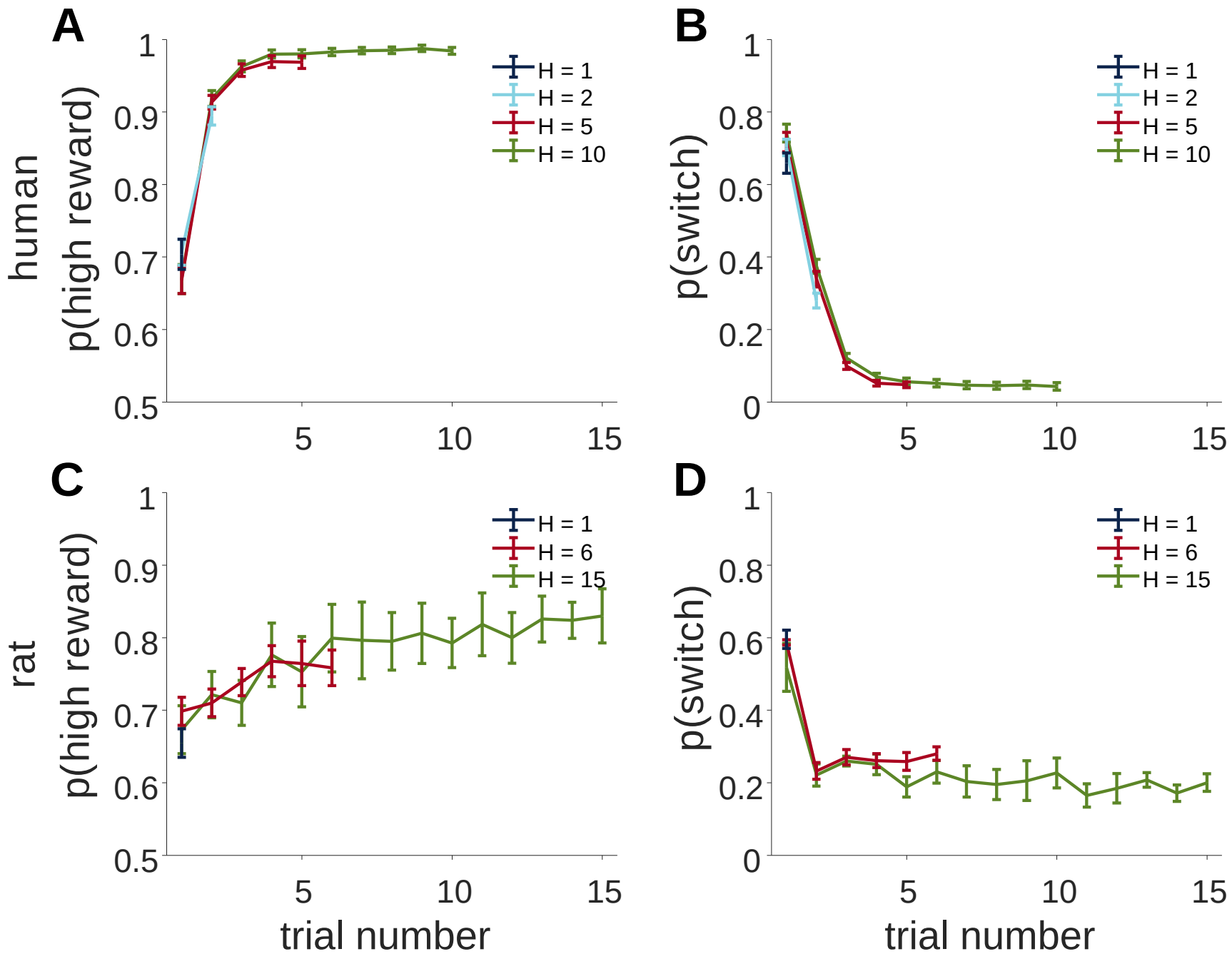

A
human
p(high reward)
1
0.9
0.8
0.7
0.6
0.5
5
10
15
H = 1
H = 2
H = 5
H = 10
B
p(switch)
1
0.8
0.6
0.4
0.2
0
5
10
15
H = 1
H = 2
H = 5
H = 10
C
rat
p(high reward)
1
0.9
0.8
0.7
0.6
0.5
5
10
15
trial number
H = 1
H = 6
H = 15
D
p(switch)
1
0.8
0.6
0.4
0.2
0
5
10
15
trial number
H = 1
H = 6
H = 15

Figure 5

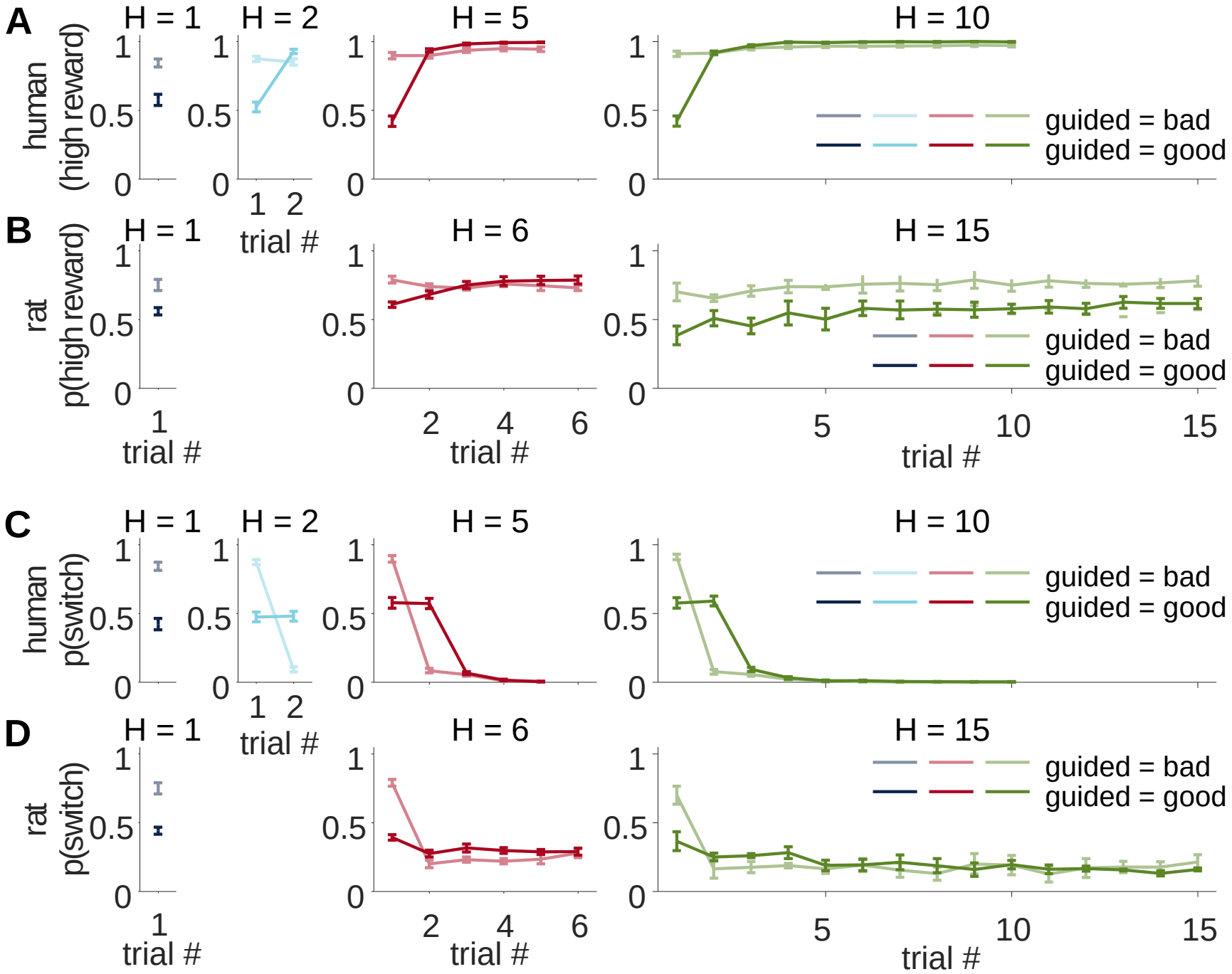

A
H = 1
H = 2
H = 5
H = 10
human (high reward)
guided = bad
guided = good
trial #
B
H = 1
H = 6
H = 15
rat p(high reward)
guided = bad
guided = good
trial #
C
H = 1
H = 2
H = 5
H = 10
human p(switch)
guided = bad
guided = good
trial #
D
H = 1
H = 6
H = 15
rat p(switch)
guided = bad
guided = good
trial #

Figure 6

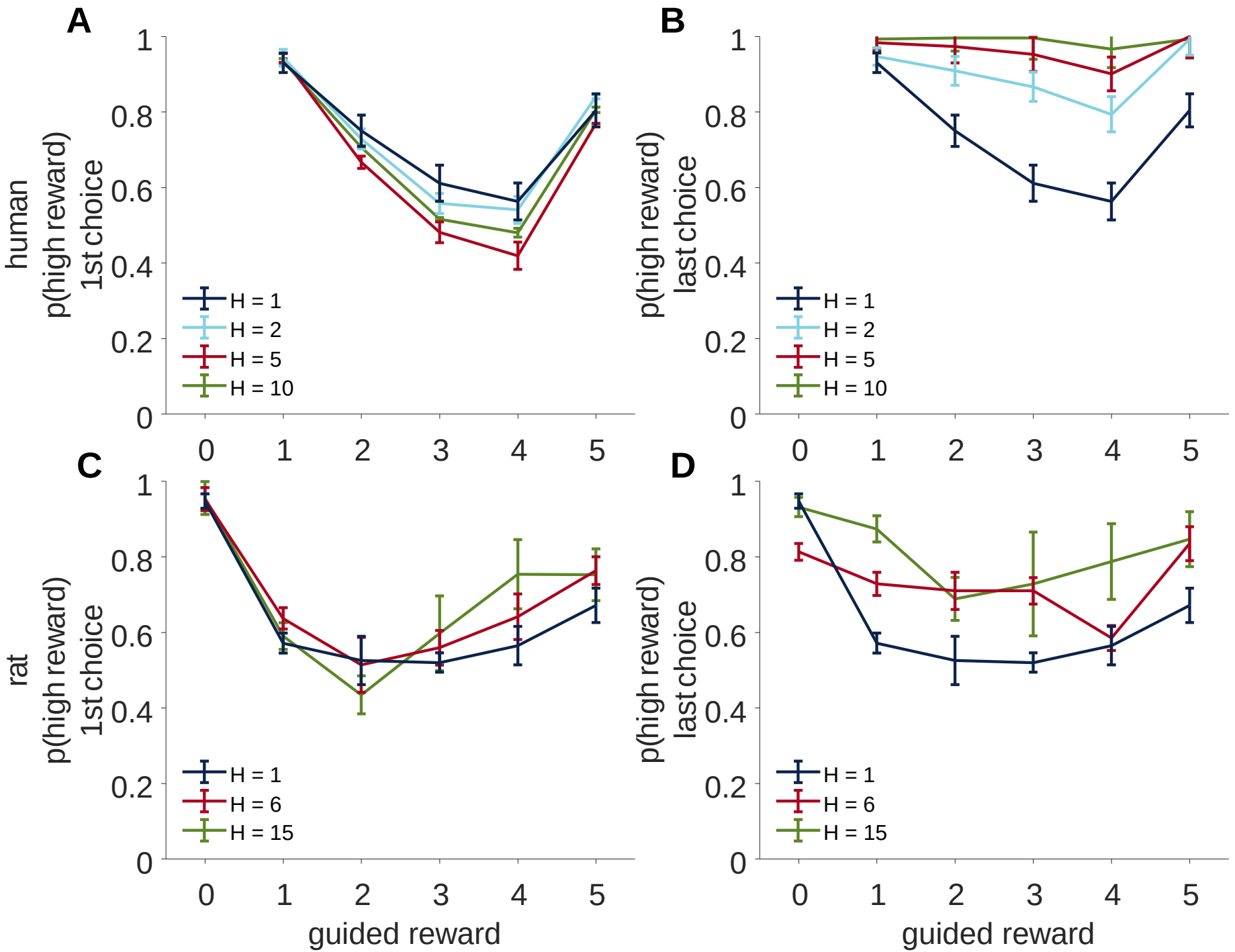

A
B
C
D
human
p(high reward)
1st choice
p(high reward)
last choice
rat
p(high reward)
1st choice
p(high reward)
last choice
H = 1
H = 2
H = 5
H = 10
H = 1
H = 6
H = 15
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5
guided reward

Figure 7

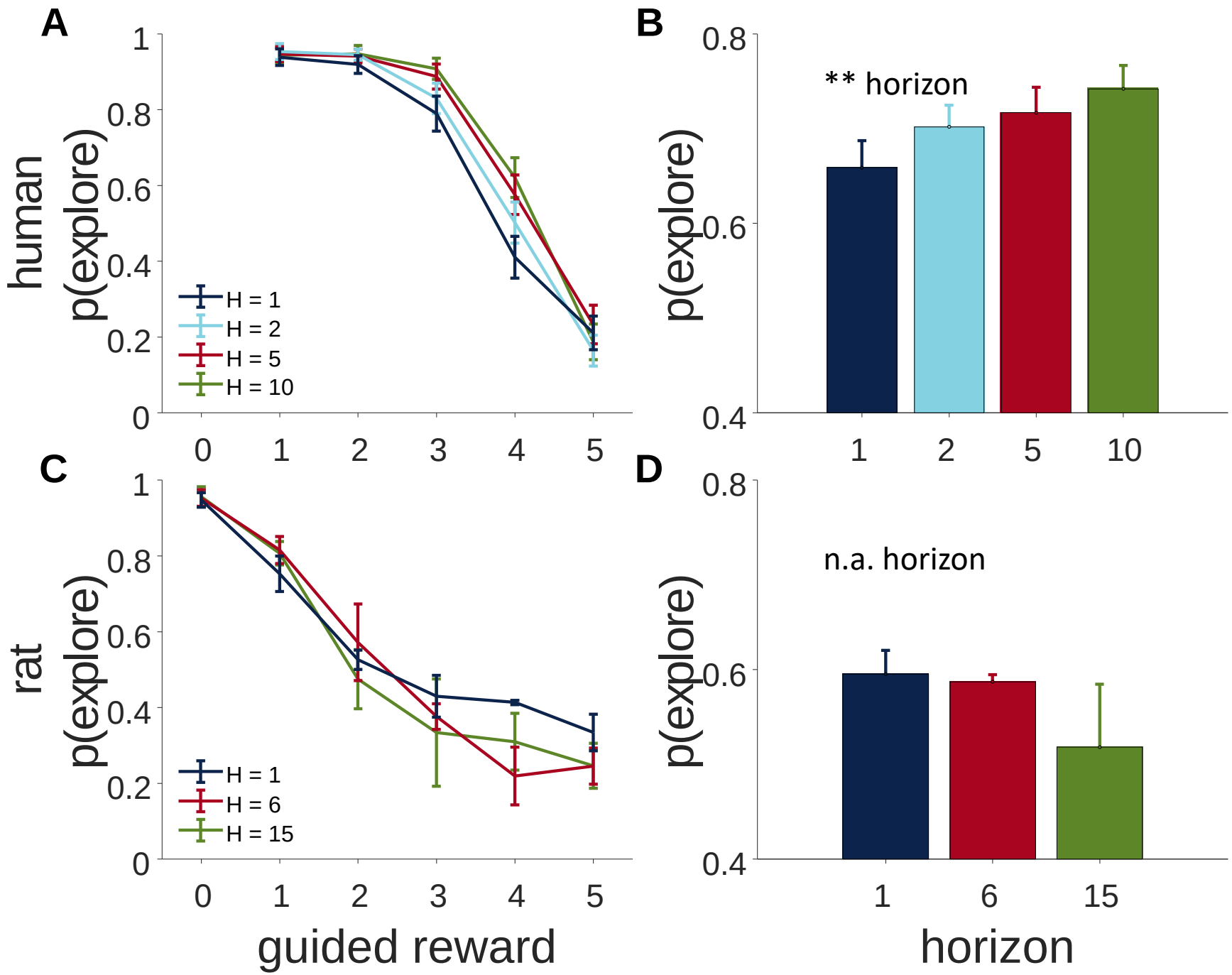

A
human p(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
H = 1
H = 2
H = 5
H = 10
B
p(explore)
0.8
0.6
0.4
** horizon
1 2 5 10
C
rat p(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
H = 1
H = 6
H = 15
guided reward
D
p(explore)
0.8
0.6
0.4
n.a. horizon
1 6 15
horizon

Figure 8

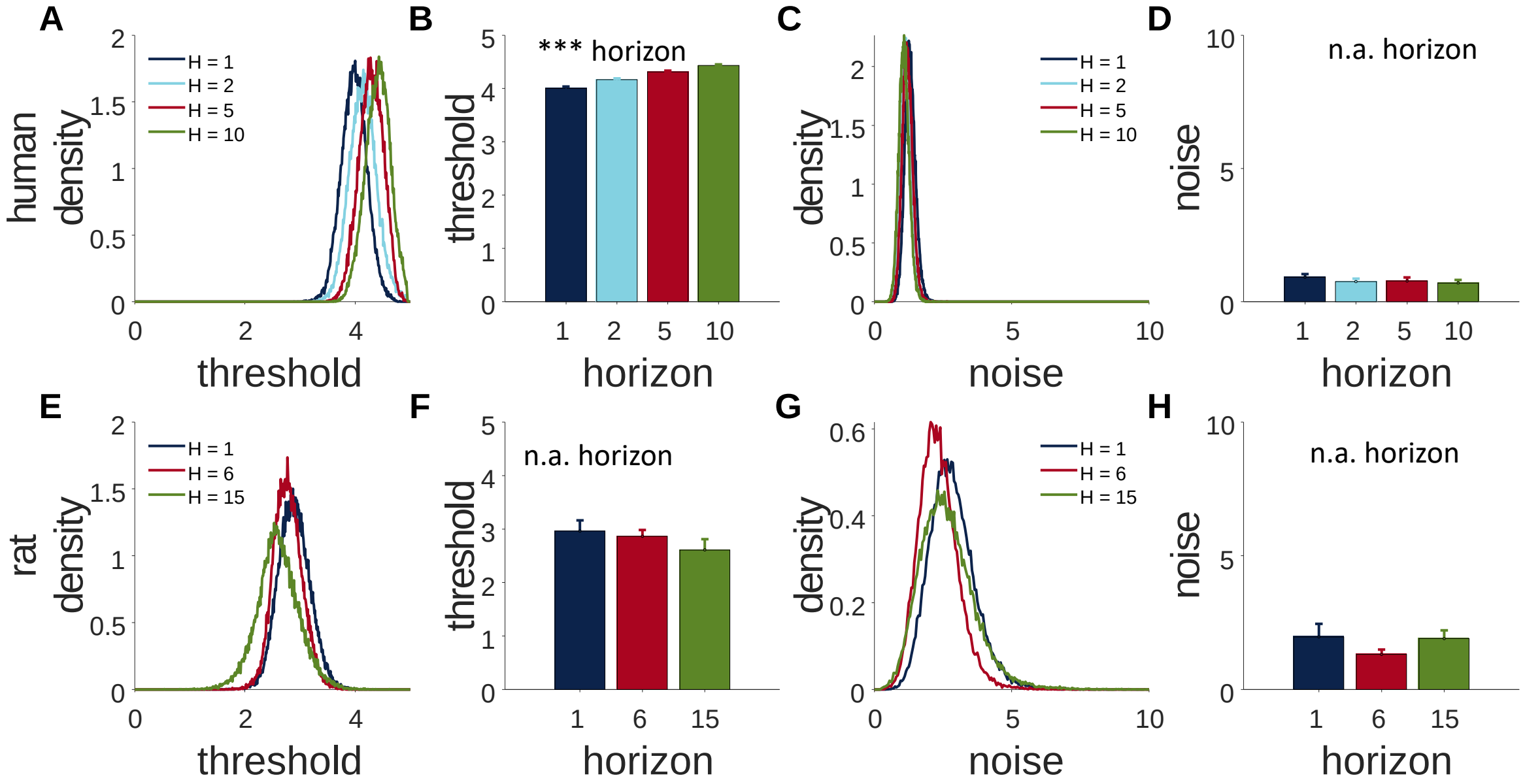

A
human density
H = 1
H = 2
H = 5
H = 10
threshold
B
*** horizon
threshold
horizon
C
density
noise
D
n.a. horizon
noise
horizon
E
rat density
H = 1
H = 6
H = 15
threshold
F
n.a. horizon
threshold
horizon
G
density
noise
H
n.a. horizon
noise
horizon

Figure 9

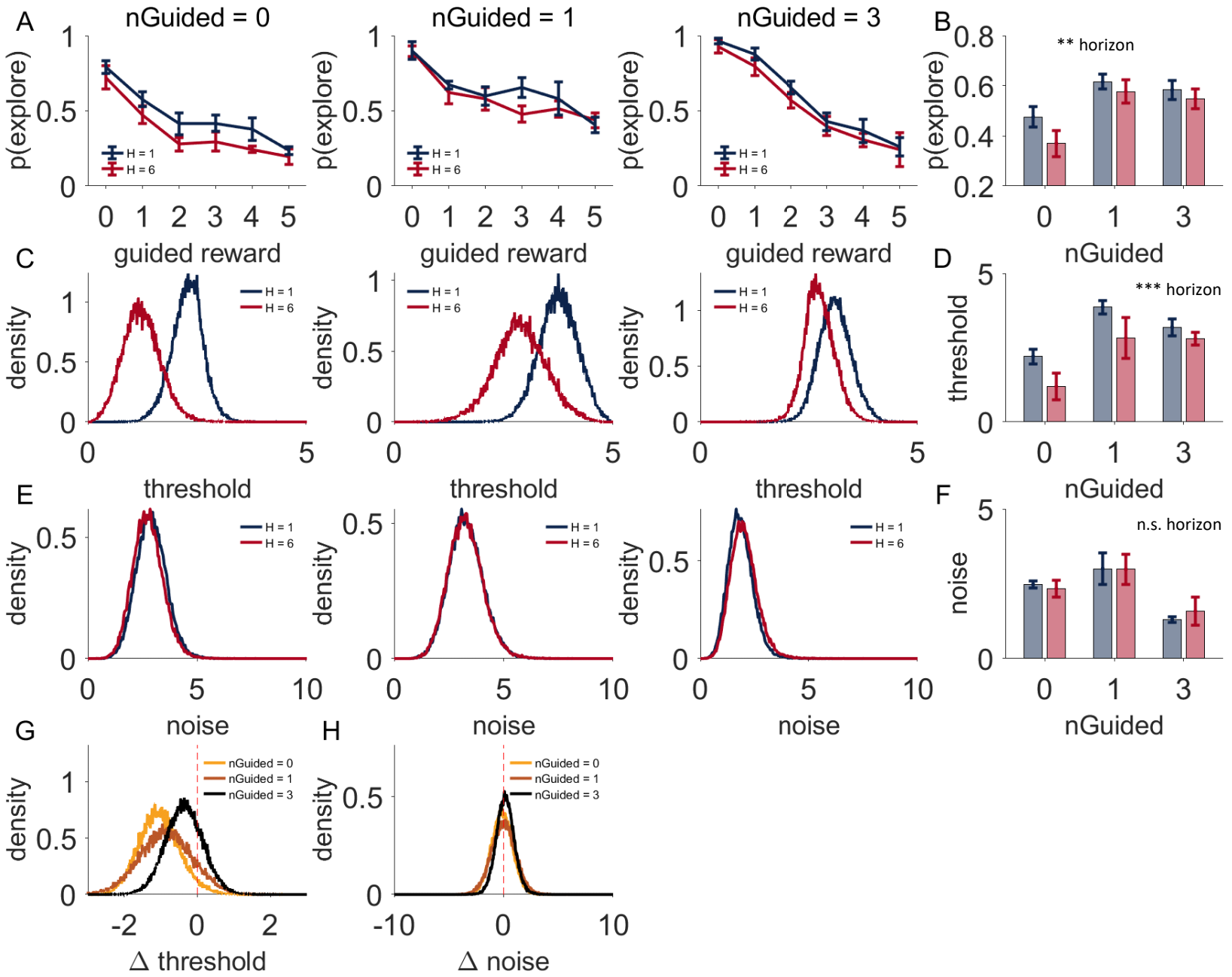

A
nGuided = 0
nGuided = 1
nGuided = 3
p(explore)
H = 1
H = 6
B
** horizon
C
guided reward
density
D
nGuided
*** horizon
threshold
E
threshold
F
nGuided
n.s. horizon
noise
G
noise
nGuided = 0
nGuided = 1
nGuided = 3
Δ threshold
H
Δ noise

Figure 10

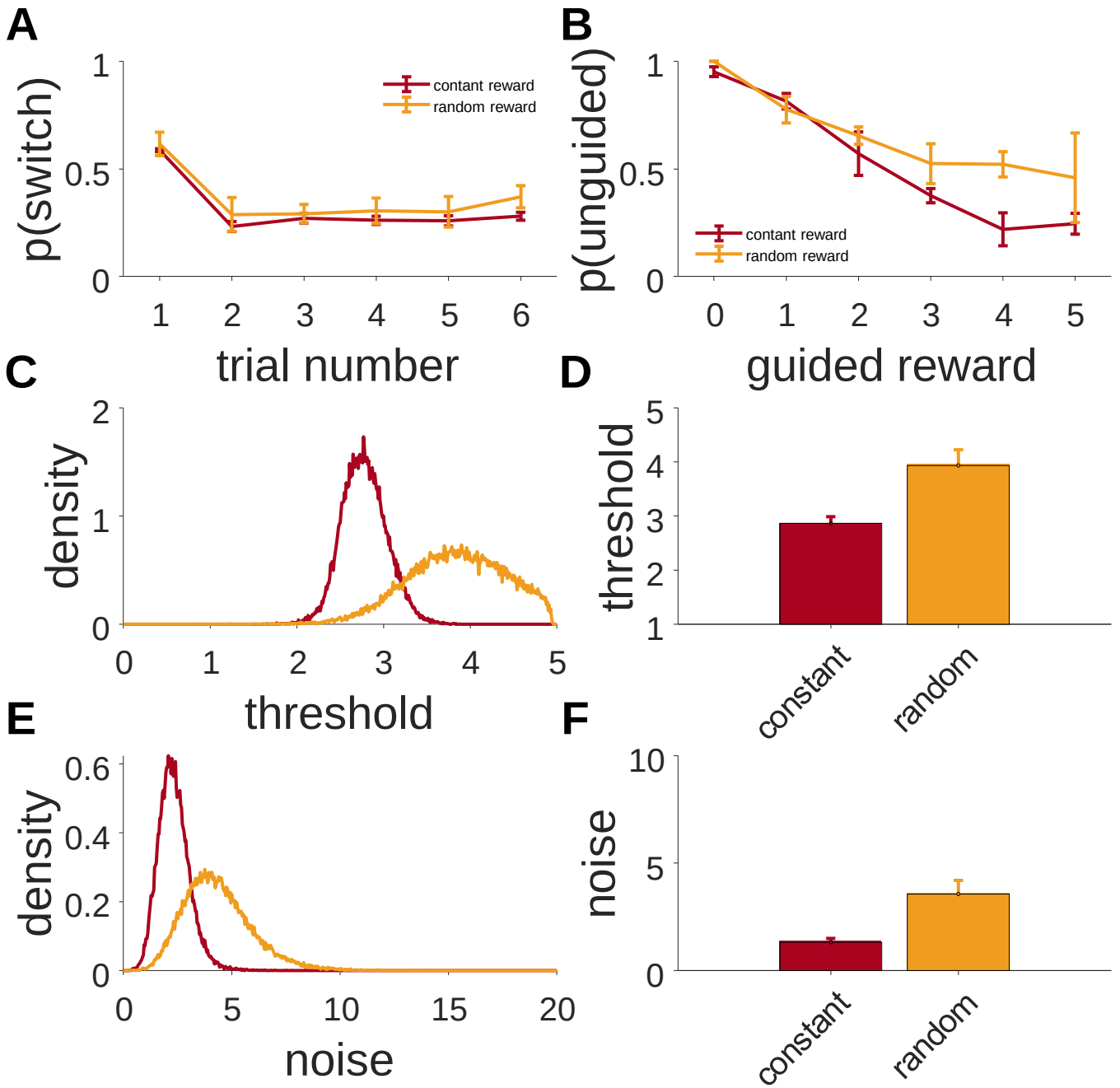

A
p(switch)
contant reward
random reward
trial number
B
p(unguided)
contant reward
random reward
guided reward
C
density
threshold
D
threshold
constant
random
E
density
noise
F
noise
constant
random

Figure 11

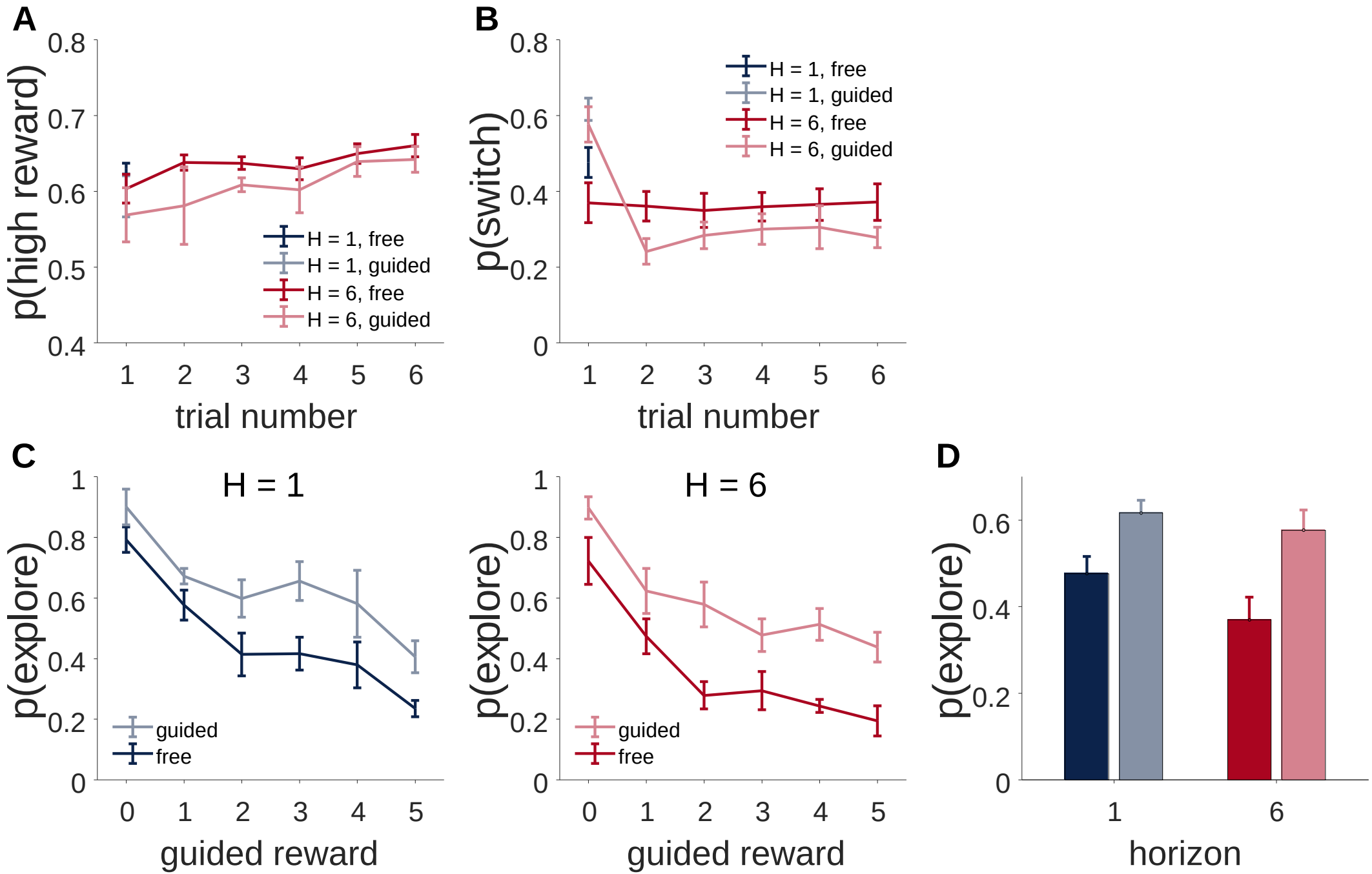
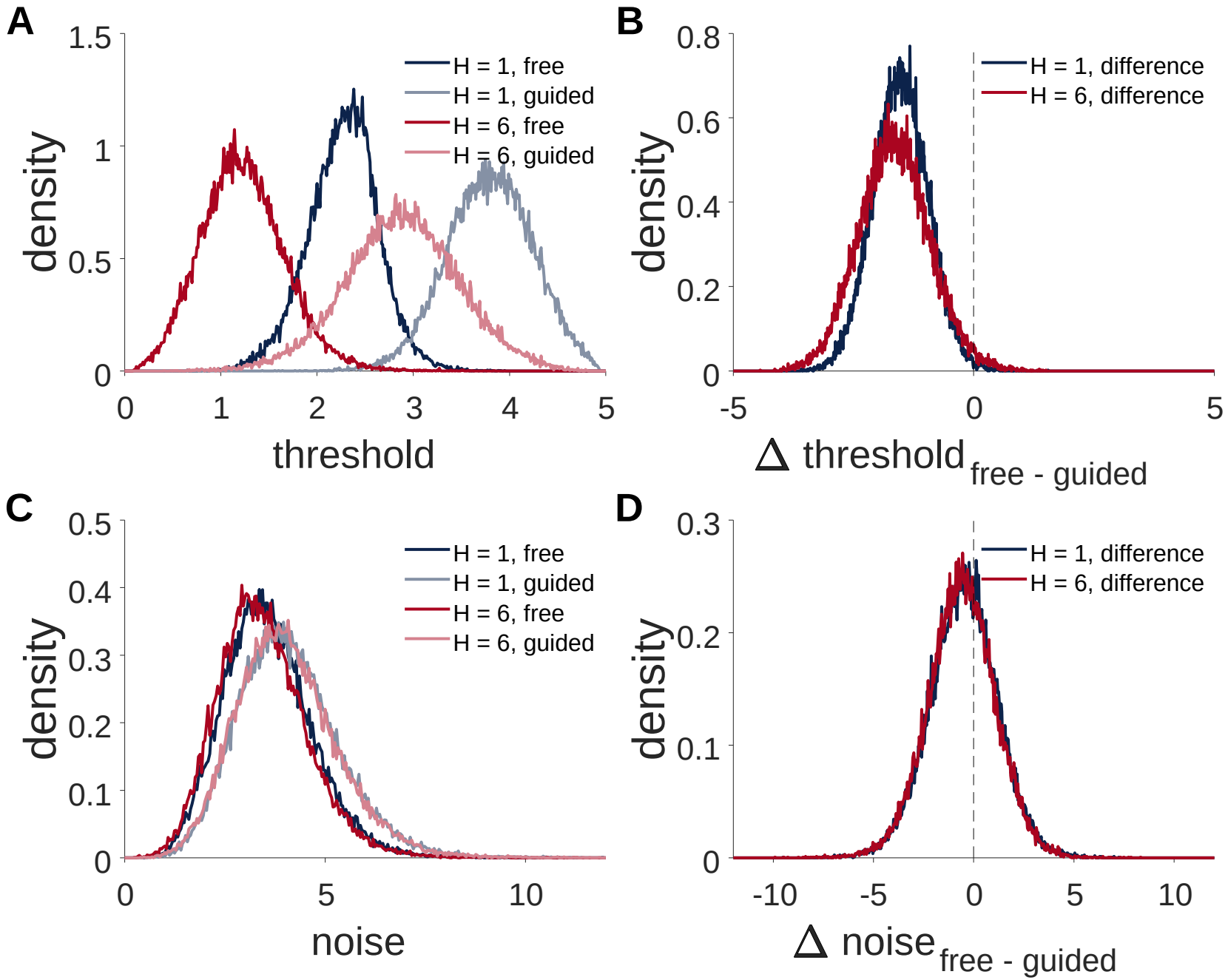

A
p(high reward)
0.8
0.7
0.6
0.5
0.4
1 2 3 4 5 6
trial number
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
B
p(switch)
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6
trial number
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
C
H = 1
p(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
guided reward
guided
free
H = 6
p(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
guided reward
guided
free
D
p(explore)
0.6
0.4
0.2
0
1 6
horizon

Figure 12


A
density
threshold
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
B
density
Δ threshold free - guided
H = 1, difference
H = 6, difference
C
density
noise
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
D
density
Δ noise free - guided
H = 1, difference
H = 6, difference

Supplementary figures

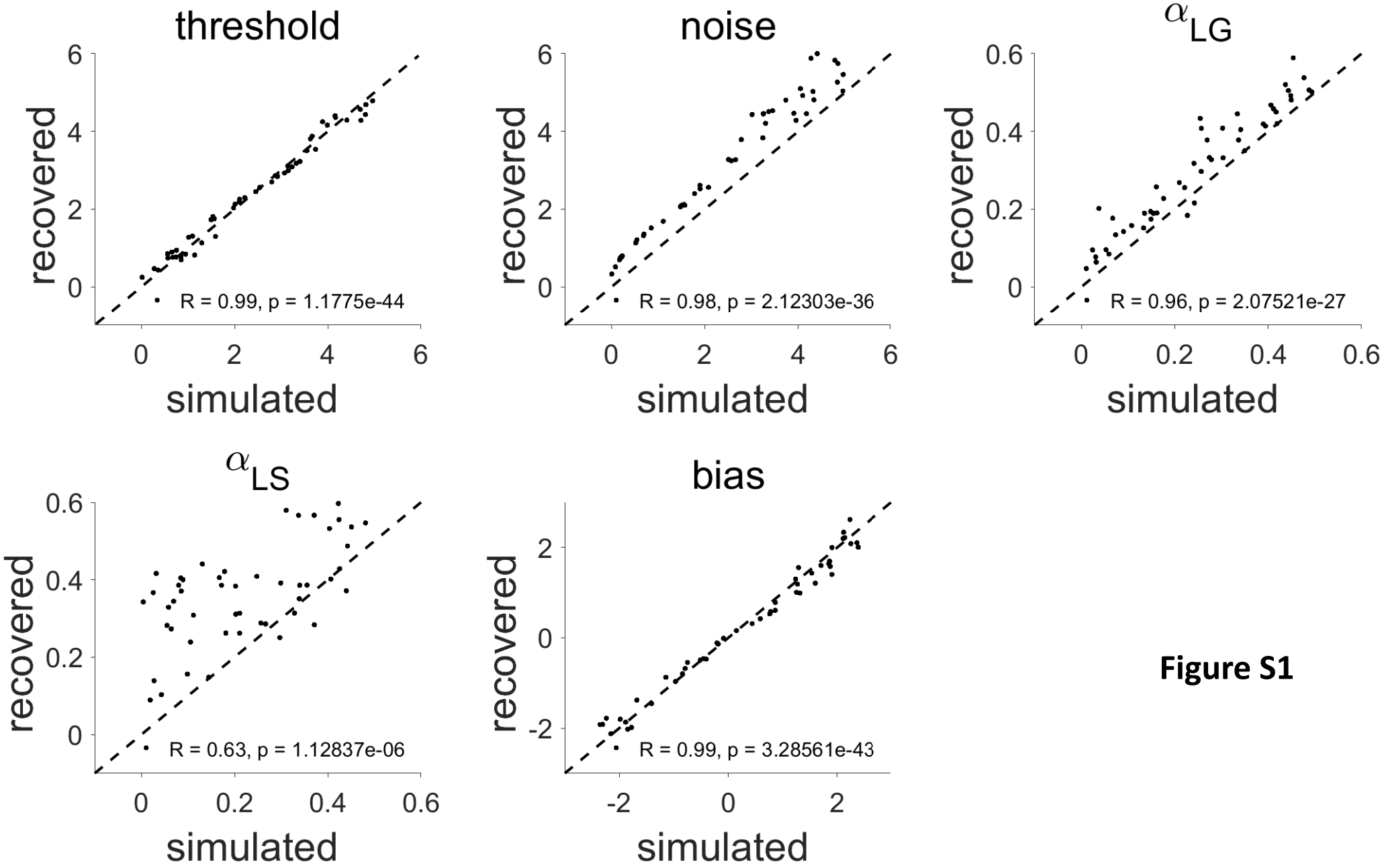

threshold
recovered
simulated
R = 0.99, p = 1.1775e-44
noise
recovered
simulated
R = 0.98, p = 2.12303e-36
α_{LG}
recovered
simulated
R = 0.96, p = 2.07521e-27
α_{LS}
recovered
simulated
R = 0.63, p = 1.12837e-06
bias
recovered
simulated
R = 0.99, p = 3.28561e-43

Figure S1

Figure S2

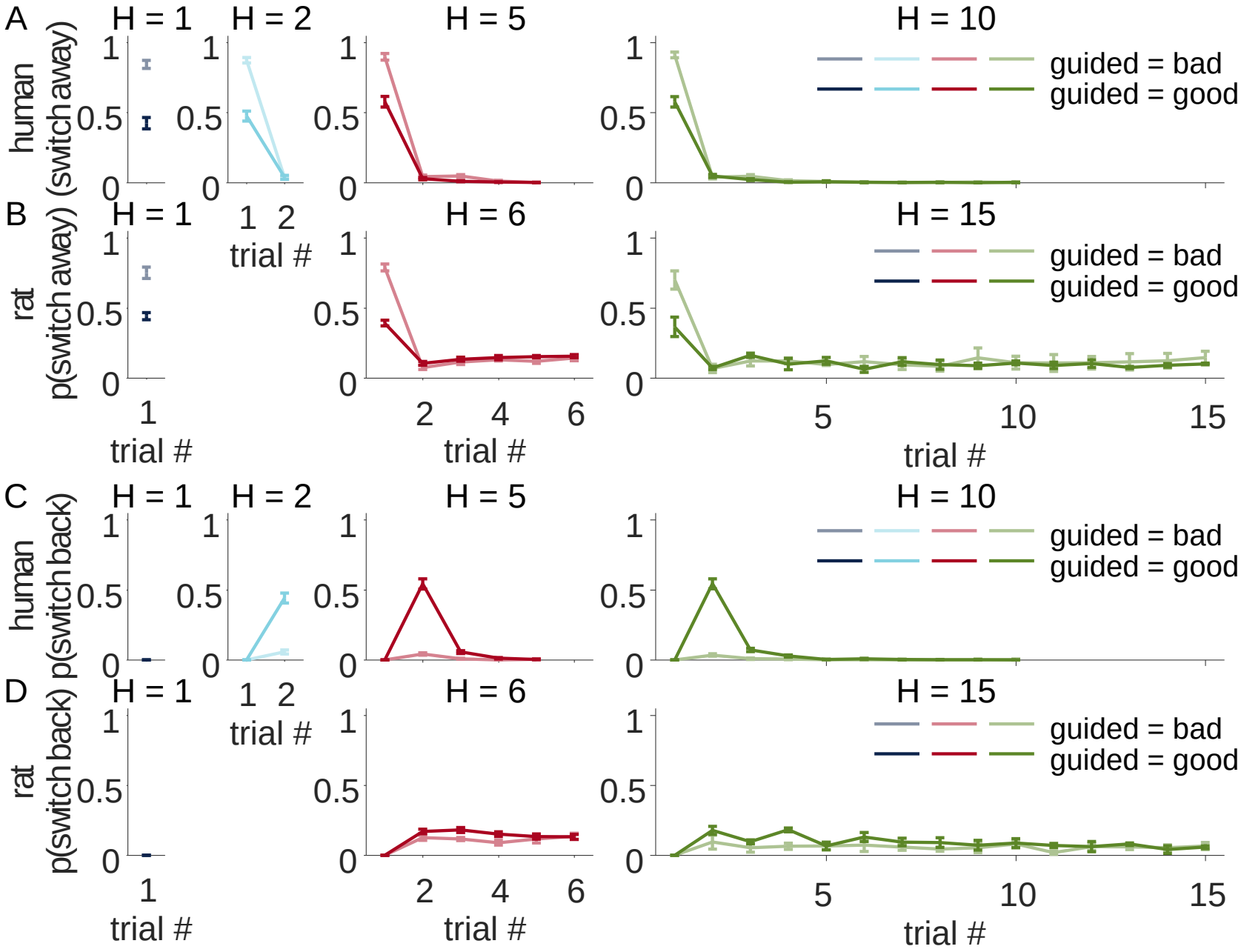

A
B
C
D
H = 1
H = 2
H = 5
H = 10
H = 6
H = 15
human p(switch away)
rat p(switch away)
human p(switch back)
rat p(switch back)
guided = bad
guided = good
trial #
0
0.5
1
2
4
6
5
10
15

Figure S3

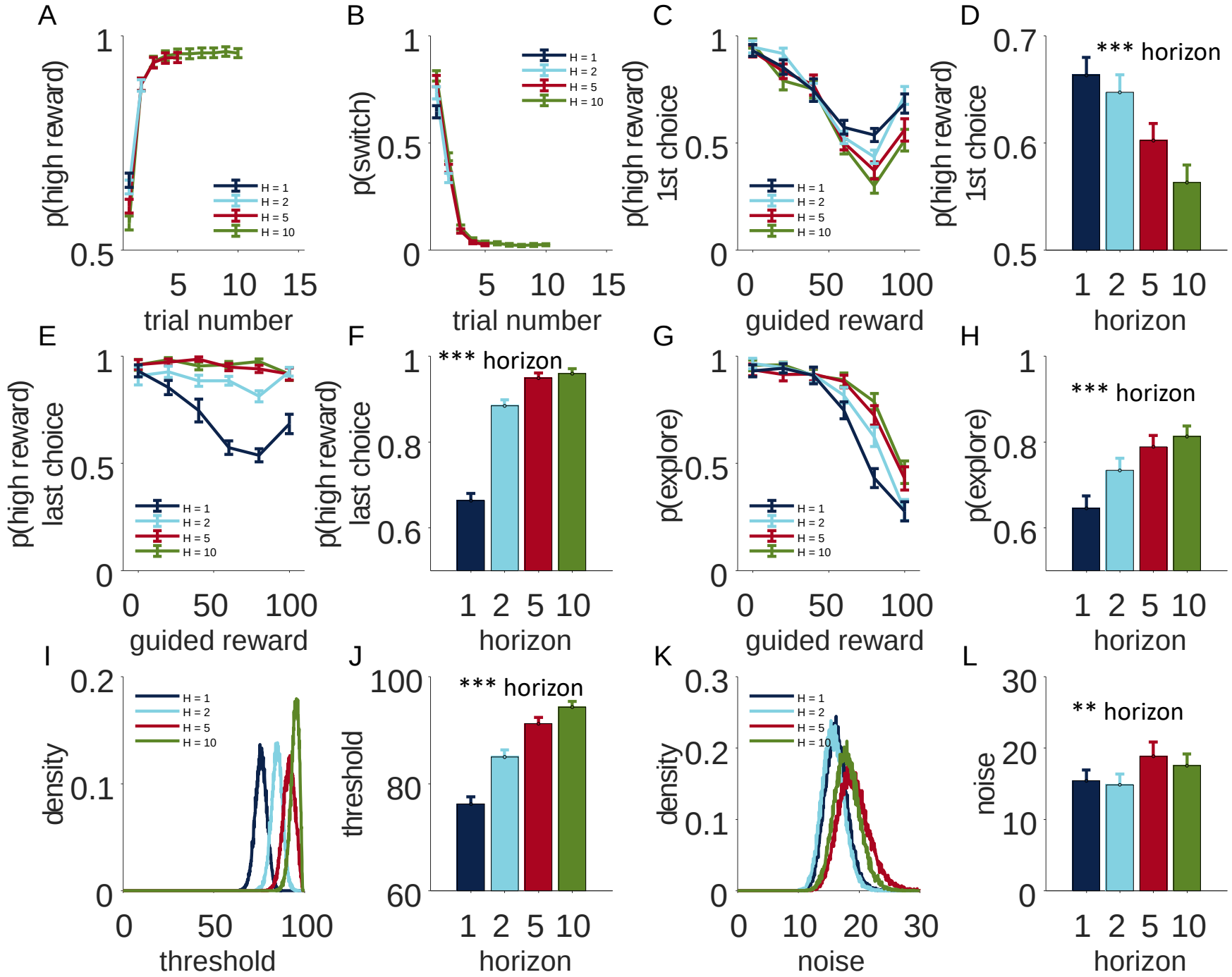

A
p(high reward)
trial number
H = 1
H = 2
H = 5
H = 10
B
p(switch)
trial number
C
p(high reward) 1st choice
guided reward
D
p(high reward) 1st choice
*** horizon
horizon
E
p(high reward) last choice
guided reward
F
p(high reward) last choice
*** horizon
horizon
G
p(explore)
guided reward
H
p(explore)
*** horizon
horizon
I
density
threshold
J
threshold
*** horizon
horizon
K
density
noise
L
noise
** horizon
horizon

Figure S4

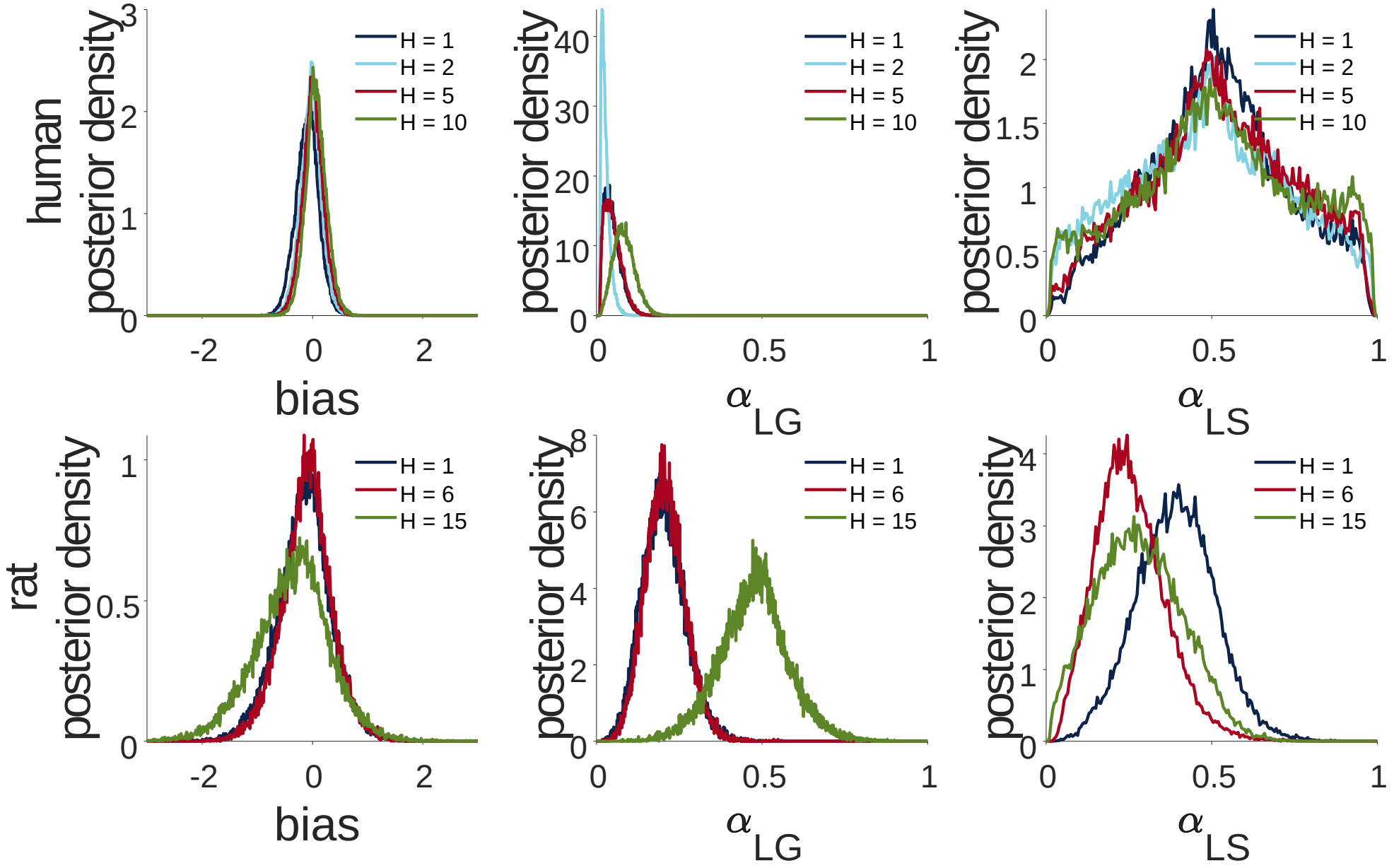

human
posterior density
bias
H = 1
H = 2
H = 5
H = 10
posterior density
α_{LG}
H = 1
H = 2
H = 5
H = 10
posterior density
α_{LS}
H = 1
H = 2
H = 5
H = 10
rat
posterior density
bias
H = 1
H = 6
H = 15
posterior density
α_{LG}
H = 1
H = 6
H = 15
posterior density
α_{LS}
H = 1
H = 6
H = 15

Figure S5

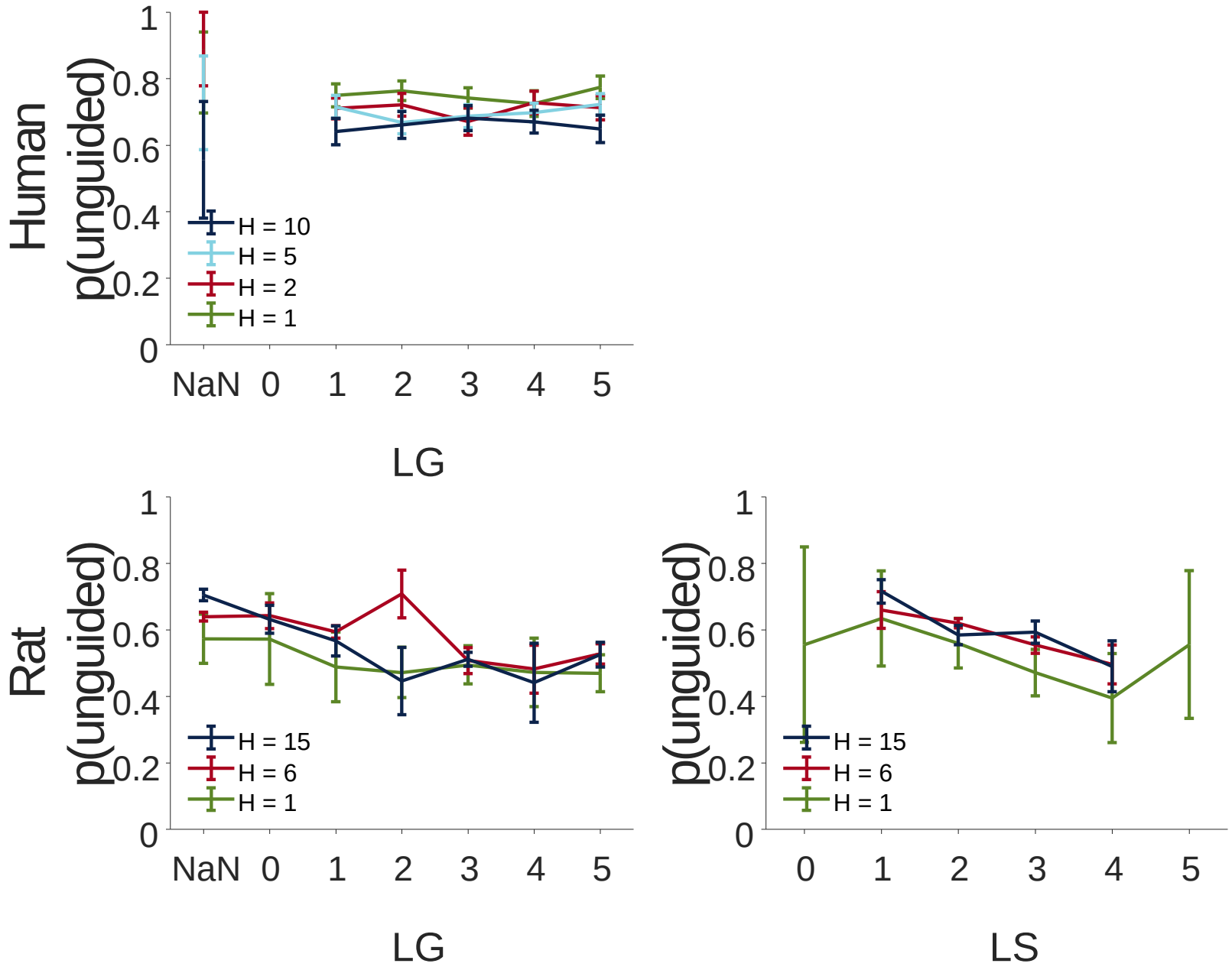

Human
p(unguided)
1
0.8
0.6
0.4
0.2
0
H = 10
H = 5
H = 2
H = 1
NaN 0 1 2 3 4 5
LG
Rat
p(unguided)
H = 15
H = 6
H = 1
NaN 0 1 2 3 4 5
LG
p(unguided)
H = 15
H = 6
H = 1
0 1 2 3 4 5
LS

Figure S6

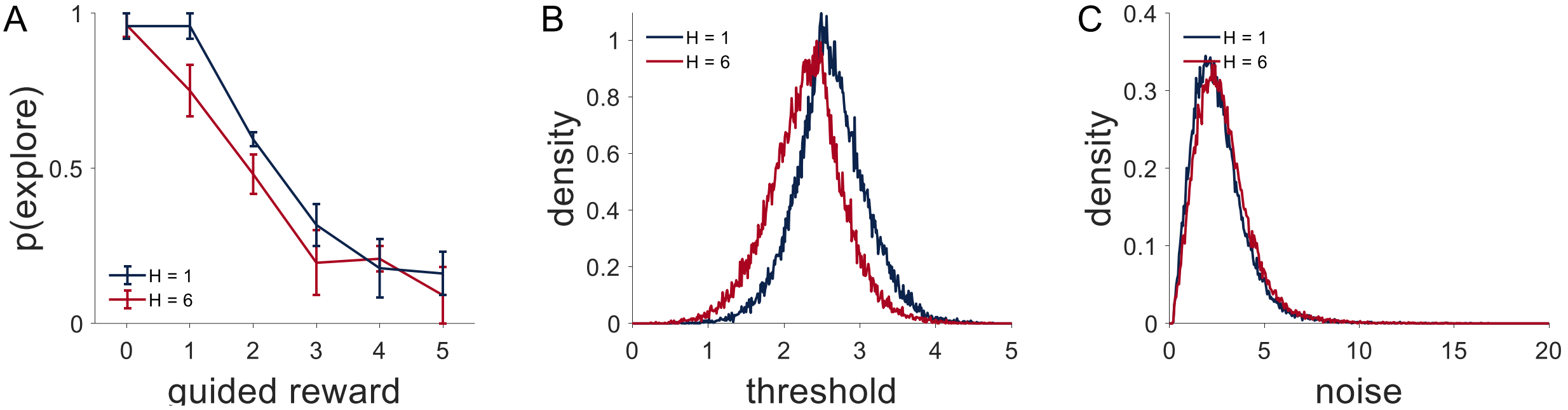
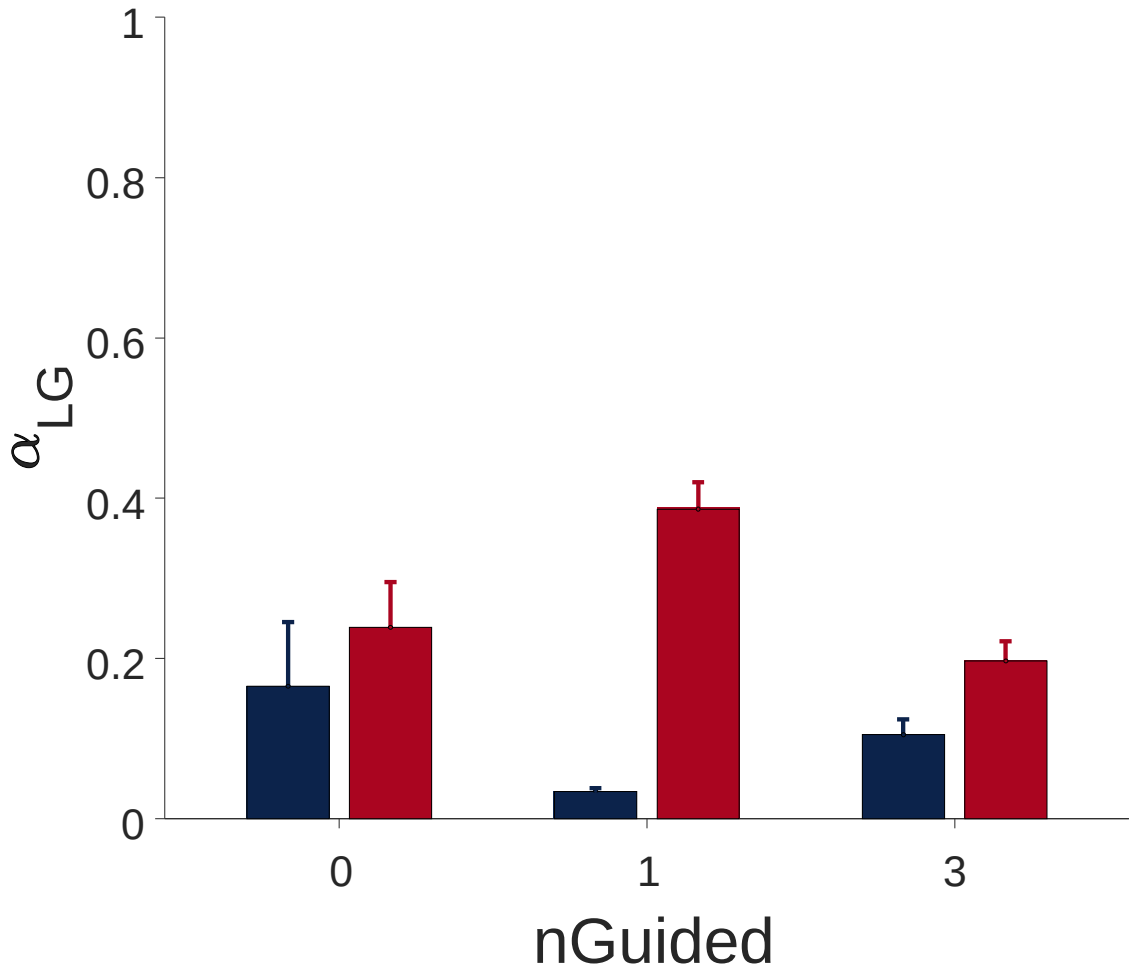

A
p(explore)
1
0.5
0
H = 1
H = 6
0
1
2
3
4
5
guided reward
B
density
1
0.8
0.6
0.4
0.2
0
H = 1
H = 6
0
1
2
3
4
5
threshold
C
density
0.4
0.3
0.2
0.1
0
H = 1
H = 6
0
5
10
15
20
noise

Figure S7


1
0.8
0.6
0.4
0.2
0
α_{LG}
0
1
3
nGuided

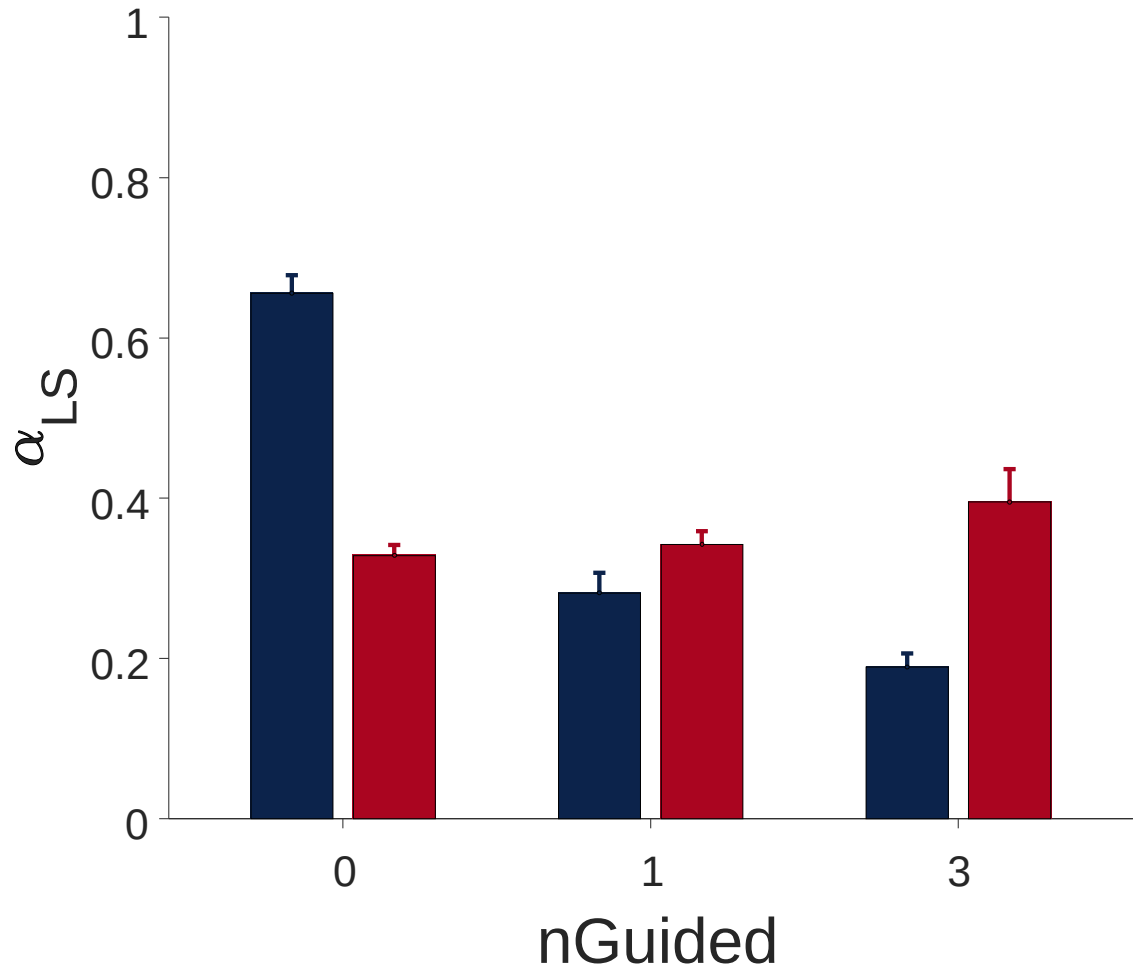
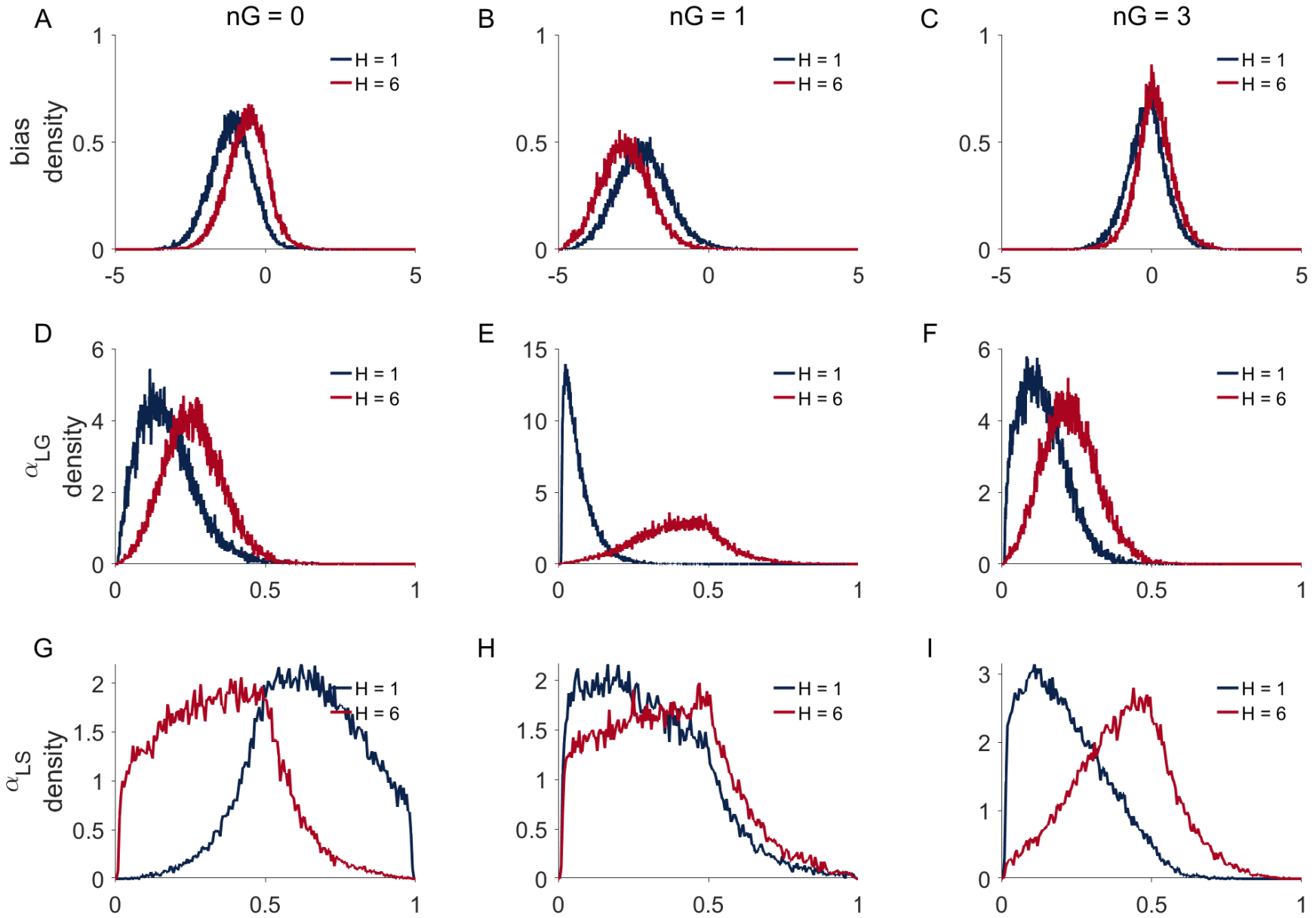

1
0.8
0.6
0.4
0.2
0
α_{LS}
0
1
3
nGuided

Figure S8


A
nG = 0
B
nG = 1
C
nG = 3
bias density
H = 1
H = 6
D
E
F
α_{LG} density
G
H
I
α_{LS} density