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: 228 words
Number of figures: 12 + 2 supplemental

Abstract

Humans and animals have to balance the need for exploring new options and 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 solve 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 addition, these paradigms do not allow for observing model-based exploration behaviors, such as utilizing prior information and adaptation to the volatility of the environment. 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 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. Given the similarities and important disparities observed between humans and rats, we revealed a more complex explore-exploit behavior than previously thought.

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

Introduction

Humans and animals constantly face the 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 like 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). Recent studies showed that humans were able to adapt the extent of their directed and random exploration with the horizon context, i.e. the number of future choices remaining (Wilson et al., 2014). Horizon adaptation is thought to be a hallmark of effective exploration, 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. 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. First, the decision to abandon a currently 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 particular, it is difficult to separate directed vs random exploration in reversal paradigms. Secondly, in exploration, both good and bad outcomes should occur. However, in reversal learning,

after the reversal point, “exploring” the previously suboptimal option will always lead to a better outcome compared to the currently bad option. 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. Fourth, there is 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. Finally, very little is known of the neural substrate of the explore-exploit decision circuits, and animal models with well-defined behavioral quantifications allowing investigations of the cellular and system mechanisms of these complex processes are sorely needed.

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 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. In addition, 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. 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

The rodent experiments were run in 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 (0.2g/ml) in the form of computer-controlled drops. A blinking LED was attached to each feeder and acted as a cue when desired. The experimental sessions were divided into 'games.' During each game, only 3 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 reward 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 reward 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 N 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 N+1st trial, they were cued to make free choices (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 and signaled the start of a new game (Game 2 start, Fig 1B). The new rewarded feeders are 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 complete. 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 figures S2.

In each session, rats were guided nG times before a free choice can 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 is actually guided (by the light cue). Four rats participated in this experiment and completed a total of 218 sessions and 5587 games (28436 trials).

Experiment 3: randomized reward

In this 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 slots machines (also referred to as bandits, Fig 3A) that gave out a fixed number of reward points uniformly drawn from 1 to 5. 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. 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 figures S1. 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(explore) is the percentage of choosing the unguided option on the first free choice, i.e. P(switch) on the first free choice. P(explore) is consistent with directed exploration, 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 explore 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

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

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

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.

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 3 free parameters, namely the exploration threshold θ, spatial bias b 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 $\theta_{hg} = \frac{a_{hg}}{a_{hg}+b_{hg}}$ and the group-level mean of decision noise $\sigma_{hg} = \frac{k_{hg}}{\lambda_{hg}}$. Posterior distributions over the exploration threshold θ_{hg} and the decision noise σ_{hg} are shown for each experiment (Fig 7, 8, 10, 11).

Table 1 Model parameters, priors and hyperpriors

Parameter	Prior	Hyperpriors
Exploration threshold θ_{hgs} $\theta_{hgs} = \theta'_{hgs} R_{max}$	$\theta'_{hgs} \sim \text{Beta}(a_{hg}, b_{hg})$	$a_{hg} \sim U(0.1, 10)$ $b_{hg} \sim U(0.1, 10)$
Decision noise σ_{hgs}	$\sigma_{hgs} \sim Gamma(k_{hg}, \lambda_{hg})$	$k_{hg} \sim Exponential(1)$ $\lambda_{hg} \sim Exponential(10)$
Spatial bias b_{hgs}	$b_{hgs} \sim Gaussian(\mu_{hg}, \epsilon_{hg})$	$\mu_{hg} \sim Gaussian(0, 0.0001)$ $\epsilon_{hg} \sim Gamma(1, 0.001)$

* R_{max} is the maximal reward in the experiment. $R_{max} = 5$ for all experiments except for human experiment 5, in which $R_{max} = 100$.
** 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.

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 higher during the last trial of longer horizons compared to shorter horizons (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% (Figure 4A) whereas at the last trial of the longest horizon condition (H = 15), rats could achieve an average accuracy of 81% (Fig 4C).

Rats switched from the last chosen option at a significantly higher rate at trial 1 (59.8% for H = 6and 60.7% for H = 15, Fig 4D) and then adopted a more constant and lower rate of switching for later trials (averaged 26.6% for H = 6 and 21.0% for H = 15, Fig 4D), whereas humans switched more at trial 1 (70.2%, 71.7% and 74.1% 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), possibly due to boredom or motor error. 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 and switched on the 1st free choice to find out that the alternative was worse, they immediately switched back on the 2nd choice (Figure 5C). It took longer for rats to switch back. The percentage of switching remained higher when guided to a good choice than to a bad choice until the 4th trial (Figure 5D). Interestingly, when guided to a good option at first, both rats and humans 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, 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). When they were guided to 0 or 5 drops (or 1 and 5 points for humans), the accuracy was the highest whereas the accuracy was lowest when they were guided to more ambiguous reward amounts such as 2 or 3 drops (or 3 or 4 drops for humans). With prior information alone, it is theoretically difficult 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 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(explore), 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). Like humans (Fig 7A), we found that rats were very likely to explore if they obtained a low reward during the guided trials (e.g. 0 drops, mean = 94.5% Fig 7C), and were very unlikely to explore if they obtained a large reward (e.g. 5 drops, mean = 29.1%, Fig 7C). Overall, when guided to the option with an objectively lower reward, rats explored the unguided feeder at around 70% on their first free choices, whereas when guided to the option with an objectively higher reward, rats only explored the unguided feeder at around 40% on their first free choices (Fig 5D, trial 1). Humans explored more than 80% on an objectively lower guided reward, and around 50% 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" 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 vs 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 (see Fig 9A, 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(explore) is akin to the p(high info) measure in previous human research and is a model-free way of measuring directed exploration (Wilson et al., 2014). In line with previous research, humans explored significantly more in long horizons than in shorter ones (Fig 7B, Experiment 4, Fig S1G, H, Experiment 5), however for rats, the long

horizon condition seemed to yield slightly lower probability of exploring than the short horizon condition (Fig 7D, h=15 vs h=1, 6, Experiment 1).

To better quantify directed vs 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. For humans, we observed an increase of threshold as horizon increases (Fig 8A, B), 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. Interestingly, in rats we observe the opposite (model-based differences of H = 1 and H = 6 will be more obvious in Experiment 2, see Fig 9). Rats decrease their thresholds as horizon increases, thus, they use less directed exploration in their first free choice when the horizon is long (Fig 8E, F).

While exploration threshold θ is theoretically tied to directed exploration, decision noise σ is tied to random exploration. It turns out that this task may not be best suited to 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 not 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 higher than decision noise (H = 1, 2) in shorter horizons (Fig S1K, L), which is in line with the horizon adaptive random exploration reported in human studies (Wilson et al., 2014). 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 (Fig 8C, D) or rats (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 in chunks of consecutive days. The amount of training a particular rat was exposed to before each condition influenced how they explore across different horizon conditions. Furthermore, the within-session version may make the difference between horizon conditions more salient to the rats. 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 is always associated with short-horizon games (H = 1) and the other home base is always associated with long-horizon games (H = 6). In this alternate design, there is therefore no confound of learning/training effect.

In Experiment 2, we showed that regardless of the number of guided trials, P(explore) was lower for Horizon 6 compared to Horizon 1 (Fig 9A, B). Through a two-way ANOVA analysis of horizon and the number of guided choices, we found a significant main effect of horizon on P(explore) with $p = 0.007$. Using the model, we confirmed that regardless of the number of guided trials, exploration threshold θ for H = 6 was lower than H = 1 (Fig 9C, D). There was a significant main effect of horizon on θ, $p = 0.003$. 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 $\Delta\theta < 0$ is 97.9%, 73.5% and 75.2% 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 $\Delta\theta > 0$ is 50.2%, 54.0% and 56.8% 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 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. 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 S2). 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.

This opposite dependence of directed exploration on horizon in rats vs humans can arise from several factors. First, the utility of 1 to 5 drops is different for humans and rats. Humans get points, whereas rats get real sugar water proportional to the number of drops. As a result, P(explore) in human subjects were at ceiling for 1 and 2 points suggesting that both reward sizes were equally salient (Fig 7A), whereas P(explore) for 4 and 5 drops were similar in rats likely due to perceptual limitations (Fig 7C). Second, the efforts humans spent in making the decision was small. As a result, they over-explored to find out the best possible action, whereas rats had to physically travel the maze to get sugar water. Rats therefore likely under-explored to secure a satisfiable amount of return for each visit. In our data, rats had lower exploration thresholds compared to humans (Fig 8A, E).

Rats explore more in more volatile environments.

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 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, 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 reward 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 be H = 6. In this version, since there was no information that could be learned and the rewards were completely random, the rat's accuracy was at chance at 54.4% (Fig 10). Possibly due to overtraining in Experiment 1 and 2, after the guided choices, rats still explored more on the first free choices than on subsequent ones, 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 exploration of the unguided feeder was higher compared to P(explore) in the constant reward scenario in Experiment 1, especially when the guided reward size was high (Fig 10B). The difference was largest when rats were guided to 4 and 5 drops. The rat still switched at over 45%, even when the average guided trial experience included the best reward condition (i.e. 4, 5 drops), significantly higher than the constant reward condition. This could account for the horizon difference in Figure 7C showing that when the guided rewards were 4 and 5 drops, P(explore) was lower in H = 6 and H = 15 compared to H = 1. 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 can potentially account for the horizon difference in P(switch) in Figure 4D where there was a lower rate of switching in H = 15 compared to H = 6, 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. It has been proposed that relative uncertainty correlates with directed exploration whereas total uncertainty correlates with random exploration (Gershman, 2019). In a completely random environment, the uncertainty of both the guided and unguided feeder increased, so there was an increase in both relative uncertainty and total uncertainty. In line with this theory, through model fitting, we observed an increase in both the exploration threshold (Fig 10C, D, $p < 0.001$) and decision noise (Fig 10E, F, $p = 0.026$) in the random reward condition compared to the constant reward condition.

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

Finally, 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 better 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 explored more on the first free choice than in subsequent choices as in other variants of the task. However, when they chose freely, the 2nd choice did not differ from subsequent choices anymore, and rats seemed to have kept a steady rate of switching throughout the game, at a rate higher than the Guided condition (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 explore 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 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 99.2%, and 99.7% for H = 1 and H = 6 respectively (Fig 12B). In other words, when rats were 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 59.7%, and 63.2% for H = 1 and H = 6 respectively (Fig 12D). When self-guided in the Free choice condition, rats behave slightly more predictably in the first choice (2nd trial in Free choice condition, 1st free choice in Guided condition) by having a lower decision noise term compare to when they were guided.

Discussion

In this study, we investigated the exploratory behaviors in 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 the reward conditions at the same set of locations/feeders as in traditional reversal learning paradigms, using an open field task, 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, the rats were guided to one of two feeder locations first, and the extent to which they explored the other unvisited feeder location in their free choices was used as a measure of exploration. This measure is an equivalent of the model-free measure of directed exploration in previous human studies (Wilson et al., 2014). In addition, rats performed the task in both a short and a long horizon conditions 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 like 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 in reversal learning that animals choose to explore more when the exploit value was low and explored less when the exploit value was high. However, unlike in reversal learning where a “win” or a “loss” is computed by comparing the current reward with the estimated 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 exploration.

In this study, we were able to separate directed exploration from random exploration. The percentage of choice of the unguided option served as a model-free measure of directed 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 directed exploration to the time horizon. For random exploration, with small range of reward size (0 – 5), we did not observe adaptations of random exploration in either humans or rats in this task. However, with a larger reward range (1

– 100), in human Experiment 5, we did observe a higher level of random exploration (Fig S1K, L) in longer horizons (H = 5 and 10) compared to shorter ones (H = 1 and 2). This can be considered a limitation of the current design. Although we are able to separate directed from random exploration, with the deterministic rewards and small reward changes, it was difficult to observe random exploration adaptation with horizon.

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. Our results do not fully explain this phenomenon, but we speculate that there may be an 'optimize vs satisfice' discrepancy in humans vs rats due to the nature of the rewards received. Humans receive hypothetical points with relatively effortless keypresses on a computer keyboard, whereas rats earned their rewards by running back and forth on a meter-long table. 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. The exploration threshold in our data is overall higher in humans compared to rats (Fig 7A, C). The drive to explore is not to optimize for rats, but to satisfice. Exploring more in longer horizon may be an optimal way of exploration, but may not be an economic one. Rats may be less motivated in short horizon conditions because they gain overall less rewards (at seen in Experiment 2 when both horizons are alternated or interleaved within the same session). This would result in an increase novelty seeking and randomness in rat's behavior which would result in exploration. Also, in short horizon, without fully understanding the structure of the task, the 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 5 supported this view, in that, by having random rewards, rats still used directed exploration (only significantly higher on the first free choice compared to subsequent choices, Fig 9A) and explored more compared to the deterministic reward case in Experiment 1 (Fig 9B). Lastly, a longer horizon means that there were many opportunities to explore the unguided option later on, making it less urgent to explore on the first trial compared to a shorter horizon.

Nevertheless, we note the significance of the fact 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. 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 how other species can adapt exploration to time horizons.

In addition, we think 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 was 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 alternative 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).This rat model has 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 are 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 indicates 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(explore) 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 LED-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. 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(explore) as a function of guided reward size by horizon. H: average p(explore) 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 S2. 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(explore) as a function of guided reward size. B. Model estimates of exploration threshold. C. Model estimates of decision noise.

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

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*.

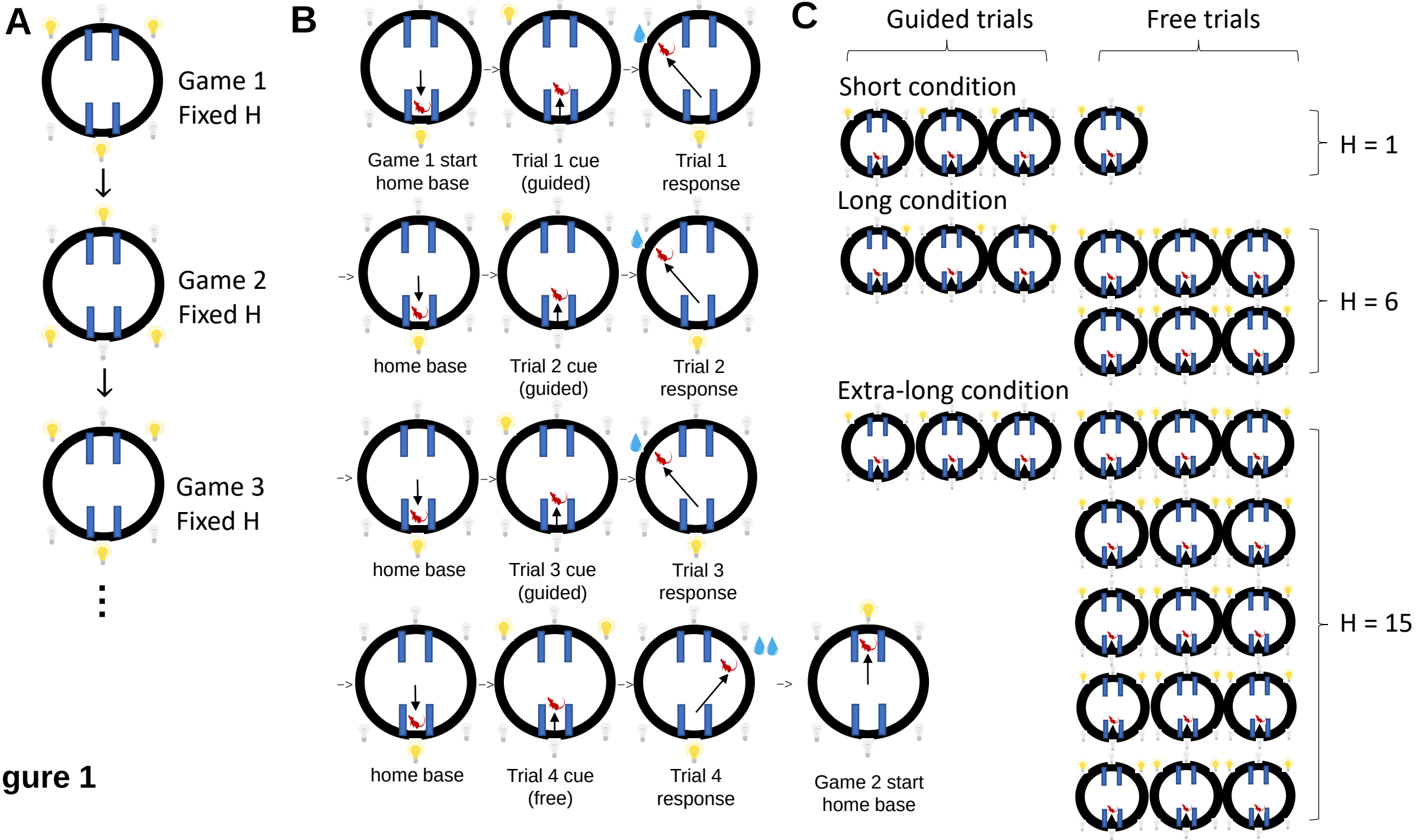
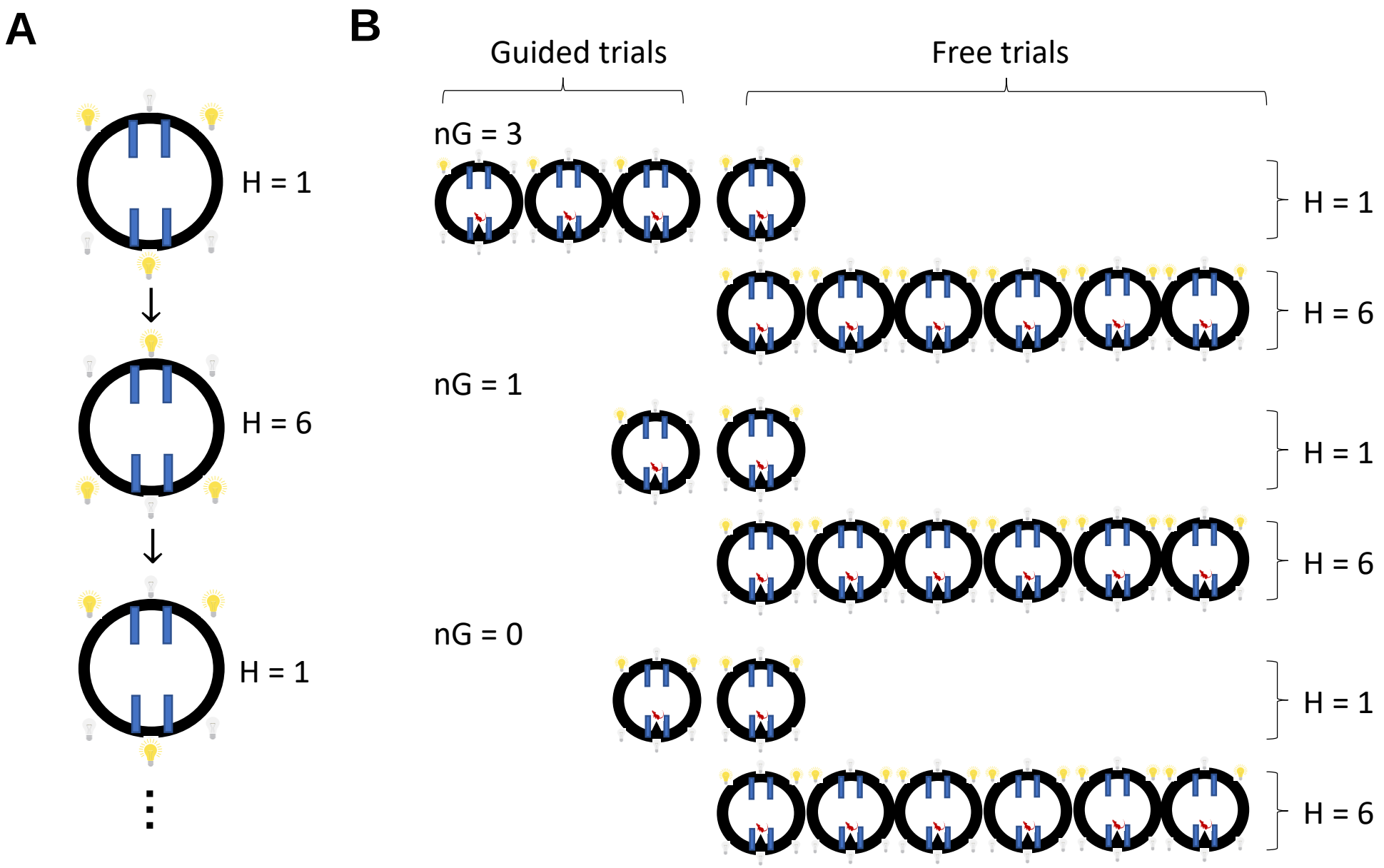

A
Game 1
Fixed H
Game 2
Fixed H
Game 3
Fixed H
B
Game 1 start
home base
Trial 1 cue
(guided)
Trial 1
response
home base
Trial 2 cue
(guided)
Trial 2
response
home base
Trial 3 cue
(guided)
Trial 3
response
home base
Trial 4 cue
(free)
Trial 4
response
Game 2 start
home base
C
Guided trials
Free trials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

Figure 1

Figure 2


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

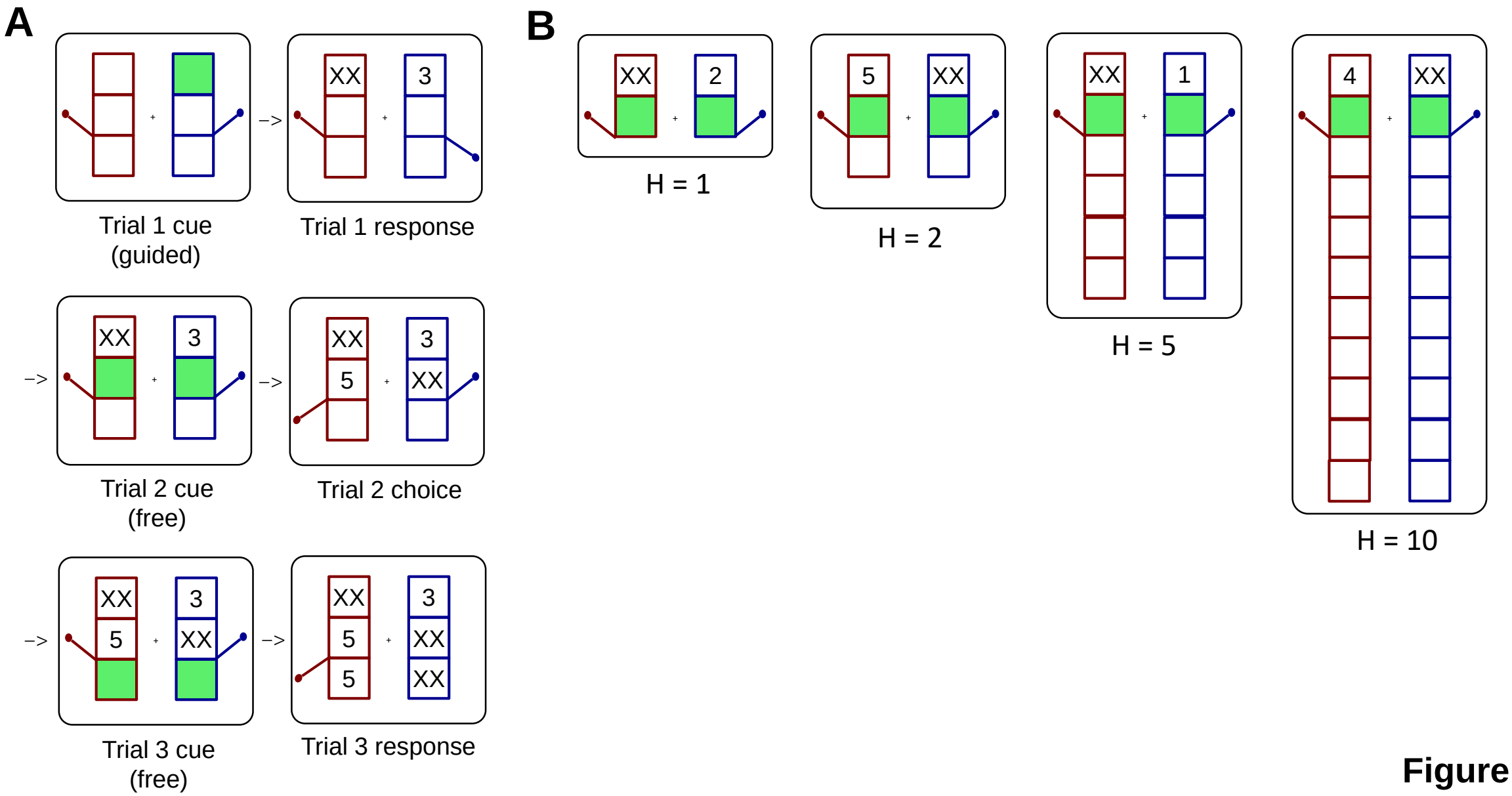

A
XX
3
Trial 1 cue
(guided)
Trial 1 response
XX
3
5
Trial 2 cue
(free)
Trial 2 choice
XX
3
5
5
Trial 3 cue
(free)
Trial 3 response
B
XX
2
H = 1
5
XX
H = 2
XX
1
H = 5
4
XX
H = 10

Figure 3

Figure 4

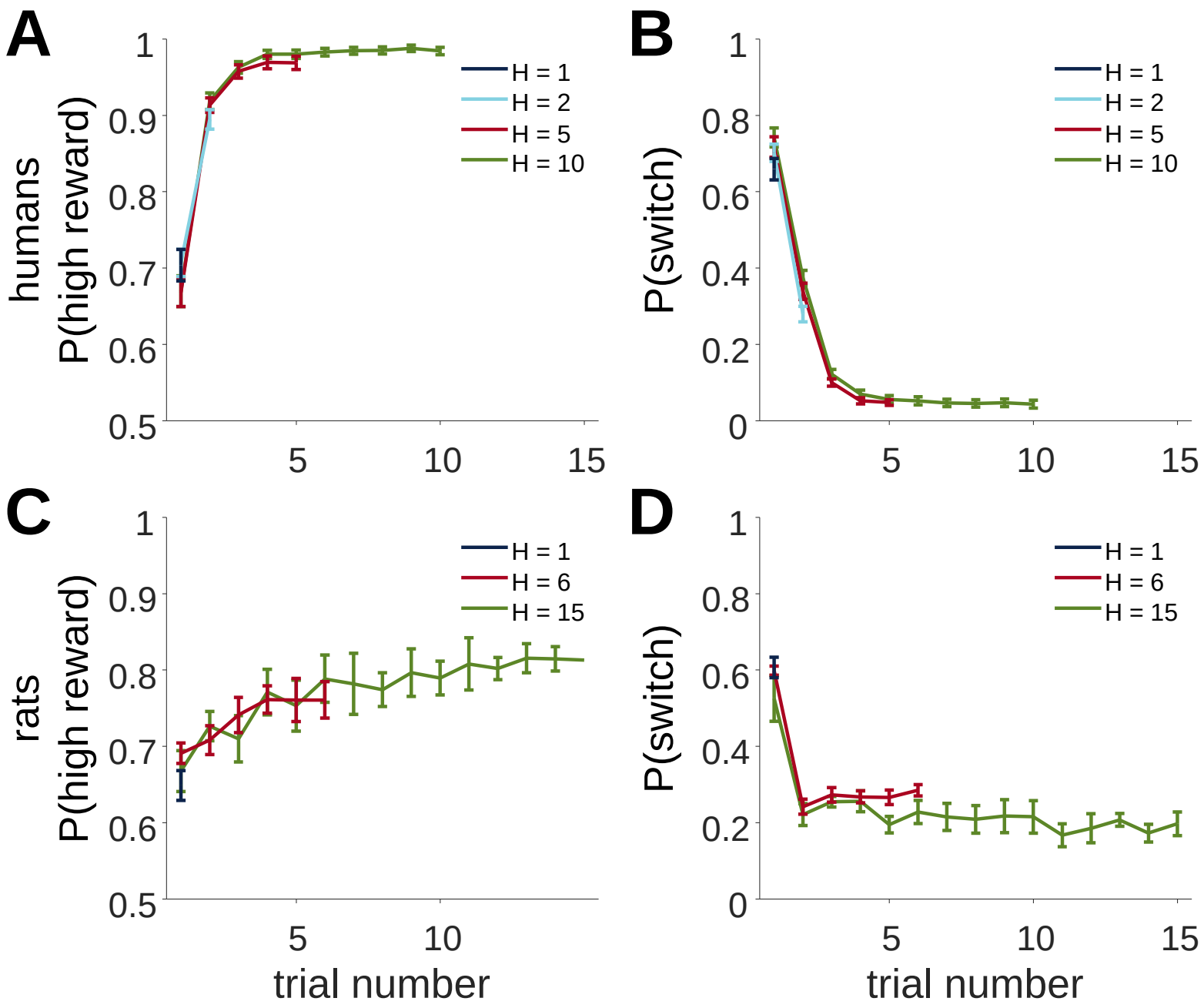

A
B
C
D
humans
P(high reward)
rats
P(switch)
trial number
H = 1
H = 2
H = 5
H = 10
H = 6
H = 15
1
0.9
0.8
0.7
0.6
0.5
0.4
0.2
0
5
10
15

Figure 5

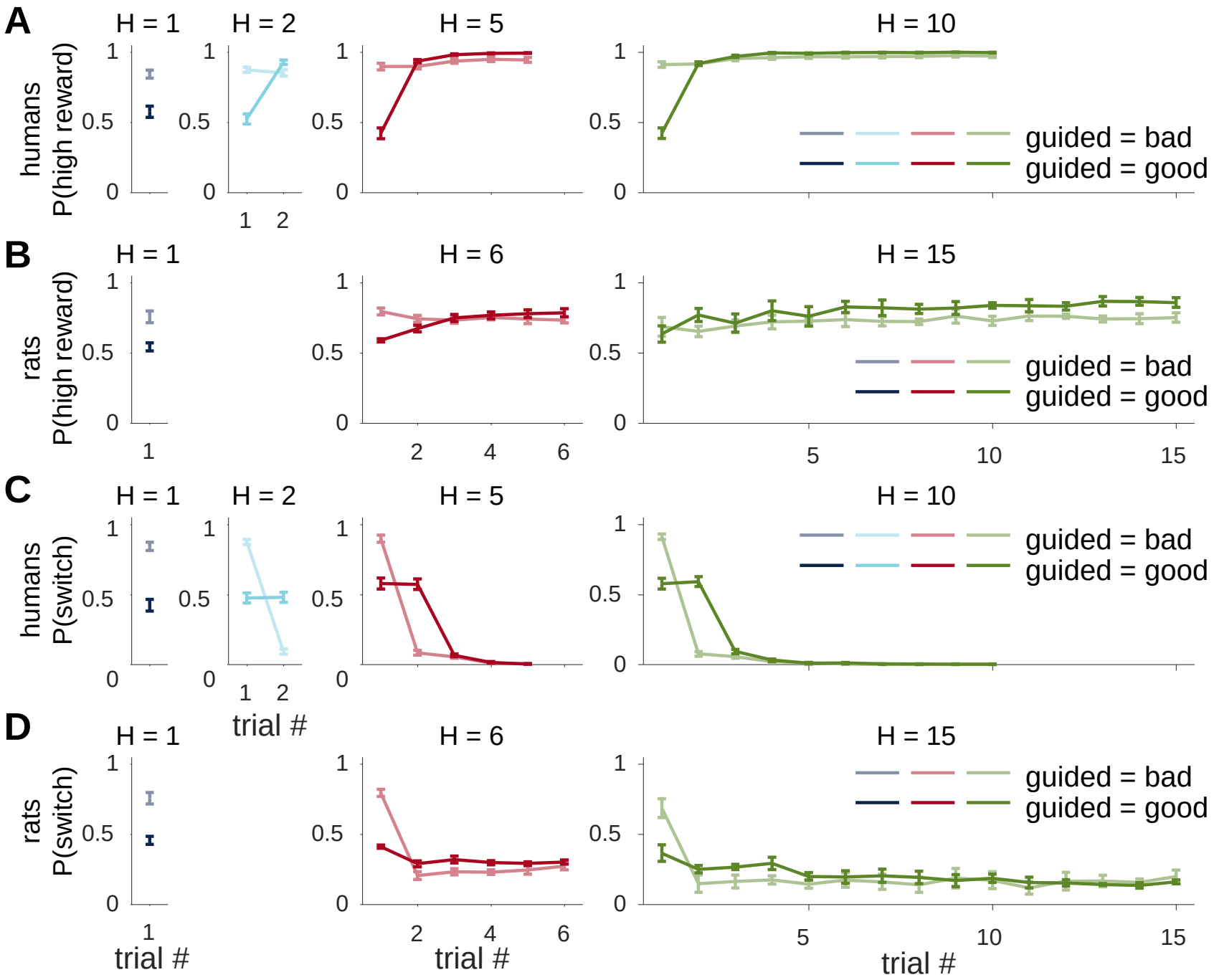

A
H = 1
H = 2
H = 5
H = 10
humans
P(high reward)
guided = bad
guided = good
B
H = 1
H = 6
H = 15
rats
P(high reward)
guided = bad
guided = good
C
H = 1
H = 2
H = 5
H = 10
humans
P(switch)
guided = bad
guided = good
trial #
D
H = 1
H = 6
H = 15
rats
P(switch)
guided = bad
guided = good
trial #
trial #
trial #

Figure 6

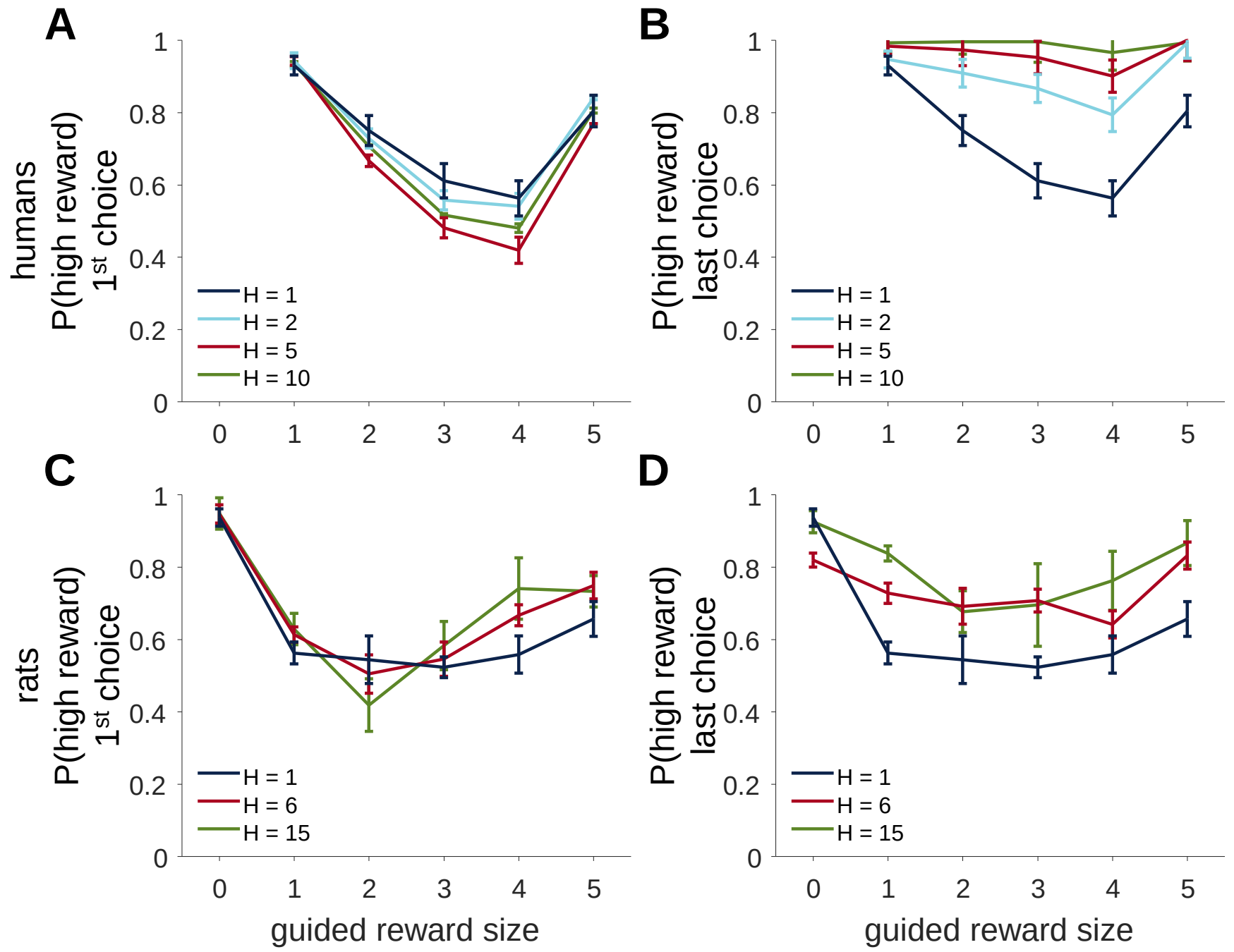

A
B
C
D
humans
rats
P(high reward)
1st choice
last choice
H = 1
H = 2
H = 5
H = 10
H = 6
H = 15
guided reward size
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5

Figure 7

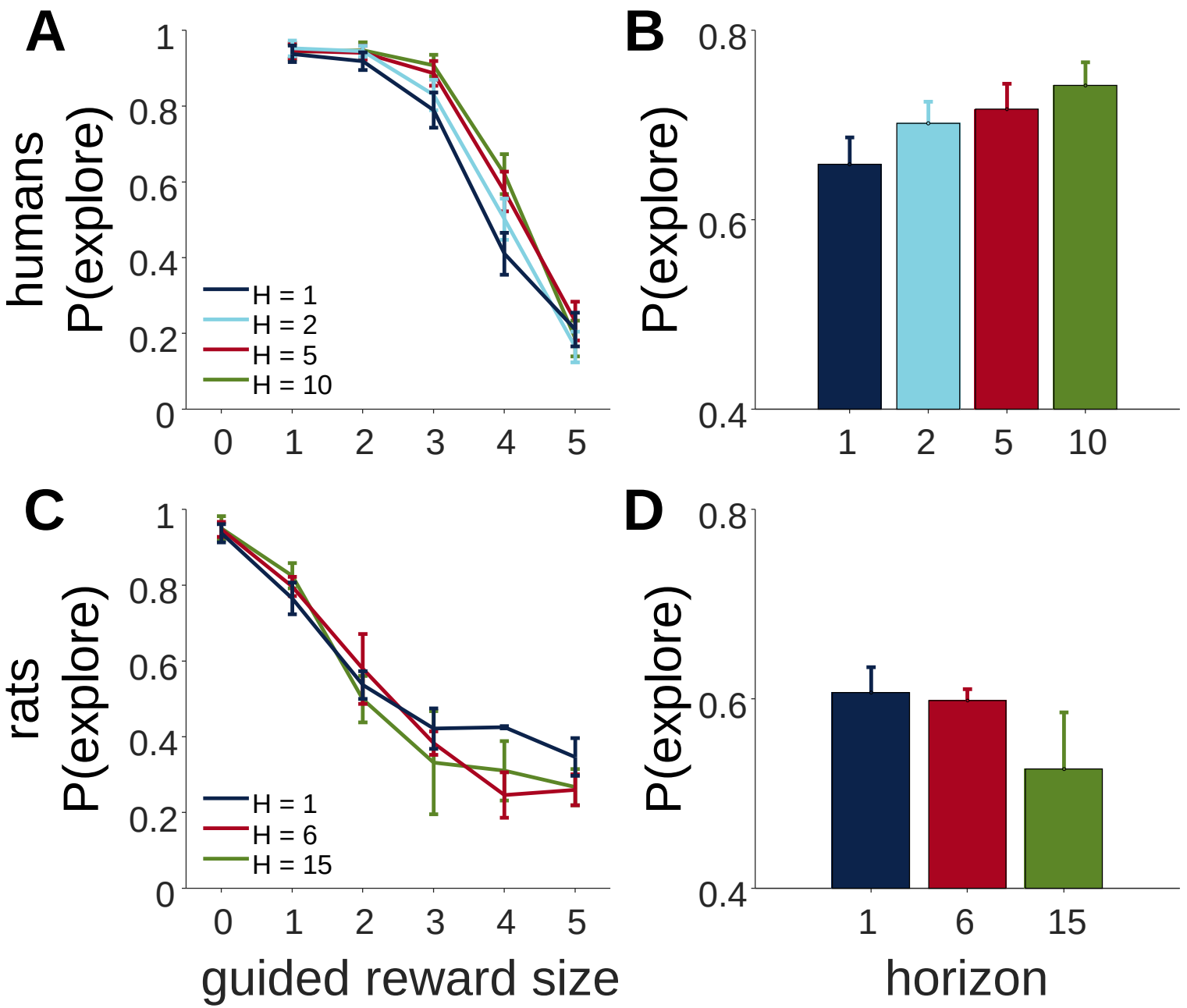

A
humans
P(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
H = 1
H = 2
H = 5
H = 10
B
P(explore)
0.8
0.6
0.4
1 2 5 10
C
rats
P(explore)
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
H = 1
H = 6
H = 15
guided reward size
D
P(explore)
0.8
0.6
0.4
1 6 15
horizon

Figure 8

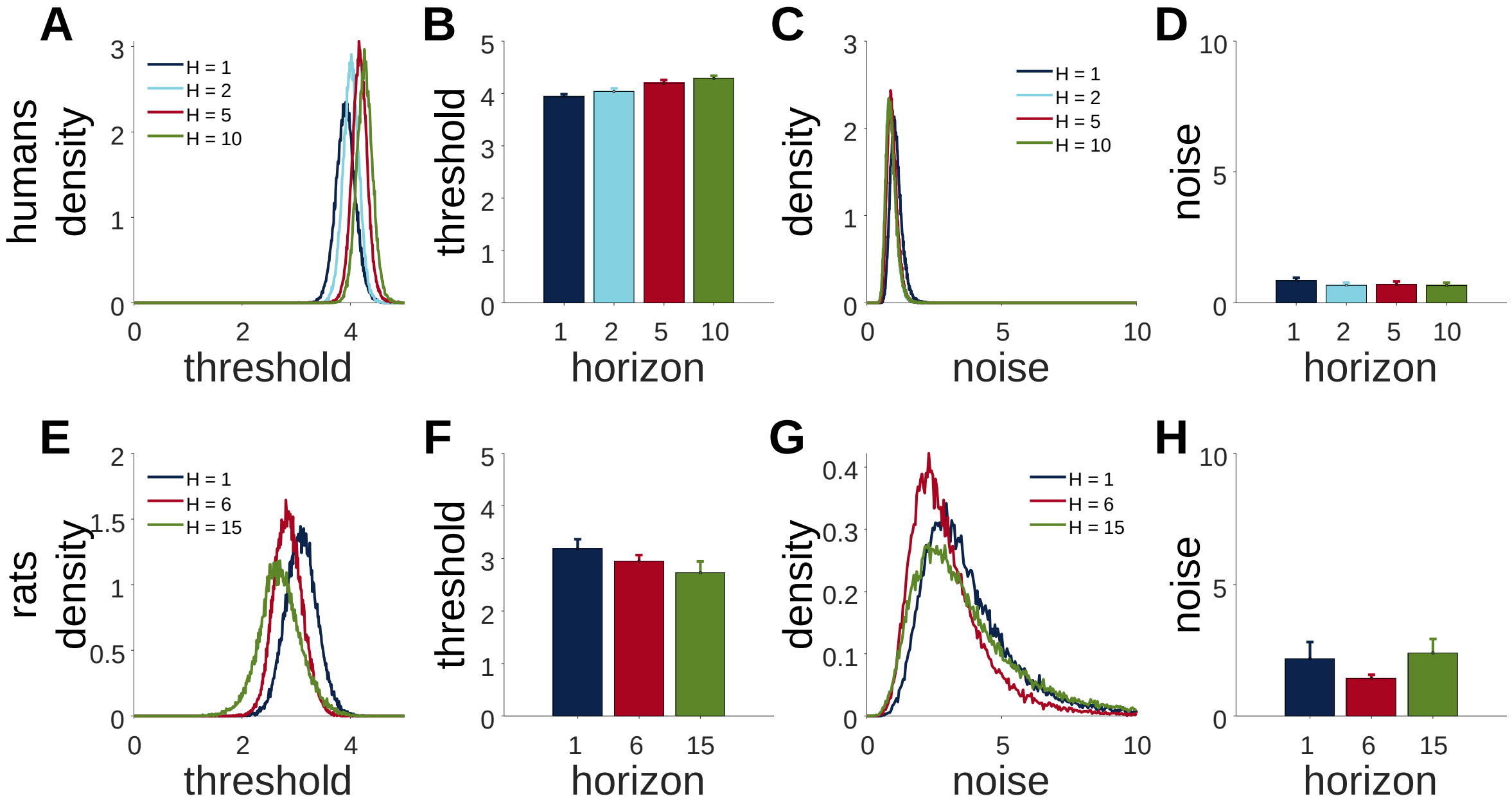

A
humans
density
H = 1
H = 2
H = 5
H = 10
threshold
B
threshold
horizon
C
density
noise
D
noise
horizon
E
rats
density
H = 1
H = 6
H = 15
threshold
F
threshold
horizon
G
density
noise
H
noise
horizon

Figure 9

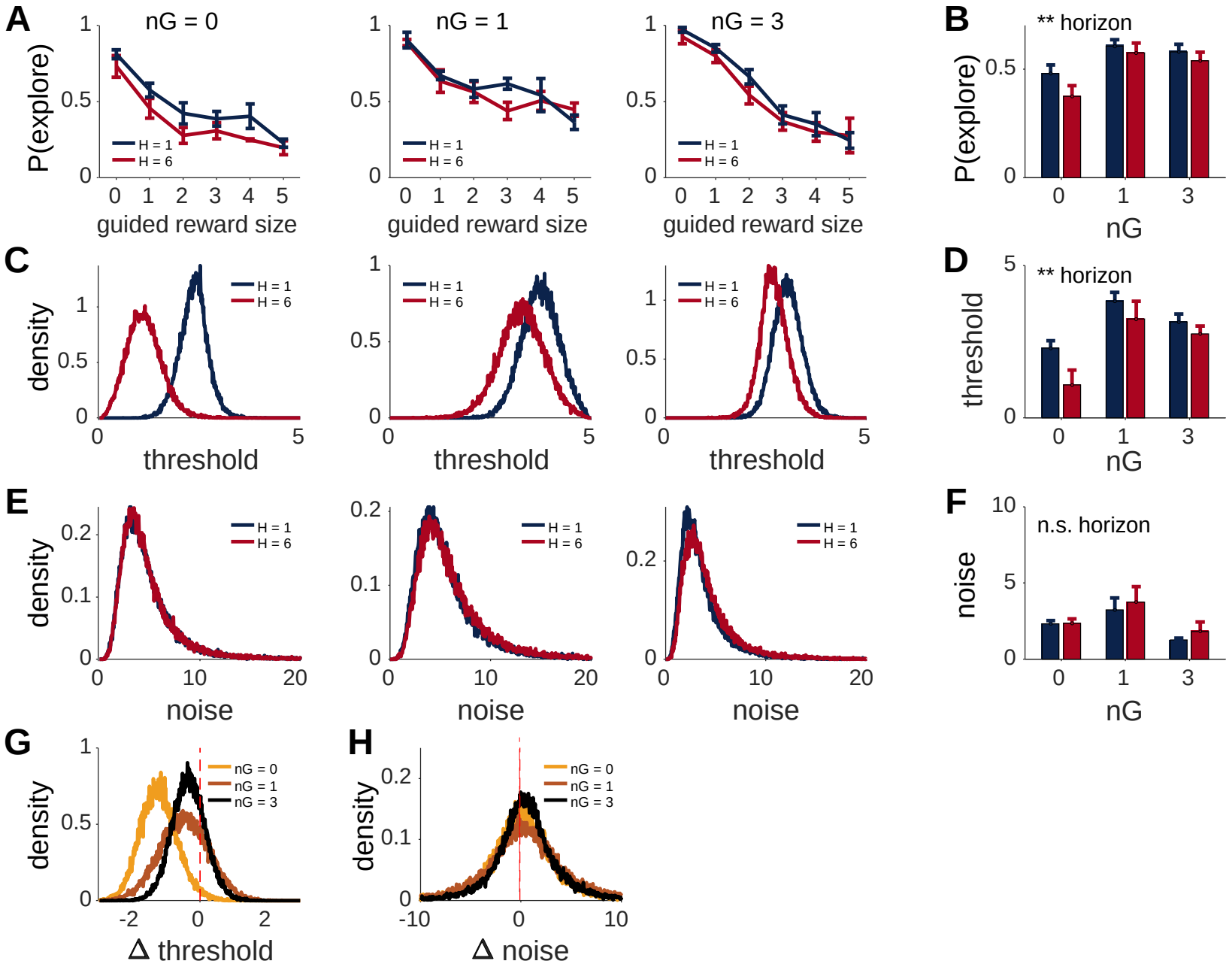

A
nG = 0
nG = 1
nG = 3
P(explore)
guided reward size
H = 1
H = 6
B
** horizon
P(explore)
nG
C
density
threshold
D
** horizon
threshold
nG
E
density
noise
F
n.s. horizon
noise
nG
G
density
Δ threshold
nG = 0
nG = 1
nG = 3
H
density
Δ noise

Figure 10

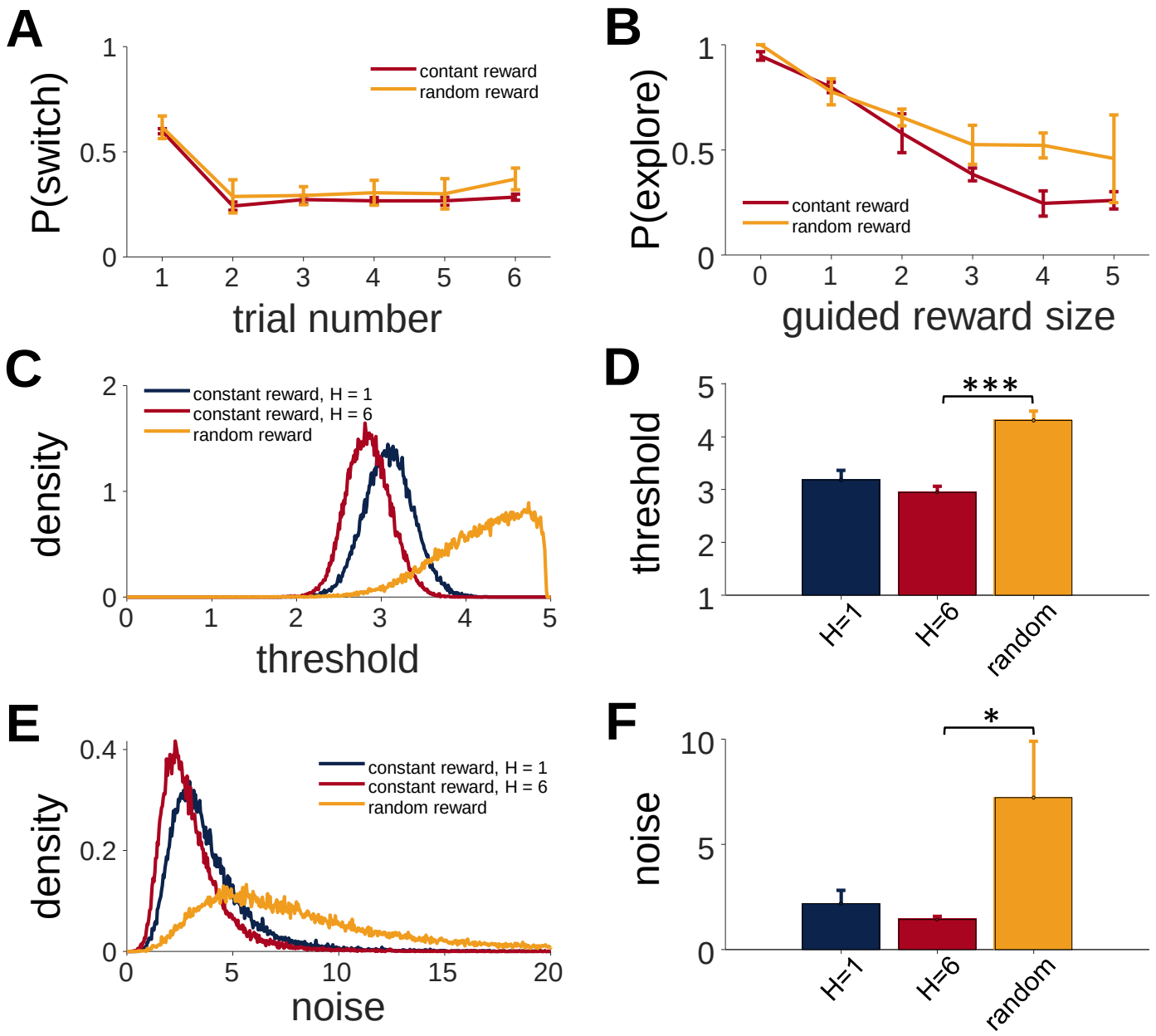

A
P(switch)
contant reward
random reward
trial number
B
P(explore)
contant reward
random reward
guided reward size
C
density
constant reward, H = 1
constant reward, H = 6
random reward
threshold
D
threshold
***
H=1
H=6
random
E
density
constant reward, H = 1
constant reward, H = 6
random reward
noise
F
noise
*
H=1
H=6
random

Figure 11

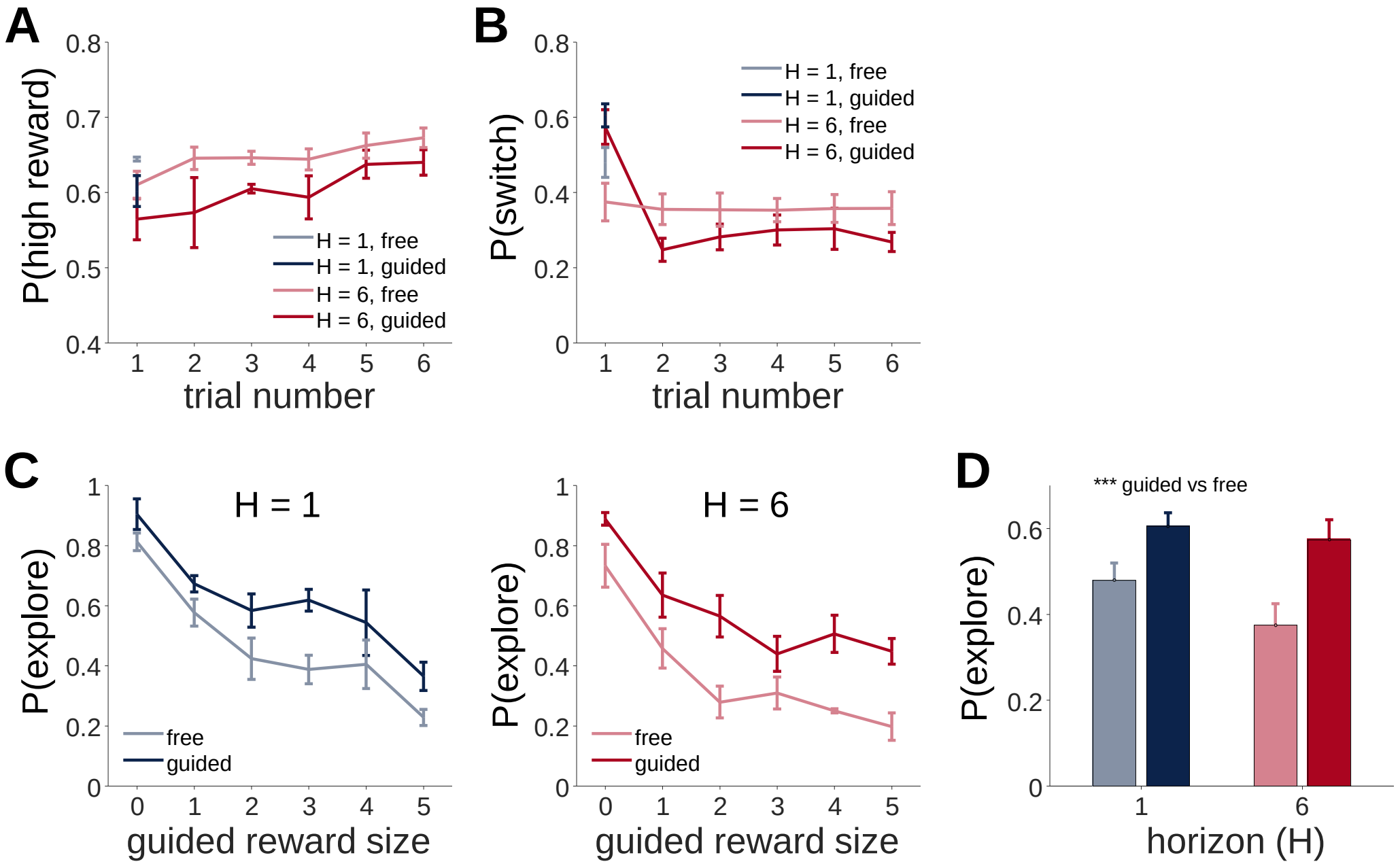
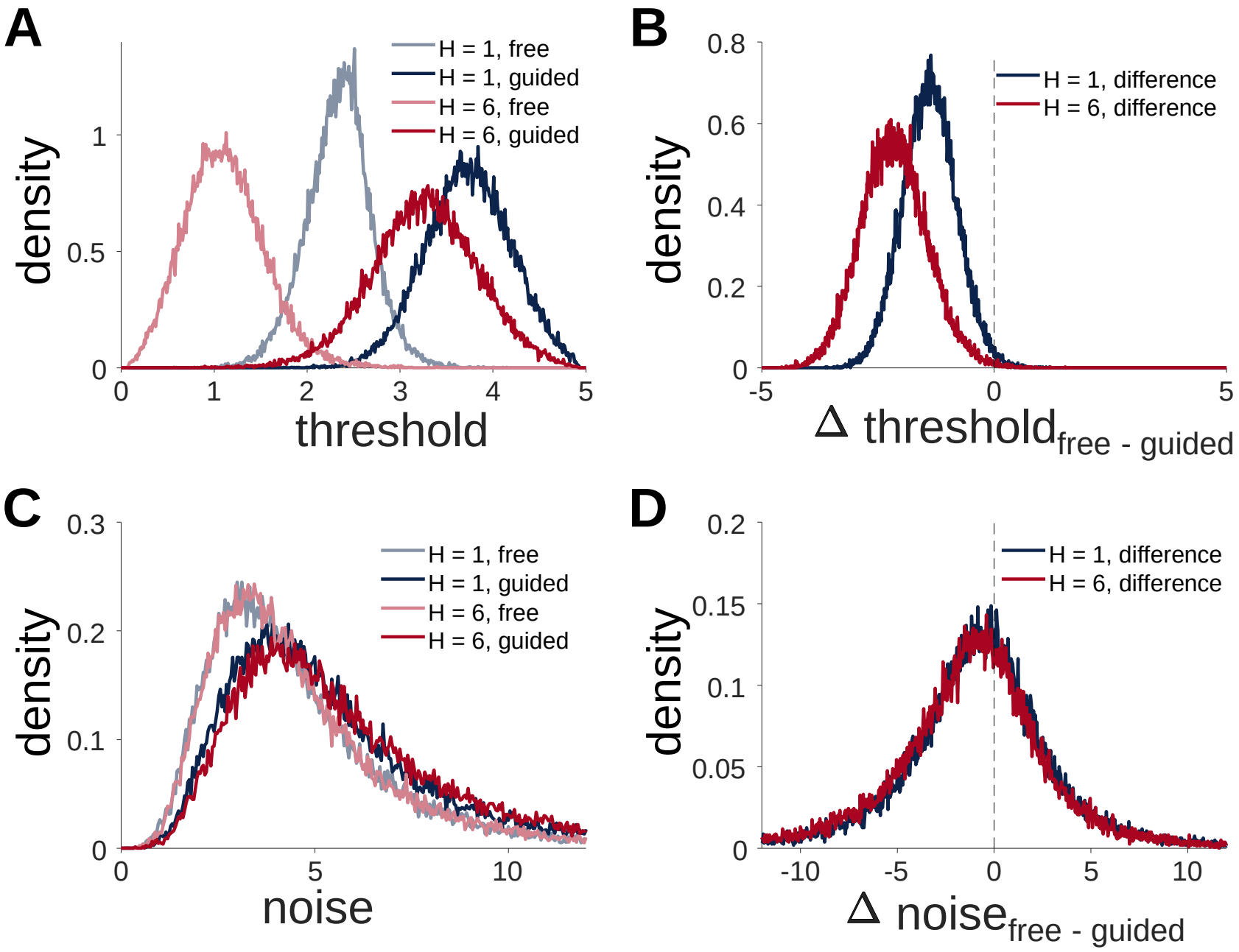

A
P(high reward)
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
6
trial number
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
B
P(switch)
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
trial number
H = 1, free
H = 1, guided
H = 6, free
H = 6, guided
C
H = 1
P(explore)
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
guided reward size
free
guided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

Figure 12


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

Supplementary figures

Figure S1

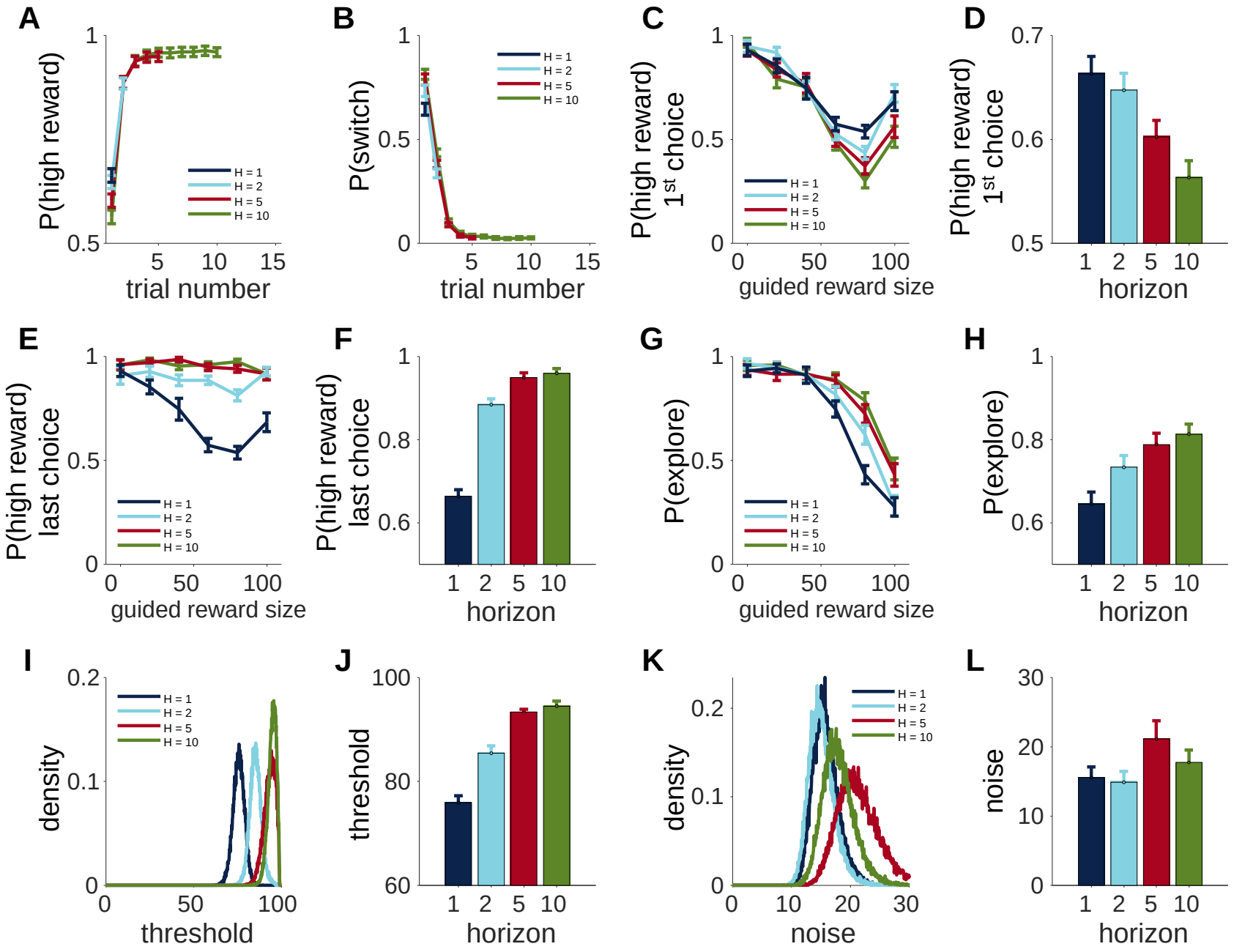
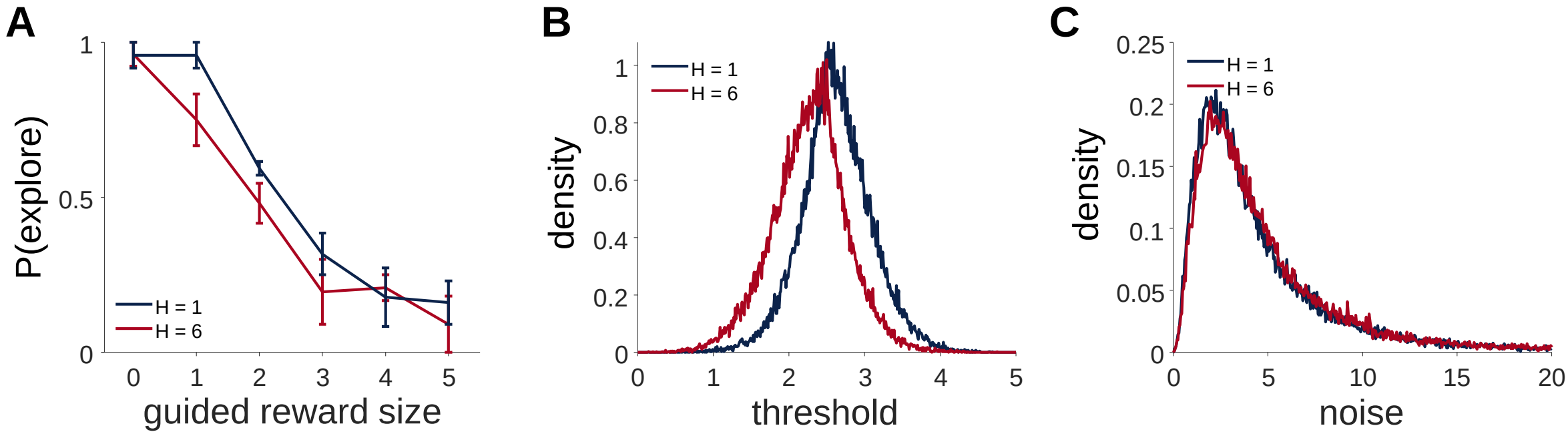

A
P(high reward)
trial number
H = 1
H = 2
H = 5
H = 10
B
P(switch)
trial number
C
P(high reward) 1st choice
guided reward size
D
P(high reward) 1st choice
horizon
E
P(high reward) last choice
guided reward size
F
P(high reward) last choice
horizon
G
P(explore)
guided reward size
H
P(explore)
horizon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

Figure S2


A
P(explore)
guided reward size
H = 1
H = 6
B
density
threshold
H = 1
H = 6
C
density
noise
H = 1
H = 6