An investigation of the effects of horizon and guided choices on explore-exploit decisions in rodents

Siyu Wang1, Blake Gerken2, Julia R. Wieland2, Robert C. Wilson1,3, and Jean-Marc Fellous1,4,5

1Department of Psychology, University of Arizona

2Neuroscience and Cognitive Science Program, University of Arizona

3Cognitive Science Program, University of Arizona

4Program in Applied Mathematics, University of Arizona

5Department 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](mailto:fellous@arizona.edu)

Abstract: 197 words

Number of figures: 8

# 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 mechanism underlying how humans and animals solve the explore-exploit dilemma, a good animal behavioral model is critical. Most previous explore-exploit studies in rodents used ethologically unrealistic operant boxes.Reversal learning paradigms in which the decision to abandon a currently bad option is confounded by the need for exploring a novel option for information collection, make it difficult to separate different drives and heuristics for exploration. Model-free win-stay lose-shift strategies are efficient in handling the reverse learning paradigms in that they discourage exploration, Such approaches however do not allow for observing model-based exploration behaviors, such as utilizing prior information or adaptation based on the volatility of the environment. In this study, we investigated exploration using a spatial navigation Horizon Task (Wilson et al., 2014) adapted to rats to address the above limitations. We compared the rat performance to that of humans using identical measures. Given the similarities and important disparity observed between human and rats, we revealed a more complex nature of explore-exploit behavior than previously thought.

Keywords: explore-exploit dilemma, directed and random exploration, win-stay lose-shift, 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 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, while animals face it when deciding whether to explore and forage for food, territory or mate. The cognitive ability to balance exploration and exploitation is vital to animal and human’s survival and success. In recent years, the study of explore-exploit decisions in humans and animals have become an active field of investigation (Mehlhorn et al., 2015, Wilson et al., 2020).

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 (Krebs et al., 1978, Meyer and Shi, 1995, Banks et al., 1997, Frank et al., 2009, Steyvers et al., 2009, Lee et al., 2011, Payzan-LeNestour and Bossaerts, 2012, Zhang and Yu, 2013, Wilson et al., 2014) . The second is an error-driven heuristic known as *random exploration*, in which exploratory actions with suboptimal estimates of reward value are chosen by chance (Kao et al., 2005, Badre et al., 2012, Wilson et al., 2014, Gershman, 2018, 2019). In particular, recent studies showed that humans were able to adapt the extent to which they explore 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.

Relatively few studies have investigated how animals, in particular rodents, make explore-exploit decisions. To study such behavior, most rodent explore-exploit studies took the approach of 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 costs (Beeler et al., 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 (Parker et al., 2016, Cinotti et al., 2019, Verharen et al., 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. Rodents have been reported to use a set of win-stay/lose-shift strategies which are effective in solving these reversal learning problems.

Reversal learning designs have however several limitations. First, the scope of ”exploration” being examined is limited, as win-stay/lose-shift is a model-free exploration strategy which works well for reversal learningbut does not allow for observing planning and model-based behavior. Second, the decision to abandon a currently bad option is confounded by the need for exploring a novel option for information collection. In such paradigms, it is impossible to dissociate directed choice from random exploration. Third, 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](#ENREF_22)). 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 general gap between the human and the rodent literature in our understanding of the explore-exploit decision processes. The complexity of the tasks and their implementation are different across species, and whether similar heuristics are in play in human 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 this complex process are sorely needed.

# Methods

## Animals

6 Brown Norway rats were used in these experiments. All rats were male between 6 and 7 months of age 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

Data from forty-six participants were used. All participants were from the undergraduate psychology subject pool and earn academic credits for their participation in the study. The human experiments were approved by the University of Arizona Institutional Review Board.

## Behavioral tasks

Rat version: The maze consisted of a circular area (1.5 m diameter) with 8 equidistant feeders at its periphery (Jones et al., 2012, Jones et al., 2015). Each feeder delivered sugar water (0.2g/L) 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. One feeder was the home base; the two others, equidistant from the home base, were the reward areas**.** The home base was never rewarded, but animals had to reach it to trigger/activate the 2 reward areas. 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 1A, blue rectangles). At the start of each game, depending on the conditions, the two reward 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 (‘Game1’, Fig 1A). Before making their free choices, rats were guided to one of the reward feeders in the first N trials (i.e. only one LED was blinking, Trial1 cue, Fig 1A). Rats performed versions where N = 0, 1, or 3. (In cases of N = 0, rats were not guided to any target feeder and started with a free choice between 2 reward feeders instead.) Fig 1A illustrates the version with N = 1. From the N+1st trial, they were cued to make free choices (the LED of the 2 rewarded feeders blinked simultaneously, ‘Trial2 cue’ Fig 1A).

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 (Game2, Trial1 Start, Fig 1A). The new rewarded feeders are the feeders opposite to the new home base (Trial1 cue, Fig 1A).

Each home base was associated with a fixed horizon in each game. Horizon is defined as the number of trials after the guided trial {Wilson, 2014 #23}. In these experiments several conditions were implemented: Horizon 1: only one free choice allowed before a new game started, Horizon 6: 6 free choices before the start of a new game, and in some cases Horizon 15. Horizon conditions were pseudo randomly chosen every day. On average, 16.1 games were run per day.

Rats performed several variations of the task described above. the different variants of the task are as follows:

1. (between-session version) Rats performed games of different horizons in blocks of weeks. In the same session, both home bases are associated with the same horizon. Different horizon conditions are run on different blocks of consecutive sessions. Rats are guided 3 times before a free choice can be made.
2. (within-session version) In the same session, one home base is always associated with the short horizon game (H = 1), whereas the other home base is always associated with the long horizon game (H = 6). Rats are guided 3 times before a free choice can be made.
3. (randomized value) Both feeders are associated with the long horizon games (H = 6). However, instead of having a fixed number of rewards drops at each of the reward areas rats obtain a random reward from 0 to 5 drops is given independently at each visit. In this case, there is nothing to learn. The reward contingency is completely random. Rats are guided 3 times before a free choice can be made.
4. (cued horizon length) Games of either short or long horizon can occur in either of the home bases. In short-horizon games, the rat will hear a low pitch sound before the free choice, whereas in the long-horizon games, the rat will hear a high pitch sound before the free choice. Rats are guided 3 times before a free choice can be made.
5. (free vs guided choice) Rats are guided either 0 or 1 times before a free choice can be made. For short horizon games, the rat either makes 2 free choices, or is guided once and makes 1 free choice. For long horizon games, the rat either makes 7 free choices, or is guided once and then makes 6 subsequent free choices.

Human version: In this task, 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 1B) 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, Horizon=2 in Figure 1B) and each row represented a trial. Before participants made their own choices, in the very first trial, they were cued to pick one of the bandits (Trial1 cued guided, Fig 1B). 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 (3- XX, Fig 1B). After 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, 9 free choices) and games with different horizons were pseudo-randomly interleaved. Humans ran 160 games per participant (with the exception of 4 early participants running 80 games).

Bayesian analysis

We used Bayesian analysis to quantify information-driven exploration and noise-driven exploration for both the human and the rodent data. We model how humans and rats make their first free choices.

To model choices on the first free-choice trial, we assume that subjects made decisions by computing the difference ∆Q between the reward value of the guided option, and an expected value of the unknown option based on prior experience, choosing to explore the unknown option when ∆Q < 0, and exploit the guided option otherwise. Specifically, we write

(1)

(2)

where, is the reward value of the guided option, is the exploitation threshold ( the higher , the more likely to explore), is the spatial bias, is 1 when the guided side is left and is -1 when guided side is right, is the decision noise.

We used hierarchical Bayesian analysis to fit the parameters of the model. In particular, we fit values of the exploitation threshold , spatial bias b and decision noise . The model was fit to the data using a Markov Chain Monte Carlo algorithm implemented in 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 8000 samples from the posterior distribution over parameters (2000 samples per chain). Each chain had a burn in period of 2000 samples, which were discarded to reduce the effects of initial conditions, and posterior samples were acquired at a thin rate of 1.

# Results

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

Both humans and rats were able to choose the best option (p(best), the option with a higher magnitude between the two available sugar water locations for rats, or the slot machine with a higher payout for humans) significantly above chance for the last choice in the long horizon games. Humans could achieve an accuracy of over 90% (Figure 2A, left) whereas rats could achieve an average accuracy around 80% (Figure 2B, left). Both improved with the number of trials given (rats: p(best, trial 6, H6) vs p(best, trial 1, H6), p = 5.89e-6).

Rats explored at a significantly higher level at trial 1 and then adopted a more constant and lower rate of exploration for later trials (~20%, Figure2B right), whereas humans explore more at trial 1 and 2, and eventually stopped switching (5%, possibly due to boredom or motor error, Figure 2A right).

These results may be partly explained by the deterministic nature of the reward delivery in the rat experiment. This suggests that after exploration, rats only explore at a constant lower rate in the exploitation phase. It is interesting to note that in a longer horizon, rodents showed a lower rate of exploration than in shorter ones. Rats explore about 30% in H = 6, and about 20% in H = 15 (Figure 2B right).

For humans, reaction time decreased with trial number and reached a plateau after the 3rd trial (0.1s, Figure 11B, right), indicating that the decision process was mostly settled during the first 2 trials. For rats, reaction time is computed as the time difference between home base arrival and target feeder release. Since travel time was included in the rats’ measure, this travel time may have overshadowed the decision reaction time difference proper, and we do not see significant changes in reaction times as trial number increased (Fig 11B, left). Although, there seems to be a decrease in reaction time between the rat’s first choice after being guided to a larger reward (4,5 drops) and that when guided to a small reward (0, 1 drops) (Fig 11A, left). This is may be due to the speed difference in travel time: with a larger reward, the rat explores less, has an overall higher motivation to run, and has a shorter travel time.

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

dA and 2BThe fact that the average accuracy was significantly above chance in the first non-guided trial showed that if prior information was available, humans and rats used it to guide subsequent exploration. In this particular task, with repeated games, humans and animals were able to assess the relative ‘goodness’ of the two bandits from the reward they obtained during the guided trial. For long horizons, both humans and rats performed above chance regardless of the reward size during the guided trial, but their performance was not uniform and displayed an inverted U shape (Figure 3A and 3B left). When they were guided to 0 or 5 drops (or 1 and 5 points for humans), the accuracy was maximized whereas the accuracy was lower when they were guided to more ambiguous reward amounts such as 2 or 3 drops.

**As with humans, rats can adapt the extent to which they explore based on the reward of the guided choice.** We computed the probability of freely choosing the bandit that was not selected when they were guided (p(switch) at the first free choice available, Figure 3B right) as a function of the reward size during the guided trial. We found that rats were very likely to explore if they obtained a low reward during the guided trials (e.g. 0 drops, mean = 95.4% Fig 3B right), and were very unlikely to switch if they obtained a large reward (e.g. 5 drops, about 30%, Fig 3B right). When rats were guided to the option with a objectively lower reward, rats switch their choices at more than 60% in the first free choice (Figure 4B left, trial 1, h=1,6,15 bad), whereas if they are guided to options with objectively higher rewards, they only switch around 40% in the first free choice (Figure 4B, left, h=1,6,15 good). Humans switch 80% on a suboptimal guided reward, and ~50% on the optimal guided reward (Figure 4A Left). Interestingly, when guided to a bad option at first, both rats and humans show a better accuracy in later trials compared to when guided to a good option (Figure 4A and 4B left). For humans, when they are guided to a good choice and switching on the 1st free choice to find out that the alternative is worse, they immediately switch back on the 2nd choice (Figure 4A Right), we did not see this in rats, perhaps due to their limited working memory capacity. Unlike in probabilistic tasks, we showed that rats can use prior information to explore in a win-stay/lose-shift fashion. Both “stay” and “shift” are outcomes of a comparison between the current reward and a prior threshold, and are not directly associated with a gain of reward vs an absence of reward, as in the probabilistic exploration tasks.

**Humans explore more than rats in longer horizons,. Humans behave more unpredictably than rats in longer horizons (but similarly for horizon 1).**

In line with previous research, humans switched more in long horizons than in shorter ones (Figure 3 3A right, h=9,5 Vs h=1,2), however for rats, the long horizon condition seemed to yield slightly lower probability of switching than the short horizon condition (Figure 3B right, h=15 vs h=1). This becomes clearer if the rat was only guided once instead of 3 times or if both horizon conditions occur during the same session (Figure 7 and 10). This can be confirmed via modeling. Posterior distributions over the exploitation threshold and decision noise for both humans and rats are shown in Figure 5. For humans, we observed an increase of threshold as horizon increases (Fig 5A, left), compatible with previous findings in the human horizon task (Wilson, 2014). In other words, in longer horizons, humans explore more 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 other variants of the rat task, see Figure 7 and 11). Rats decrease their thresholds as horizon increases, thus, they explore less in their first free choice when the horizon is long.

This opposite dependence of directed exploration on horizon in rats can arise from several factors. First, the utility of 1 to 5 drops is different for humans vs rats. For humans, they get points, whereas rats are getting real sugar water proportional to the number of drops. As a result, human subjects were at ceiling for 1 and 2 points indicating that both reward sizes were equally salient (Figure 3A, right), whereas 4 and 5 drops are similar to rats likely due to perceptual senses (Figure 3B right). Second, the efforts humans spend in making the decision is negligible, as a result, they over-explore to find out the best possible action, whereas rats have to physically travel the maze to get sugar water, they therefore likely under-explore to secure a satisfiable amount of return for each visit.

While is theoretically tied to directed exploration, random exploration is tied to . For humans, except for H = 1, as reported in Wilson et al 2014, decision noise increases as a function of horizon. For the shortest horizon however, decision noise is actually higher than the intermediate horizons. For rats, the change of decision noise with horizon is more complex, possibly because of the manner they are trained. Because rats, unlike humans, had to perform the different horizon conditions in chunks of consecutive days, the amount of training a particular rat is exposed to before each condition will influence decision noise. Both completely random behavior and noise-driven random exploration are potentially affected. As a result, the within-session version of the task was designed to address this concern.

In the within-session variant of the rat task, within the same session, one home base is always associated with short-horizon games (h = 1) whereas the other home base is always associated with long-horizon games (h = 6). In this design, there is no confound of learning/training effect. The model-based results show that the threshold is lower in horizon 6 compared to horizon 1 whereas decision noise is not modulated by horizon (Figure 6C left).

All horizons are interleaved in the human version. To make the rat version even closer to the human version, we played a sound before the start of each game and during the guided trials to cue the rat the horizon condition associated with a home base. By using the sound cue, we no longer have the fixed association between a particular feeder and a given horizon. Within a session, each home base can be associated with different horizon conditions. In this version, we also observed that threshold decreases as a function of horizon (Fig 7C, left). Decision noise is also quantitatively smaller.

**The verall level of uncertainty boosts exploration in cases of “win”.**

It has been proposed that relative uncertainty correlates with directed exploration whereas total uncertainty correlates with random exploration (Gershman et al). In order to assess the effect of uncertainty in driving exploration, in one variant of the task, instead of using deterministic rewards, each feeder will give a random reward from 0 to 5 drops that change independently from trial to trial.

In this version, since there is no information that can be learned and the rewards are completely random, rat’s accuracy is only at chance around 50% (Figure 8A, left). After the guided choices, they still explore more on the first free choices than on subsequent ones, suggesting that the novelty of the unknown feeder itself rather than the potential better reward can drive exploration (Fig 8B left). Regardless of the average guided reward, p(switch) is higher across all guided conditions(See Figure 8B right). The difference is largest when rats are guided to 4 and 5 drops, in the unpredictable environment. The rat still switches at about 50%, even when the average guided trial experience includes the best reward condition (i.e. 4,5), significantly higher than the deterministic case. For later choices, the overall level of switching is also higher compared to case of deterministic rewards (Fig 8A right). In an unpredictable environment, rats do increase their rate of switching. This can potentially account for the horizon difference in p(switch) in Figure 2B rightwhere there is a lower rate of switching in H = 15 compared to H = 6, possibly due the fact that the environment is more predictable and certain in the H = 15 case.

If rats are guided to the low-reward feeder, their accuracy seems to be slightly higher than if they experienced high-rewards during the guided choices. (Figure 8B Left)

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

Finally, we investigated whether self-driven exploration is any different from guided exploration during the first trial of a game. Do rats behave any differently if they are guided by sensory cues on the first trials, or if they are instead invited to choose randomly? Specifically, in two separate weeks, rats performed both a version in which they were guided to one feeder once before freely choosing between the 2 options (Guided), and a version in which they start off with 2 options to choose from (Free initial choice). In the analysis, we treated the 1st choice in the Free choice version as if it was guided (i.e. self-guided by the rat itself, instead of by the blinking LED), and treated the 2nd choice as choice number 1 (Fig 9,10).

We found that rats overall explored more on their first free choices if they were guided (Fig 9B right), but their overall accuracy was higher if they were allowed to make free choices compared to when they were guided (Fig 9A left). As a result of less switching in the Free choice version, the accuracy when guided to 4 or 5 drops was higher (Fig 9B left). In the Guided version, even if rats were guided to a 5 drop feeder, they still switched 40% of the times as opposed to 20% in the free-choice version. Moreover, when they were guided, they explore more on the first choice as in other variants of the task. However, when they choose freely, the 2nd choice did not look different any more, and rats seemed to have reached a steady rate of exploration, at a rate higher than the Guided condition (Figure 9A right).

Regardless of whether they were guided or not, rats tended to explore more in a short horizon context than in the long horizon context, which is the opposite of what a human participant would do (Fig 9A and 9B right). By quantifying threshold and decision noise using the Bayesian model, we found that threshold in horizon 6 is lower than horizon 1, regardless of whether the rat is guided or not. In addition, threshold in Free-choice condition is lower than in Guided condition. In other words, when they are guided, they explore more in the first free choice. Interestingly, although decision noise doesn’t change much with horizon, when self-guided, rats behave less randomly by having a lower decision noise term compare to when they are guided, their bias is also smaller.

<something need to be said about Fig 10 here>

Discussion

In this study, we investigated the behavioral performance of rats in a new model of the Horizon task adapted to rats. We addressed the limitations of previous rodent studies by designing a novel open-field task in which rodents choose between two locations that offer different amounts of rewards. To dissociate the uncertainty in estimation of value from the ambiguity of a novel option, we manipulated the magnitudes of rewards rather than their delivery probabilities. Indeed, choosing the a different option in the probabilistic case can be, at least in part, exploratory whereas repeating the same choice is more exploitative in a purely deterministic reward setting. In our design, the rats were guided to one of the 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. In this way, we are also able to quantify directed exploration in a model-free manner. In addition, rats were set to perform the task in both a short and a long horizon condition to assess whether they explored differently in different horizon contexts. Using an open field task, we were able to use two sets of different locations alternatively as new games start as opposed to switching the reward conditions at the same set of locations (as in reversal learning paradigms). The two sets of feeders were associated with different horizon contexts. Moreover, we recruited human subjects to perform a version that is comparable to the rat task, and we compared the performance between humans and rats.

As with humans, we showed that rats were able to use prior information about the distribution of rewards to guide future exploration. However, rats did not alter their exploration strategies based on the horizon context to the same extent as humans did.

In line with previous research using the reversal learning paradigm in an operant box, we showed that rats demonstrated the model-free win-stay/lose-shift strategy in making explore-exploit decisions. Specifically, they chose to explore more when the exploit value was low and explored less when the exploit value was high. However, in our design, in order to assess whether the exploit value was low or high, instead of using short-term memory to recall the value at the exploit option before reversal, rats had to use their long-term memory from 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 addition, in the reversal learning paradigm, the level of exploration had to be evaluated on the course of several trials (in the probabilistic case), therefore the exact timing of “exploration” decision was difficult to estimate. In our case, however, exploration can be seen in a single trial (visiting the unknown option), which is advantageous in studying the neurophysiological mechanisms underlying real-time explore-exploit decisions and its neural substrate.

We applied Bayesian statistics to quantify directed vs random exploration in both the rat and the human datasets. Unlike humans, rats showed an opposite adaptation of behavioral strategy to the horizon context. The use of horizon context to facilitate exploration requires planning and model-based reasoning. Both an optimal agent and our human participants would explore more in long horizon contexts. However, rats explored more in short horizon context. This may be a combination of their tendency to use a model-free learning strategy (that shows no horizon adaptation) and low motivation in short horizon contexts because they overall gain less reward from the short horizon context (that increases randomness in rat’s behavior which would result in exploration). For humans, it took over 40 participants to show a robust horizon effect. With only a small number of rats, statistical power with the limited sample size limits the interpretation of the horizon modulation of exploration in rats, and more work is needed.

Finally, we observed an interesting difference in the exploration strategy between when the first choice is self-driven vs 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 more in the first free choice when they were guided first, but this was not observed when they were exploring by themselves. A similar phenomenon was recently reported in a human explore-exploit study (Sadeghiyeh et al., 2018). 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.

# References

Badre D, Doll BB, Long NM, Frank MJ (2012) Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron.

Banks J, Olson M, Porter D (1997) An experimental analysis of the bandit problem. Economic Theory.

Beeler JA, Daw N, Frazier CRM, Zhuang X (2010) Tonic dopamine modulates exploitation of reward learning. Frontiers in Behavioral Neuroscience 4:1-14.

Bellman R (1954) The Theory of Dynamic Programming. Bulletin of the American Mathematical Society.

Cinotti F, Fresno V, Aklil N, Coutureau E, Girard B, Marchand AR, Khamassi M (2019) Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific Reports 9:1-14.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience.

Gershman SJ (2018) Deconstructing the human algorithms for exploration. Cognition 173:34-42.

Gershman SJ (2019) Uncertainty and exploration. Decision.

Jones B, Bukoski E, Nadel L, Fellous JM (2012) Remaking memories: reconsolidation updates positively motivated spatial memory in rats. Learning & memory 19:91-98.

Jones BJ, Pest SM, Vargas IM, Glisky EL, Fellous JM (2015) Contextual reminders fail to trigger memory reconsolidation in aged rats and aged humans. Neurobiology of learning and memory 120:7-15.

Kao MH, Doupe AJ, Brainard MS (2005) {C}ontributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Nature 433:638-643.

Krebs JR, Kacelnik A, Taylor P (1978) Test of optimal sampling by foraging great tits. Nature 275:27-31.

Laskowski CS, Williams RJ, Martens KM, Gruber AJ, Fisher KG, Euston DR (2016) The role of the medial prefrontal cortex in updating reward value and avoiding perseveration. Behavioural Brain Research 306:52-63.

Lee MD, Zhang S, Munro M, Steyvers M (2011) Psychological models of human and optimal performance in bandit problems. Cognitive Systems Research.

Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, Braithwaite VA, Hausmann D, Fiedler K, Gonzalez C (2015) Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. Decision.

Meyer RJ, Shi Y (1995) Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. Management Science.

Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, Daw ND, Witten IB (2016) Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nature Neuroscience 19:845-854.

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.

Sadeghiyeh H, Wang S, Wilson RC (2018) Lessons from a “failed” replication: The importance of taking action in exploration. PsyArXiv doi 10.

Steyvers M, Lee MD, Wagenmakers EJ (2009) A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology.

Verharen JPH, den Ouden HEM, Adan RAH, Vanderschuren LJMJ (2020) Modulation of value-based decision making behavior by subregions of the rat prefrontal cortex. Psychopharmacology 237:1267-1280.

Vodicka EL, Chung MH, Zimmermann MR, Kosgei RJ, Lee F, Mugo NR, Okech TC, Sakr SR, Stergachis A, Garrison LP, Jr., Babigumira JB (2019) Estimating the costs of HIV clinic integrated versus non-integrated treatment of pre-cancerous cervical lesions and costs of cervical cancer treatment in Kenya. PloS one 14:e0217331.

Wilson RC, Bonawitz E, Costa VD (2020) Balancing exploration and exploitation with information and randomization. 1-18.

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of Experimental Psychology: General.

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