Behavioral investigations of the explore-exploit tradeoff in rats

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: XXX words

Number of figures: XXX

# Abstract

Humans and animals have to balance the need for exploring new options and exploiting existing options that are known to be good. 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 the approach of a reversal learning paradigm which has certain limitations. In this paradigm, from 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. A model-free win-stay lose-shift strategy is efficient in handling the reverse learning which discourages and doesn’t allow for observing model-based planning behavior in exploration. Most of the previous experiments were done in chamber boxes which were not natural decisions for the rats ecologically. In this study, we investigated the exploration in rats using a Horizon Task ([Wilson et al., 2014](#_ENREF_24" \o "Wilson, 2014 #23)) adapted to rats to address the above limitations.

Keywords: explore-exploit dilemma, directed and random exploration, win-stay lose-shift, planning horizon

# Introduction

Humans and animals constantly face the dilemma of choosing between exploiting options that are known to be good and exploring unknown options in the hope of discovering better options for the future. Humans face it in scenarios from simple choices like deciding whether to explore a new restaurant for dinner, to important life decisions like 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 ([Mehlhorn et al., 2015](#_ENREF_15), [Wilson et al., 2020](#_ENREF_23)).

Although optimal solution to explore-exploit decisions is in general computationally intractable ([Bellman, 1954](#_ENREF_4)). Humans and animals are thought to use approximations or heuristics in making explore-exploit decisions. Previous research suggested both an information-driven heuristic known as directed exploration in which action is biased towards the more uncertain option ([Krebs et al., 1978](#_ENREF_12), [Meyer and Shi, 1995](#_ENREF_16), [Banks et al., 1997](#_ENREF_2), [Frank et al., 2009](#_ENREF_6), [Steyvers et al., 2009](#_ENREF_20), [Lee et al., 2011](#_ENREF_14), [Payzan-LeNestour and Bossaerts, 2012](#_ENREF_18), [Zhang and Yu, 2013](#_ENREF_25), [Wilson et al., 2014](#_ENREF_24)) and an error-driven heuristic known as random exploration in which exploratory actions with suboptimal estimates of value will be chosen by chance ([Kao et al., 2005](#_ENREF_11), [Badre et al., 2012](#_ENREF_1), [Wilson et al., 2014](#_ENREF_24), [Gershman, 2018](#_ENREF_7), [2019](#_ENREF_8)). In particular, ([Wilson et al., 2014](#_ENREF_24)) showed that humans are able to adapt the extent to which they explore with the horizon context, i.e. the number of future choices remaining. Horizon adaptation is thought to be a hallmark of exploration.

Relatively few studies have investigated how animals, in particular rodents, make explore-exploit decisions. To study such behavior, most existing 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, this can be options with high vs low costs ([Beeler et al., 2010](#_ENREF_3)), options with large reward and short delay vs small reward and long delay ([Laskowski et al., 2016](#_ENREF_13)) , or binary reward options with high vs low probabilities ([Parker et al., 2016](#_ENREF_17), [Cinotti et al., 2019](#_ENREF_5), [Verharen et al., 2020](#_ENREF_21)). As animals explore the two options they will eventually converge to the better option and keep exploiting that option, until the outcome of the two options are swapped. Deviating from the previously exploit option after reversal is considered exploration in these tasks. Rodents are reported to use a win-stay lose-shift strategies which is effective in solving these reversal learning problems.

However, these reversal learning designs have several limitations. Firstly, the scope of ”exploration” being examined using such designs is limited, as win-stay lose-shift is a model-free exploration strategy which works well for reversal learning, however, such design doesn’t allow for observing planning and model-based behavior. Secondly, from 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. Thirdly, most of the tasks mentioned above are implemented in operant boxes that are not natural environments for a rat. As ([Vodicka et al., 2019](#_ENREF_22)) pointed out recently, head-fixed monkeys have the opposite risk preference compared to freely moving monkeys during the same task, suggesting that decision making may be directly influenced by the physical constraints of the experimental paradigms. One of the most fundamental and natural behavior of rat is to spatially navigate. It is unknown how rats would behave in a setting in which the explore-exploit dilemma taps into their spatial navigation abilities. Fourthly, there is a general gap between the human and rodent literature in the context of the explore-exploit decision processes. The complexity of the tasks and their implementation are different across species, and that remains an open question whether similar heuristics are in play in human and rodents when making explore-exploit decisions. Finally, very little is known of the neural substrate of the explore-exploit decision circuits, and animal models allowing details investigations of the cellular and system mechanisms of this process are sorely needed.

# Methods

## Animals

4 Brown Norway rats were used in the experiment. 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 committee at University of Arizona and followed NIH guidelines.

## Human participants

Data from forty six participants were used in the experiments. All participants were from the undergraduate psychology subject pool who earn credits for participation in this study. The human experiment was approved by the University of Arizona Institutional Review Board.

## Behavioral Tasks

Rat version: The maze consisted of a circular area (1.5m diameter) with 8 equidistant feeders at its periphery ([Jones et al., 2012](#_ENREF_9), [Jones et al., 2015](#_ENREF_10)). Each feeder delivered sugar water (0.2g/L) in the form of computer controlled drops. A blinking LED was attached to each feeder ad acted as 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 the 2 rewards areas. The home base was flanked by two lego blocks so as to force the animal to start its navigation to the 2 choice feeder without bias (Fig 1A, 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 in each game, 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 target feeders in the first N trials (only one LED was blinking, Trial1 cue, Fig 1A). Rats performed versions where N = 0, 1, or 3. (in case of N = 0, rats were not guided and started with a choice between 2 unknown feeders). Fig 1A illustrates the version with N = 1. From the N+1th trial, they were cued to make free choices (the LED of the 2 rewarded feeders blinked simultaneously, ‘Trial2 cue’ Fig 1A) for either 1 trial (short horizon condition) or 6 trials (long horizon condition).

After the first game was completed, a 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 home base was now activated and signaled the start of a new game (Game2, Trial1 Start, Fig 1A). The new rewarded feeders are the activated 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 (ref). In these experiments two conditions were implemented: Horizon 1: only once free choice allowed before a new game started, and Horizon 6: 6 free choices before the start of a new game. Horizon conditions were pseudo randomly chosen every day. On average, 16.1 games were ran per day.

Human version: In this task, participants were sitting in a booth, in front of a computer. They were asked to choose between two slots machines (will also refer to as bandits, Fig 1B) that gave out a fixed number of rewards uniformly drawn from 1 to 5. Participants were instructed to maximize the total number rewards. The height of the boxes indicates the number of choices allowed in the current game (i.e. the horizon condition, Horizon=2 in Figure 1 B), each row represents a trial. Before participants make their own choices, in the very first trial, they were cued to pick one of the bandits. The options available was cued with a green background color. Participants indicate their choices by pressing the arrow keys on a keyboard. Their response was followed by an indication of how much rewards they obtained (3, XX, Fig 1B). From the 2nd trial, both bandits were available and participants were free to make their own choices. There wre four horizon conditions (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 doing 80 games).

# Results

Both humans and rats are able to choose the best option (p(correct), 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 in the last choice in the long horizon context. Humans can achieve an accuracy of over 90% whereas rats can achieve an average accuracy around 80% (Figure 2). Both improve with the number of trial given. Interestingly, in the first free choice, both humans and rats can achieve a similar average accuracy of 70% (Figure 2).

The fact that the average accuracy is significantly above chance of 50% in the first non-guided trial shows that prior information is available and used by human and rats to guide subsequent exploration. Both humans and rats perform above chance regardless of the reward size during the guided trial (Figure 3A). When they are guided 0 or 5 drops (or 1 and 5 points for humans), the accuracy is at ceiling whereas the accuracy is lower when they were guided to more ambiguous reward amounts like 2 or 3 drops. In longer horizons, by the last game, the accuracy also increased for positive drop sizes 1 to 5 with the exception of drop size of 0 for rats (Figure 3B).

We computed the probability of choosing a different bandit than the one they were guided to in their first free chocies (p(switch), Fig 4) as a function of the reward size during the guided trial. We find that rats were very likely to explore the other feeder if they obtained a low reward during the guided trials (e.g. 0 drops, mean = 95.4% Fig 4A), and were very unlikely to switch if they obtained a large reward (e.g. 5 drops Fig 4A). Interestingly, the long horizon conditions seemed to yield lower probability of switching than in the short horizon condition. This becomes more obvious if the rat is only guided once instead of 3 times (Figure 8B). Figure XX shows the rate of switching from the previous choice as a function of trial number in the game, rats seem to explore more on the first free choice and then revert to a constant rate of switching from the 2nd choice. The overall shape of the curve was similar in humans with some notable differences. First, the utility of 1 to 5 drops seems different between humans and rats that human subjects were at ceiling for 1 and 2 points indicating that both rewards size were equally salient (Fig 4). Second, increasing the horizon seem to increase the probability to switch for intermediate reward size 3,4 in Fig 4. This is in the same direction as what an optimal agent would behave while at the opposite direction of that of a rat. The effect size of the decrease of exploration with horizon is more obvious when rats were only guided once (Fig 8B).

We also looked at the percentage of switching from the previous choice as a function of trial number within a game (Fig 5). Rats explore at a significantly higher level at trial 1 and then adopt a constant rate of exploration for later trials (20%), whereas humans explore more at trial 1 and 2, and eventually stop switching (5%s, possibly due to boredom or motor error). For humans, reaction time also decreased with trial number and became constant from the 3rd trial (0.1s), indicating the decision process is mostly during the first 2 trials. For rats, since travel time is included in the rats’ reaction times, they may have overshadowed the reaction time difference in decision making, and we don’t see any significant change of reaction time as trial number increases (Fig 6B). Although, there seems to be a decrease in reaction time in the rat’s first choice after being guided to a larger reward (4,5 drops) compared to being guided to a small reward (0, 1 drops) (Fig 6A). This is likely due to the speed difference in travel time, with a larger reward, the rat explores less and has an overall higher motivation to run, and reaction time is also shorter.

Finally, we investigated whether self-driven exploration is any different from guided exploration. Specifically, in two separate weeks, rats performed both a version A in which they are guided once before choosing between the 2 options, and a version B in which they start off with 2 options to choose from. In the analysis, we treat the 1st choice in version B as if it was guided (self-guided by the rat itself), and treat the 2nd choice as trial number 1 (Fig 7,8). We found that rats overall explore more on their first free choices if they are guided (Fig 7B, 8B) and their overall accuracy is higher if guided (Fig 7A). Regardless of being guided or not, rats explore more in a short horizon context than in the long horizon context, which is the opposite of what an optimal agent or a human participant would do (Fig 8B).

Discussion

In this study, we investigated the behavioral performance of rats in a new model of the Horizon task adapted to rats. We addressed these limitations of previous rodent studies by designing a novel open-field task in which rodents choose between two locations that offer different amount of rewards. To dissociate the uncertainty in estimation of value from the ambiguity of a novel option, we used magnitudes of rewards instead of probabilistic rewards. Indeed, choosing the same option again 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 location first, and the extent to which they explore the other unvisited feeder location in their free choices is used as a purer 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 rats explore differently in different horizon contexts. Using an open field, we were able to use two sets of different locations alternatively as new games start as opposed to having to reverse the reward conditions at the same set of locations (as in reversal learning paradigms). The two sets of feeders are associated with different horizon contexts. Moreover, we recruited human subjects to perform a similar version that is comparable to the task that the rats, and we compared the performance in exploration 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.

In line with previous researches using the reversal learning paradigm in an operant box setting, we showed that rats demonstrated the model-free win-stay lose-shift strategy in making explore-exploit decisions. Specifically, they choose to explore more when the exploit value is low and explore less when the exploit value is high. However, in our design, in order to assess whether the exploit value is 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 has to be evaluated on the course of several trials (in the probabilistic case), the exact timing of “exploration” decision is difficult to estimate, whereas in our case, 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.

Further analysis needs to be done to quantify directed vs random exploration in both the rat and the human datasets. In a model-free manner, the probability of choosing the unknown options can be viewed as a tendency for directed exploration. However, as with the reversal learning paradigm, our design cannot fully dissociate directed exploration from random exploration without model fitting. Instead, we can quantify both directed and random exploration by incorporating an “information bonus” and a “decision noise” term using the softmax action selection rule to dissociate the tendency to choose the unknown option vs. the variability in behavior.

Here is the reward they are guided to, is the expectation based on an estimate of the average payout using prior information, IB is the “information bonus” which is a quantification of how biased the animal towards choosing the unguided option, bias is the spatial bias of choosing a particular physical side, = 1 if the guided trial is on the left, and = -1 if the guided trial is on the right, is “decision noise” which is a quantification of the level of randomness in the behavior. With this model fitting, we are able to evaluate how directed and random exploration are modulated differently by horizon context in both rats and humans.

Unlike humans, rats seemed to show an opposite adaptation of behavioral strategy to the horizon context. The use of horizon context to facilitate exploration requires planning and model-based reasoning. An optimal agent would explore more in long horizon context. 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 context as they overall gain less reward from the short horizon context (that increases randomness in rat’s behavior which would show up as exploration). For humans, it takes over 40 participants to show a robust horizon effect, with only a small number of rats, statistical power with the limited sample size will limit the interpretation of the horizon modulation of exploration in rats.

Finally, we observed an interesting difference in the exploration strategy between when the first choice is self-driven vs guided. This suggests a different neural mechanism underlying voluntary vs guided learning. Rats explore in the first free choice more when they are guided first, but this is not observed when they were exploring by themselves. A similar phenomenon is recently reported in a human explore-exploit study ([Sadeghiyeh et al., 2018](#_ENREF_19)). This rat model has the potential of probing the differential neural mechanism underlying active vs passive learning.

Overall, our novel design provides a potentially better behavioral paradigm to investigate explore-exploit tradeoffs in future electrophysiological studies.

# 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.