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

Keywords:

# 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, Wilson et al., 2020).

Although optimal solution to explore-exploit decisions is in general computationally intractable (Bellman, 1954), 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 (Banks et al., 1997, Frank et al., 2009, Krebs et al., 1978, Lee et al., 2011, Meyer and Shi, 1995, Payzan-LeNestour and Bossaerts, 2012, Steyvers et al., 2009, Wilson et al., 2014, Zhang and Yu, 2013), and an error-driven heuristic known as random exploration in which exploratory actions with suboptimal estimates of value will be chosen by chance (Brainard and Doupe, 2002, Gershman, 2018, 2019, Kao et al., 2005, Wilson et al., 2014). In particular, Wilson et al. (2014) 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), 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 (Cinotti et al., 2019, Parker et al., 2016, Verharen et al., 2020). 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 Chung et al (2020) 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 dilemmataps 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, andt 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-expoit 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 fourty 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 (Cite Jones et al 2012, 2015). 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 first in the first trial (only one LED was blinking, Trial1 cue, Fig 1A). From the 2nd 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. Up to XX 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.

# 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. In this particular task, with repeated games, humans and animals are able to assess the relative ‘goodness’ of the two bandits from the reward they obtained during the guided trial. We computed the probability of choosing a different bandit than the one they were guided to (p(explore), Fig 3) 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, 9X.Y% +- ZZZ Fig 3A), and were very unlikely to switch if they obtained a large reward (e.g. 5 drops Fig 3A). Interestingly, when drop size were 3 or 5, the long horizon conditions seemd to yield lower probability of switching than in the short horizon condition.

The overall shape of the curve was similar in humans with some notable differences. First, human subjects were at ceiling for 1 and 2 points indicating that both rewards size were equally salient (Fig 3).. Second, increasing the horizon seem to increase the probability to switch for intermediate reward size (3,4 in Fig 3).

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 for exploring the alternative option, bias is the spatial bias of choosing a particular physical side ( = 1 if the guided trial is on the left, and

Finally, unlike humans, rats did not seem to show an adaptation of behavioral strategy to the horizon context. This may reflect their tendency to use a model-free learning strategy because horizon adaptive behavior requires planning and model-based reasoning. For humans, it takes about 50 participants to show a robust horizon effect, with only a small number of rats, the lack of horizon adaptively exploration may also be due to a lack of statistical power with the limited sample size.

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

# References

Jeffrey Banks, Mark Olson, and David Porter. An experimental analysis of the bandit problem. *Economic Theory*, 1997. ISSN 09382259. doi: 10.1007/s001990050146.

Jeff A. Beeler, Nathaniel Daw, Cristianne R.M. Frazier, and Xiaoxi Zhuang. Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience*, 4(NOV):1–14, 2010. ISSN 16625153. doi: 10.3389/fnbeh.2010.00170.

Richard Bellman. The Theory of Dynamic Programming. *Bulletin of the American Mathematical Society*, 1954. ISSN 02730979. doi: 10.1090/S0002-9904-1954-09848-8.

M. S. Brainard and A. J. Doupe. What songbirds teach us about learning. *Nature*, 417(6886):351–358, May 2002.

Franc ̧ois Cinotti, Virginie Fresno, Nassim Aklil, Etienne Coutureau, Benoˆıt Girard, Alain R. Marchand, and Mehdi Khamassi. Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, 9(1):1–14, 2019. ISSN 20452322. doi: 10.1038/s41598-019-43245-z.

Michael J. Frank, Bradley B. Doll, Jen Oas-Terpstra, and Francisco Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuro- science*, 2009. ISSN 10976256. doi: 10.1038/nn.2342.

Samuel J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, apr 2018. ISSN 00100277. doi: 10.1016/j.cognition.2017.12.014.

Samuel J. Gershman. Uncertainty and exploration. *Decision*, 2019. ISSN 23259973. doi: 10.1037/dec0000101.

M. H. Kao, A. J. Doupe, and M. S. Brainard. Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature*, 433(7026):638–643, Feb 2005.

John R. Krebs, Alejandro Kacelnik, and Peter Taylor. Test of optimal sampling by foraging great tits. *Nature*, 275(5675):27–31, 1978. ISSN 00280836. doi: 10.1038/275027a0.

C. S. Laskowski, R. J. Williams, K. M. Martens, A. J. Gruber, K. G. Fisher, and D. R. Euston. The role of the medial prefrontal cortex in updating reward value and avoiding perseveration. *Be- havioural Brain Research*, 306:52–63, 2016. ISSN 18727549. doi: 10.1016/j.bbr.2016.03.007. URL http://dx.doi.org/10.1016/j.bbr.2016.03.007.

Michael D. Lee, Shunan Zhang, Miles Munro, and Mark Steyvers. Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, 2011. ISSN 13890417. doi: 10.1016/j.cogsys.2010.07.007.

Katja Mehlhorn, Ben R. Newell, Peter M. Todd, Michael D. Lee, Kate Morgan, Victoria A. Braithwaite, Daniel Hausmann, Klaus Fiedler, and Cleotilde Gonzalez. Unpacking the exploration-exploitation

tradeoff: A synthesis of human and animal literatures. *Decision*, 2015. ISSN 23259973. doi: 10.1037/dec0000033.

Robert J. Meyer and Yong Shi. Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed- Bandit Problem. *Management Science*, 1995. ISSN 0025-1909. doi: 10.1287/mnsc.41.5.817.

Nathan F. Parker, Courtney M. Cameron, Joshua P. Taliaferro, Junuk Lee, Jung Yoon Choi, Thomas J. Davidson, Nathaniel D. Daw, and Ilana B. Witten. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience*, 19(6):845–854, 2016. ISSN 15461726. doi: 10.1038/nn.4287.

́  
Elise Payzan-LeNestour and Peter Bossaerts. Do not bet on the unknown versus try to find out more:

Estimation uncertainty and ”unexpected uncertainty” both modulate exploration. *Frontiers in Neuro- science*, 2012. ISSN 16624548. doi: 10.3389/fnins.2012.00150.

Mark Steyvers, Michael D. Lee, and Eric Jan Wagenmakers. A Bayesian analysis of human decision- making on bandit problems. *Journal of Mathematical Psychology*, 2009. ISSN 00222496. doi: 10.1016/j.jmp.2008.11.002.

Jeroen P.H. Verharen, Hanneke E.M. den Ouden, Roger A.H. Adan, and Louk J.M.J. Vanderschuren. Modulation of value-based decision making behavior by subregions of the rat prefrontal cortex. *Psy- chopharmacology*, 237(5):1267–1280, 2020. ISSN 14322072. doi: 10.1007/s00213-020-05454-7.

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

Robert C Wilson, Elizabeth Bonawitz, and Vincent D Costa. Balancing exploration and exploitation with information and randomization. pages 1–18, 2020.

Shunan Zhang and Angela J. Yu. Forgetful Bayes and myopic planning: Human learning and decision- making in a bandit setting. In *Advances in Neural Information Processing Systems*, 2013.

Steverson, K, Chung, H-K, Zimmermann, J, Louie, K & Glimcher, P 2019, 'Sensitivity of reaction time to the magnitude of rewards reveals the cost-structure of time', Scientific reports, vol. 9, no. 1, pp. 20053. https://doi.org/10.1038/s41598-019-56392-0