A behavioral investigation of how rodents solve the explore-exploit tradeoff

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 5Neuroscience Graduate Interdisciplinary Program, University of Arizona

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 big 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 strategy 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 very model-free exploration strategy which works well for reversal learning, however, such design doesn’t allow for observing planning and model-based behavior in. Secondly, going away from a current bad option is confounded with exploring a novel option for information, it is impossible to dissociate directed from random exploration using conventional reversal learning. Thirdly, most of these tasks mentioned above are implemented in chamber boxes which are not natural to a rat, as Chung et al (2020) pointed out recently that head-fixed monkeys have the opposite risk preference compared to freely moving monkeys using the same task, suggesting that whether rats would behave in a more natural setting facing the explore-exploit dilemma remains unclear. Fourthly, there is a gap between the human and rodent literature, the complexity in the task and the presentation of the task is relatively different across species. It remains an open question whether similar heuristics are in human vs rodents when making explore-exploit decisions.

In the current study, we addressed these limitations by designing a novel open-field task in which rodents choose between two locations that offer fixed different amount of sugar water. To dissociate the uncertainty in estimation of value and the ambiguity of a novel option, we used magnitudes of rewards instead of probabilistic rewards, choosing the same option again in the probabilistic case can be exploratory whereas repeating the same choice is more exploitative in a pure 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 are used as a purer measure of exploration. In this way, we are also able to quantify directed exploration in a model-free way. 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 are 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. 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 did, and we compared the performance in exploration between humans and rats.

Methods Animals

4 Brown Norwegian 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 obtained from vendor XXX. All rats were housed under reverse 12:12 light cycles. All animal procedures were approved by the IRB committee at University of Arizona.

Human participants

46 participants participated in the experiment. All participants are from the undergraduate psychology subject pool who earn credits for participation in this study. The human experiment is approved by the University of Arizona Institutional Review Board.

Behavioral Task

Rat version

In this task, rats chose between two feeder locations that give out different amount of sugar water released in drops. The two feeders are associated with a fixed number of sugar water drawn uniformly from 0 to 5 in each game, and will always give the same number of drops of sugar water during a game whereas the reward magnitude (number of drops) will be reset before the start of a new game. In each game, before making their free choices, they were guided to one of the target feeders first in the first trial. From the 2nd trial, they are cued to make free choices for either 1 trial (short horizon condition) or 6 trials (long horizon condition).

There are 4 feeder locations and 2 home bases in the experiment setup (See Figure 1, A). The home base is placed between 2 lego blocks to separate them from the target feeders. Each home base is paired with the 2 feeders at the opposite side of the table. Each game will be played with one of the home bases and its associated target feeders. The other set of home base and feeders will be used for the next game. An LED light is set up at all 6 locations to cue the rats, and rats are pre-trained to follow lights. During the first trial of each game, the LED light will blink at one of the home bases, once the rat goes to the home base, the light will turn off and LED light at one of the target feeders will blink, and the rat is guided to visit that feeder, after which the cue light will come off and a certain number of sugar water drops will be released. From the second trial, after the rat is cued to go back to the home base, LED lights at both target feeders will blink simultaneously and the rat is free to go to either one. As soon as the rat makes its decision, both lights will come off and the rat will get the associated reward at the feeder visited. After the last choice is made, the rat is guided back to the home base and an 8s increasing tone is played to indicate the start of a new game, after 10s (2s without sound) the light at the opposite will blink, and the next game begins.

Each home base is associated with a fixed horizon on each single session, and the association horizon condition will be reset pseudo randomly from session to session.

Human version

In this task, participants are asked to choose between two slots machines (will also refer to as bandits) that give out a fixed number of rewards uniformly drawn from 1 to 5. Participants are instructed to maximize the total points they get. The height of the boxes indicates the number of choices to make (i.e. the horizon condition) in a game (See Figure 1 B), each row represents a trial. Before participants make their own choices, in the very first trial, they are asked to pick one of the bandits. The options available is highlighted with a green background color. For the first trial, only one of the two options will be highlighted and be available to choose. Participants indicate their choices by pressing the arrow keys on a keyboard. From the 2nd trial, both of the options will be available and participants are free to make their own choices. There are four horizon conditions (1, 2, 5, 9 free choices) and games with different horizons are interleaved.

Results

Both humans and rats are able to choose the best option (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 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).

Interesting, in the first free choice, both humans and rats can achieve an average accuracy of 70% (Figure 2). The fact that the average accuracy is significantly above chance of 50% at the first shows that there is prior information that is available and used by human and rats to guide exploration. In this particular task, with repeated training, humans and animals are able to estimate the distribution of rewards across games and use that information to guide their judgement. In particular, if they are guided to a low reward, they are more likely to explore whereas if they are guided to a high reward, they are more likely to exploit (Figure 3).

Humans showed an overall increase in the percentage of exploration in a long horizon context compared to a short horizon context. However, rats did not show any horizon difference in the strategy that they use to explore (Figure 3).

Discussion

In this study, we investigated how rodents solve the explore-exploit tradeoff in a more naturalistic setting using a novel open field task. Like in humans, we showed that rats are able to use prior information about the distribution of rewards to guide exploration. However, rats did not seem to alter their exploration strategies based on the horizon context.

In line with previous researches using the reversal learning paradigm in a chamber box setting, we showed that rats demonstrated the model-free win-stay lose-shift strategy in making explore-exploit decisions in our open field task. 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 judge 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 have to recall from sessions from previous days to estimate the distribution of possible rewards. We showed that rats are indeed able to incorporate prior information in guiding exploration.

Further analysis needs to be done to quantify directed vs random exploration in both the r at and the human dataset. In a model-free way, the percentage of choosing the unknown option can be viewed as a tendency for directed exploration. However, like 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 in the softmax action selection formula, to dissociate the tendency to choose the unknown option vs. the variability in behavior.

Finally, unlike humans, rats did not show an adaptation of behavioral strategy to the horizon context. This may reflect a tendency of using a model-free way of learning in their behavior as 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 potential better behavioral paradigm to investigate explore-exploit tradeoffs in future electrophysiology studies. 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 hard 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.

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