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

Siyu Wang[1], Blake Gerken[2], Julia R. Wieland[2], Robert C. Wilson[1,3], and Jean-Marc Fellous[1,4,5]

[1]Department of Psychology, University of Arizona

[2]Neuroscience and Cognitive Science Program, University of Arizona

[3]Cognitive Science Program, University of Arizona

[4]Program in Applied Mathematics, University of Arizona

[5]Neuroscience 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 when making decisions 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 strategy 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 strategy 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, almost existing rodent explore-exploit studies used 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 high reward magnitudes associated with low delay time vs low reward associated with high delay (Laskowski et al., 2016), 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 reversed. 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. Most of these tasks are implemented in a chamber box.

However, these reversal learning designs have several limitations. Firstly, the scope of ”exploration” being studied is limited, as win-stay lose-shift is a very model-free exploration strategy which works well for reversal learning, planning and model-based exploration can not be observed in such designs. Secondly, going away from a current bad option is confounded with exploring a novel option for information. Thirdly, there is a gap between the human and rodent literature, few papers have used comparable tasks in humans and rodents. It is important to understand whether the same exploration strategy is used between humans and rodents to understand how translatable rodent neural results are to humans.

In the current study, we partially 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. They are guided to one location first, and the extent to which they explore the unvisited location in their initial choice serves as a measure of exploration. In particular, rats performed the task in both a short and a long horizon condition. 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. Using rewards of different magnitudes instead of probability, we are taking the uncertainty in the value estimate of actions out of the equation, and as a result, exploration can be defined in a clean way that dose not depend on modeling. By looking at the extent to which rats go to the unvisited location at their initial choices, we are able to directly quantify information-driven exploration. We also ran a human version of an identical task except that humans are choosing between two slot machines that give out different amount of reward points, so we are able to compare rodent behavior directly with humans in this task. Particular, we assessed whether rats use model-based planning in exploration by evaluating whether they explore differently in a short vs long horizon context.

# 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 XXX at University of Arizona.

## Human participants

XXX participants (age XXX) 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

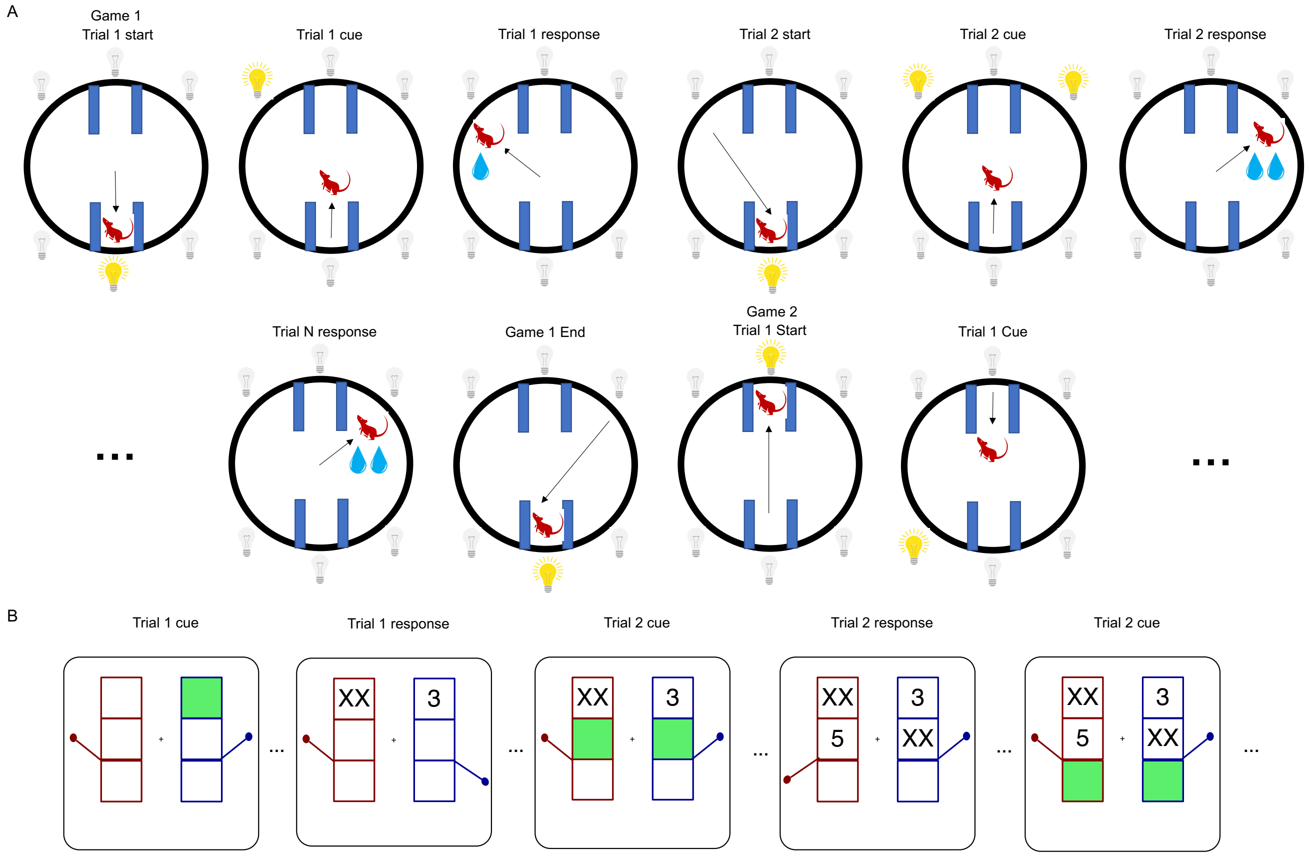

Figure 1: A. Rat version of the exploration task. B. Human version of the exploration task.

### 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 pseudorandomly 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 indicate 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

## Performance on the task

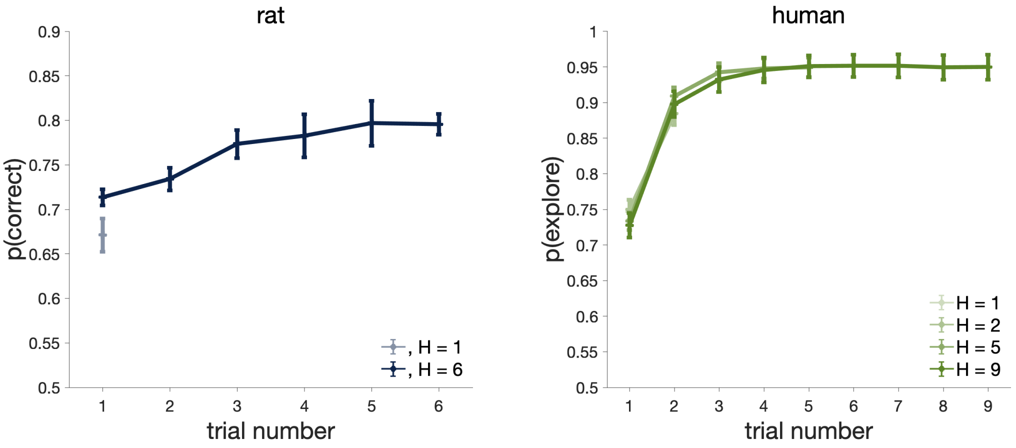

Figure 2: Performance of rats and humans in the exploration task.

## Exploration as a function of exploit value

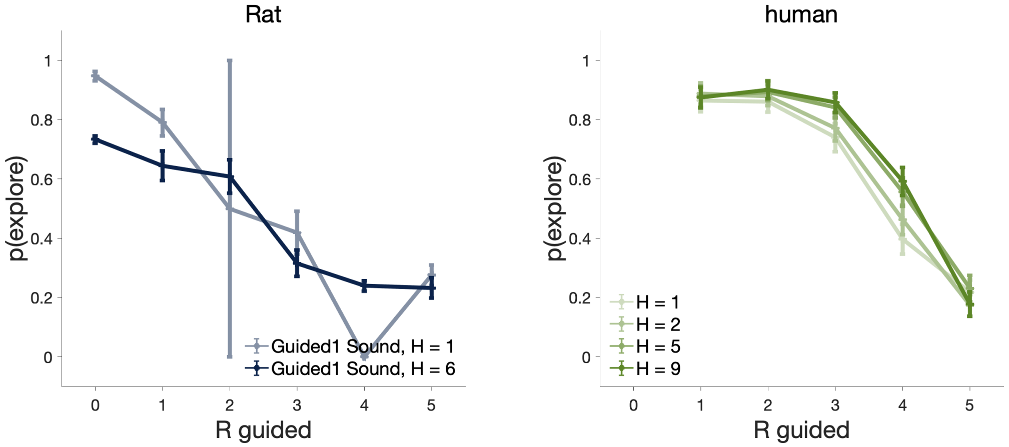

Figure 3: P(explore).

This task is very easy for humans, they can achieve an accuracy of 95%, rats can learn it as well with an accuracy of 80%. Humans achieve a 90% accuracy from the 2nd trial whereas rats learn gradually as they go in the task.

Both rats and humans chose to explore more when the exploit value is low. This is an indication of using the win-stay-lose-shift strategy. Humans increase the degree of exploration as horizon increases. However, the level of exploration doesn't seem to be modulated by horizon in rats.

# Discussion

In this study, we designed a novelty seeking task to study explore-exploit decisions in rats. In this task, rats choose between a known option with fixed amount of reward, and an unknown option with an unknown fixed amount of reward. The extent to which they explore the unknown location in their initial choice serves as a measure of exploration. In particular, rats performed the task in both a short and a long horizon condition. 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. Using rewards of different magnitudes instead of probability, we are taking the uncertainty in the value estimate of actions out of the equation, and as a result, exploration can be defined in a clean way that dose not depend on modeling. By looking at the extent to which rats go to the unvisited location at their initial choices, we are able to directly quantify information-driven exploration. We also ran a human version of an identical task except that humans are choosing between two slot machines that give out different amount of reward points, so we are able to compare rodent behavior directly with humans in this task. Humans explore more in longer horizon context, whereas rats don't adapt the level of exploration to the horizon condition.

# 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.

François Cinotti, Virginie Fresno, Nassim Aklil, Etienne Coutureau, Benoît 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 Neuroscience*, 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. *Behavioural 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.

Élise 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 Neuroscience*, 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. *Psychopharmacology*, 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.