**Face Page**

Personnel:

Co-PI: Jean-Marc Fellous (UA Psychology), Co-PI Robert Wilson (UA Psychology)  
Consultant: Dr. Siyu Wang (NIH)  
Undergraduate students: Julia Wieland (Fellous lab),

Grant Mechanism: Faculty Pilot Grant Award

Project performance sites: Fellous (rodent) and Wilson (human) laboratories

Abstract (30 lines):

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 mechanisms underlying explore-exploit decisions, a good animal behavioral model is critical. Most previous rodents explore-exploit studies used ethologically unrealistic operant boxes and reversal learning paradigms in which the decision to abandon a 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. In addition, these paradigms do not allow for observing model-based exploration behaviors, such as utilizing prior information and adaptation to the volatility of the environment. In this proposed study, we will investigate how rodents make explore-exploit decisions using a novel spatial navigation task that is designed to address the above limitations in combination with a computational model based on reinforcement learning theory. We will also record from distal hippocampal neurons known to be selective conjunctively to reward and spatial location (Xiao, Lin & Fellous 2020) to determine whether these neurons also code the value of exploration – a critical component in human exploration (Wilson et al. 2014). Importantly, we will compare the rats’ behavioral performance to that of humans in a desktop navigation task with identical parameters. Success in our project will demonstrate the feasibility of measuring and modeling explore-exploit behavior in the two species and the possibility of recording neural activity associated with exploratory decisions.

**Research Plan**

**I Specific aims**

The tradeoff between exploring unknown options and exploiting known resources is a fundamental behavioral dilemma faced by all adaptive organisms. Exploring yields information and a better understanding of our environment but comes at a cost of potentially missing out on exploiting known resources. Recent work suggests that humans solve the explore-exploit dilemma with a mixture of two strategies, directed exploration and random exploration, that appear to rely on dissociable cognitive and neural systems (reviewed in Wilson et al. 2021). However, while the large-scale neural systems underlying human explore-exploit decisions have been exposed by fMRI, relatively little is known about the small-scale neural computations underlying explore-exploit behavior. One reason for this lack of knowledge has been the difficulty in developing behavioral tasks that capture the explore-exploit dilemma in rodents. This difficulty

In this grant, we propose

Specific Aim1: Computational modeling of human and rodent behavior in the Changing Bandits Task

Specific Aim 2: Electrophysiology of

**II Research Strategy (4 pages)**

**Significance**

**Innovation**

**Approach**

**General Methods**

The Changing Bandits Task

The basic structure of the Changing Bandits Task is the same for both rats and humans. In this task, participants make a series of decisions between two options, one on the left and one on the right, that are loosely based on slot machines (or “one-armed bandits”) found in a casino. When chosen, each bandit pays out a reward between 0 and 5 drops of juice (for rodents) or 0 and 5 points (for humans). The participants’ goal (implicit for rats, explicit for humans) is to maximize their reward by choosing the option that pays out the most juice or points.

Rewards in the Changing Bandits Task are determined such that the past is usually a good predictor of the future, but not always. Most of the time the reward from each bandit stays the same from one trial to the next. However, with small probability (the “hazard rate,” or in our studies), the reward value can change such that it is randomly resampled from a uniform distribution between 0 and 5 (see Figure XXX for example reward schedule).

Crucially, the bandits change their rewards independently of each other such that when one bandit changes its reward, the other may not. In addition, the options change their reward payout regardless of whether they are chosen on a trial or not. This reward dynamic sets up an explore-exploit dilemma in the Changing Bandits Task. The longer participants exploit one bandit, the more likely the other (exploratory) bandit will have changed, and the more uncertain they will be about its value. Thus, participants must constantly assess whether it is worth continuing to exploit the known bandit or explore the lesser-known bandit to see if it is better.

While the reward structure of the task is identical for rats and humans, the specifics are different for each species. Rats will perform the task in an open field for juice rewards, while humans will perform the task in desktop virtual reality for points that will be converted to money at the end of the experiment.

In the rodent version of the task, rats initiate a trial by returning to their “home base,” a feeder that provides them with one drop of juice at the start of each trial. They then choose between the two “bandits,” which are feeders on the other side of the enclosure, one on the left and one on the right. To choose a bandit, rats must navigate to their chosen feeder, which will then deliver between 0 and 5 drops of juice. After they have completed a trial, they must then return to the home base feeder to initiate the next trial.

In the human version of the task, participants will perform the same task in a desktop virtual reality environment coded in Unity. In this task, participants initiate a trial by returning to a home base, in which they press a virtual lever to begin the trial. They then must navigate to one of two one-armed bandits, one on the left and one on the right, and pull the lever on the bandit to earn their reward. After completing a trial, they must then return to the home base to initiate the next trial.

Computational models of behavior in the Changing Bandits Task

A major advantage of the Changing Bandits Task over many other explore-exploit tasks is its amenability to computational modeling. This includes modeling both the behavior of rats and humans as well as computing the optimal strategy in the task, which allows us to predict how rat and human behavior may change in the task.

To model rat and human behavior in the Changing Bandits Task we focus on a class of models based on logistic regression. In one of the simplest versions of these models we assume that participants base their decision to explore or exploit on the last reward they saw from each option and an “information bonus” () for the explore option. In this case, the probability of exploring is given by

Where and determine the relative weighting of the explore and exploit rewards on the choice and (when negative) implements a bias towards the explore option consistent with directed exploration.

Different values of , and lead to quite different patterns of behavior as shown in Figure XXX. In this figure, we plot the probability of exploring, , as a function of the explore and exploit values for different parameter values.

|  |
| --- |
| Figure XXX – Example behaviors for different parameter values in the model. In each plot, each cell corresponds to a particular combination of and . The color and number in each cell corresponds to the probability of exploring, with brighter colors indicating a higher probability of exploring. (A) the reward weights and have opposite signs and large magnitude, leading to a purely greedy strategy in which the model always explores when the explore value is higher, exploits when the exploit value is higher, and chooses randomly when explore and exploit values are equal. (B) High magnitude reward weights and non-zero information bonus leads to pure directed exploration with a bias for exploring even when . (C) Reward weights of unequal magnitude causes the model to ignore the reward from the explore option. (D) Small reward weights and no information bonus leads to pure random exploration in which choice probabilities are in between 0 and 1 for all combinations of reward. (E) Small reward weights with non-zero information bonus leads to a mixture of directed and random exploration. (F) Small reward weights of unequal magnitude mixes directed and random exploration based only on the value of the exploit option. |

In addition to modeling rat and human behavior, we can also compute what optimal performance would look like on the task. This is helpful as it gives us a hint of the kinds of patterns of behavior we might see in our experiments. In particular, the optimal model predicts changes in the pattern of behavior as a function of both the hazard rate, , and the number of trials that the participant has been exploiting the same option, . An example when is shown in Figure XXX. When the model has only been exploiting for trial, the model is approximately greedy, choosing the option with the highest value (Figure XXXA). However, after trials of exploiting the same option, the uncertainty in the explore option is high (because it is more likely to have changed since it was last played) and consequently the behavior of the model has a large information bonus and is almost independent of (Figure XXXB).

|  |
| --- |
| Figure XXX – Probability that the optimal model explores after 1 trial of exploiting the same option (A) and 5 trials of exploiting the same option (B). Observe how the strategy evolves from an almost pure greedy strategy when (c.f. Figure XXXA) to a directed exploration strategy with almost no weighting on when (c.f. Figure XXXC). |

Of course, the optimal model has no randomness in its choices, having choice probabilities of 0 or 1. Nevertheless, this general pattern of responding, albeit with more randomness, is what we expect to see in rat and human behavior. Moreover, by quantifying how the reward weightings and information bonus change as a function of , **work in Aim 1 will allow us to determine the relative contribution of directed and random exploration to rat and human behavior**.

*Pitfalls and caveats*

It is possible that other heuristics not related to exploration may bias rat and human choices including side biases (e.g., a preference for choosing left), non-linear utilities (CITE Prospect Theory), and lapsing (i.e. choosing completely randomly between the two options regardless of their utilities, CITE). Such biases are easy to incorporate by adding additional parameters to the model. As shown in (CITE WILSON COLLINS), including these “irrational” biases in our models allows for better estimation of the effects of

## Subjects: Animals

A total of 14-16 Brown Norway rats will be used in these experiments. Six rats have already been run (behavioral experiments only, no electrophysiology). All rats will be male and female between 6 and 7 months old at the start of the experiment. We do not expect sex differences in these experiments, but data will be analyzed separately to confirm this hypothesis. All rats will be housed under reverse 12:12 light cycles. All animal procedures have been approved by the IACUC of the University of Arizona and follow NIH guidelines.

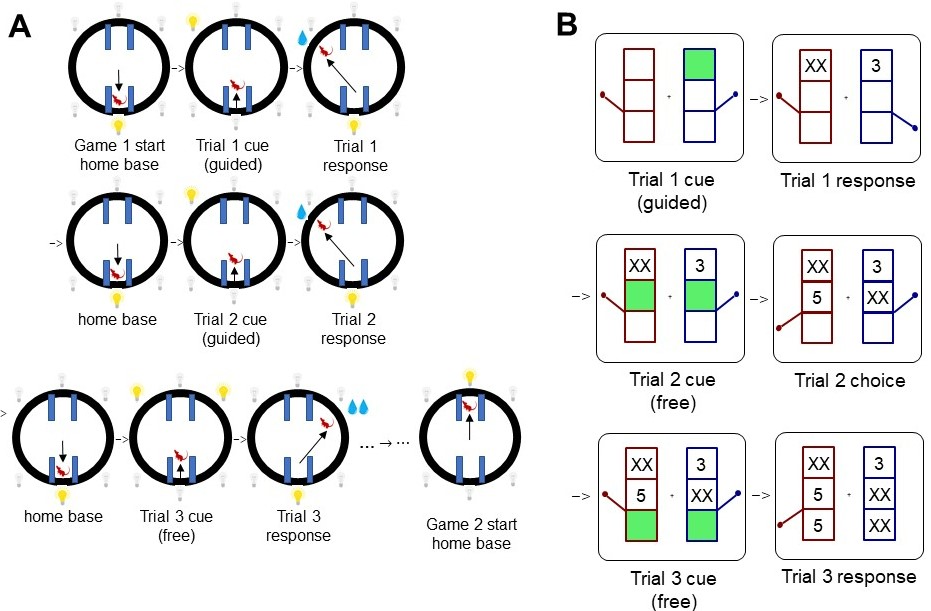


Figure 1: A: Timeline of the rat experiments. Rats were trained to start each trial by reaching the home base (no reward). They were then given a small number (here nG = 2) of guided trial (e.g. Trial 1-2, one blinking light, here 1 drop). Subsequent trials (Trial 3, free) consisted in 2 simultaneously blinking lights, The number of these trials determine the Horizon. The end of a game was signaled by a sweeping tone and a change of home base. B: Timeline of the human experiments: Human subjects were presented with a 2-armed bandit display of explicit time horizon (here Horizon = 2). They were guided to the first bandit and obtained a visible reward (here 3 points). Subsequent trials consisted in simultaneously colored squares indicating free choices between the two bandits.

## Subjects: Human participants

Male and female undergraduates subjects 18 years old or above from the University of Arizona will participate in this study. Participants who will not perform significantly above chance will be excluded. All participants will be from the undergraduate psychology subject pool and will earn academic credits for their participation in the study. The human experiments have been approved by the University of Arizona Institutional Review Board.

**Specific Aim1. Experiments - rats**

Behavioral: The rodent experiments will be run in an open field maze that consists of a circular area (1.5 m diameter) with 8 equidistant feeders at its periphery (B. Jones, Bukoski, Nadel, & Fellous, 2012; B. J. Jones, Pest, Vargas, Glisky, & Fellous, 2015). Each feeder delivers sugar water (0.2g/ml) in the form of computer-controlled drops. A blinking LED is attached to each feeder and acts as a cue when desired. The experimental sessions are divided into ‘games.’ During each game, only 3 feeders are activated (Fig 1A, yellow light bulbs). One feeder is the home base; the two others, equidistant from the home base, are the reward feeders**.** The home base was never rewarded, but animals had to reach it to trigger/activate the 2 rewarded feeders. The home base is 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 rewarded feeders will be associated with a fixed number of sugar water drops drawn uniformly from 0 to 5 and will always give the same number of drops during that game. Before making their free choices, rats will be guided to one of the rewarded feeders in the first nG trials (i.e. only one LED blinking, nG=2, ‘Trial 1 cue’ to ‘Trial 2 cue’, Fig 1A). Critically, only one of the two rewarded feeders will be cued during the guided trials, leaving the value of the other rewarded feeder unknown to the rat before making free choices. Rats will perform versions where nG = 0, 1, or 3 (In cases of nG = 0, rats will not be guided to any target feeder and will start with a free choice between the 2 rewarded feeders instead.). Fig 1A illustrates the version with nG = 2. From the nG+1st trial, they will be cued to make free choices (the LED of the 2 rewarded feeders blinked simultaneously, ‘Trial 3 cue’ Fig 1A). The guided trials will be followed by H free choices between the 2 rewarded feeders. Rats will perform versions with H = 1, 6, or 15. After the first game is completed, an 8s increasing sweep tone will be played to indicate the start of a new game. The layout will then switch, and the feeder directly opposite to the initial home base will be activated as the new home base and will signal the start of a new game (Game 2 start, Fig 1A). The new rewarded feeders are the feeders opposite to the new home base. The number of free choices H is also referred to as the ‘horizon’.

Electrophysiology: Male and female Brown Norway (6-8 months old) rats will be food restricted and pre-trained as previously done in the Fellous laboratory (refs). Animals will be implanted with a 18-tetrodes device targeted at the dorsal CA1 area of the hippocampus (3.3-3.8 mm posterior to bregma, 2.0 mm lateral to midline). This targeting will be made possible by 3D printed implant exit tips produced in-house. Data will be collected with a second-generation Digital FreeLynx wireless system connected to a Halo-18 tetrode drive. Both spike and LFP data (e.g. theta, sharpwaves ripples) will be collected. Single neurons will be isolated offline and the quality of the isolation will be assessed (refs). The data will be analyzed with custom written Matlab code (refs). Animals will be tracked using an overhead video camera at 25-30 Hz. Place fields are computed making sure to exclude sharpwave-elicited activity (refs). The position of the tetrode tips will be determined in all animals by electrolytic lesions and standard Nissl staining.

**Experiments 1.1 - How do rats process the horizon information after guided trials?**

Our preliminary data show that rats explore (direct exploration) less in long horizons than in shorter ones, but it is unclear as to why. One possibility is that, in long horizons, rats know that they have more chances to visit the alternative feeder, and as a result, they reduce their rate of switching on the first unguided trial. Another possibility is that rats treat the horizon condition as a cue to the stability of their environment; in long horizon, the environment is more stable and learnable, so they spent time exploring it, but short horizons are too volatile and they choose more randomly and as a result visit the unguided feeder more.

1.1A: To test the first possibility, we will encourage rats to explore the unguided feeder by running a pre-condition such that if the rat does not pick the unguided feeder, that feeder will no longer appear as a choice in the future, decreasing the overall outcome of the current game. Normal games will be run immediately after this pre-condition. The results will be compared to games that were run without this precondition. We expect that rats will explore short and long horizon in the same fashion if the pre-conditioning is run.

1.1B: To test the second possibility, we will run a condition in which the long horizons are made unstable by dynamically changing the size or probability of the rewards during each game.

1.1C:  Electrophysiology: 4-6 rats will be used for electrophysiological experiments. To increase our yield, and simplify the behavioral conditions, the rat will either be guided to 1 drop or 5 drops in separate games, but during the same day (same cells). In the case of 1 drop, we will be testing whether rats are willing to visit the unguided feeder to figure out if it is a 3 or 0 drop. In the case of 5 drops, we expect the animals not to explore the other feeder. We will use the interleave design of our preliminary experiments and use a sound cue. We will measure the activity of conjunctive reward-location CA1 cells at or near the 3 feeders, and predict that 1) the firing rate of the cells near the 1 drop guided feeder will be significantly different from those of the 5 drop feeders, 2) the firing rate or co-activity levels (e.g. cross-correlation, in case of recordings yielding multiple conjunctive cells) at the home base will be predictive of the next feeder choice. We will compare the activity of these conjunctive cells across two horizon conditions (short: 3 and long:10).

**Experiments 1.2 - guided vs free choice**

Our preliminary results show that in the case where a rat is guided once vs choose freely once on the first trial, rats will then visit the same feeder more if they acted freely on the 1st choice compared to when they were guided. Free choice on the 1st trial yields less exploration in subsequent trials than if the first trial was guided. This intriguing result was unexpected and requires further investigation.

1.2A: We will record the activity of hippocampal cells in this paradigm. Will conjunctive reward-location cells at or near the home base, or at the two target feeder locations encode guided information differently from information gained from free choices? In these experiments, we will also use a simplified version of the task in which the rat will always make 10 visits to one pair of feeders, 5 of them will be guided, 5 of them will be free choices (make the 1st choice 50% guided as well). This will maximize the statistical power to study the contrast between guided vs free choices.

Caveats and alternative plans for the rodent experiments:

**Specific Aim 2: Human experiments**

General methods:

In this experiment, participants will be sitting in a booth, in front of a computer screen. They will be asked to choose between two slots machines (also referred to as bandits, Fig 1B) that give out a fixed number of reward points uniformly drawn from 1 to 5. Participants will be instructed to maximize the total number of points. The height of the boxes indicate the number of choices allowed in the current game (i.e. the horizon condition, H=2 in Fig 1B) and each row represents a trial. Before participants make their own choices, in the very first trial, they will be guided to pick one of the bandits (Trial 1 cue, nG=1, Fig 1B). The option available will be cued with a green background color. Participants indicate their choices by pressing an arrow key on the keyboard. Their response will be followed by an indication of how many rewards they obtained, the reward of the unchosen option will not be shown and will showed up as ‘XX’ (Trial 1 response, Fig 1B). From the 2nd trial, both bandits will be available and participants will be free to make their own choices. There will be four horizon conditions (H=1, 2, 5, 10 free choices), and games with different horizons will be pseudo-randomly interleaved. Fourty human participants would complete a total of 6080 games (33440 trials).

For human experiments, the purpose should be to match the rats (in order to test the difference), right? We should get rid of the visual history and horizon cue for humans. I think we should even use dots instead of points for humans, to make it a perceptual task. (Rats can revisit a feeder to confirm the perception that the reward was 2 drops, if the reward is based on perceptual cue as opposed to numbers, human will need to revisit to confirm the perceptual judgement as well).

So the proper comparison that I am thinking is like this:

Humans can choose left or right, say choose left, then a empty box will show up, there will be random green dots in the box, the density of green dots show reward magnitude (however humans have to judge by eye), we can even present this in 200ms so humans really need to revisit to be sure of a judgement. then the box will disappear, a new trial begins. This way, we should have a pretty fair comparison between humans and rats. Horizon cues can use sound cues as well.

Another experiment I think is interesting to try is to run humans without instruction, rats never got verbal instruction, and only learned through trial and error.

Caveats and alternative plans for the human experiments:

**III Plan for Submission of External Grant Application**

We plan to submit an RO1 application to NIMH or NINDS. By that time, we expect the behavioral data obtained so far to be published. The application will include preliminary results from the electrophysiological experiments collected as part of this pilot grant (the collection of these data cannot be achieved otherwise in the Fellous lab as of summer 2022)

**IV Timeline**

Months 1-12: behavioral data collection (10 rats). Electrophysiology data collection (4 rats).  
Months 5 -16: rodent data analyses, figure production for inclusion in grant application.  
Months 17-18: submission of grant application (estimated summer 2024)