Temporal Persistence Explains Mice Exploration in a Labyrinth

Umesh K Singla (usingla@princeton.edu)
Princeton Neuroscience Institute, Princeton University
Department of Computer Science, UC San Diego

Marcelo G Mattar (marcelo.mattar@nyu.edu)
Department of Psychology and Center for Neural Science, New York University
Department of Cognitive Science, UC San Diego

Abstract

Exploration in sequential decision problems is a computationally challenging problem. Yet, animals exhibit effective exploration strategies, discovering shortcuts and efficient routes toward rewarding sites. Characterizing this efficiency in animal exploration is an important goal in many areas of research, from ecology to psychology and neuroscience to machine learning. In this study, we aim to understand the exploration behavior of animals freely navigating a complex maze with many decision points. We propose an algorithm based on a few simple principles of animal movement from foraging studies in ecology and formalized using reinforcement learning. Our approach not only captures the search efficiency and turning biases of real animals but also uncovers longer spatial and temporal dependencies in the decisions of animals during their exploration of the maze. Through this work, we aspire to unveil a novel approach in cognitive science of drawing interdisciplinary inspiration to advancing the field's understanding of complex decision-making.

Keywords: Exploration; Reinforcement Learning; Temporal Abstraction; Animal Behavior; Ecology; Foraging

Introduction

Exploration plays a fundamental role in animal survival. Understanding the exploratory and search behavior of humans and animals is a key focus in diverse scientific fields. The dynamics of exploration in neuroscience have been studied across mostly shorter temporal scales: from characterizing choice behavior in two-alternative forced choice tasks (Costa, Mitz, & Averbeck, 2019), to studying head turns on encountering a novel object (Gharbawie, Whishaw, & Whishaw, 2004; Gordon, Fonio, & Ahissar, 2014), to studying kinematics of exploration in closed circular arenas (Tchernichovski, Benjamini, & Golani, 1998). Yet, relatively few studies have tried to model animals' exploratory behavior in larger or more complex environments, such as the ones actually faced by animals in the real world. This study aims to fill that gap.

The natural world is full of complex environments that require animals to navigate through intricate paths and make decisions based on their surroundings. As such, exploration and search strategies ought to be much more complicated in real settings. Neuroscience and psychology experiments often fall short of replicating that complexity and involve substantial human interference, limiting what we can learn about true animal behavior. This is in contrast to the field of spatial ecology, which has focused extensively on studying animal movement in naturalistic settings, from prey hunting in plain fields to bird migrations across oceans (Viswanathan, Da Luz, Raposo, & Stanley, 2011). For instance, Lévy walks in ecology are known to capture animal movement at larger spatial scales or longer temporal scales (Viswanathan et al., 1999). During Lévy walks, animals persist in a certain direction, which helps them expand their search territory faster and avoid getting stuck in a local region. However, there is a gap in the exchange of ideas between ecology and neuroscience (Berman, Kardan, Kotabe, Nusbaum, & London, 2019), partly because of the differences in the scale of investigation of the two fields.

Recent advances in machine learning have enabled animal tracking with high precision, enabling an increase in the use of complex environments that contain many choice points to study animal behavior (Vallianatou, Alonso, Aleman, Genzel, & Stella, 2021; Nagy et al., 2020). One such experiment recently conducted by Rosenberg, Zhang, Perona, and Meister (2021) involves ten mice, each exploring a complex labyrinth for close to seven hours without any human interference whatsoever. Animals had access to sufficient food and water in the home cage connected to the maze. Curiously, even though the maze offered no explicit reward, animals continued to enter and exit the maze throughout the night to explore (Figure 1). While this behavior supports the role of intrinsic motivation in driving animals to explore, the structure and remarkable efficiency exhibited in their exploration strategies constitute a perfect example of the complex and naturalistic behavior that remains poorly understood in the behavioral sciences. In their original paper, Rosenberg et al. (2021) characterized the animals' exploratory behavior using a computational model composed of four parameters that governed the probabilities of actions at each junction. However, this model was tailored to the specific dataset and maze layout. As such, it remains unclear if there are general computational principles capable of explaining the efficiency of animal exploration in this and other complex environments. Such principles should, ideally, also relate to known tenets of animal movement in spatial ecology.

In this study, we propose a candidate principle meeting these desiderata. Using the maze exploration data from Rosenberg et al. (2021) as a case study, we built an exploration agent based on a few simple principles of animal movement from foraging studies and formalized using the framework of reinforcement learning (RL). Our main hypothesis is that, during exploration, animals rely on temporal abstraction

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

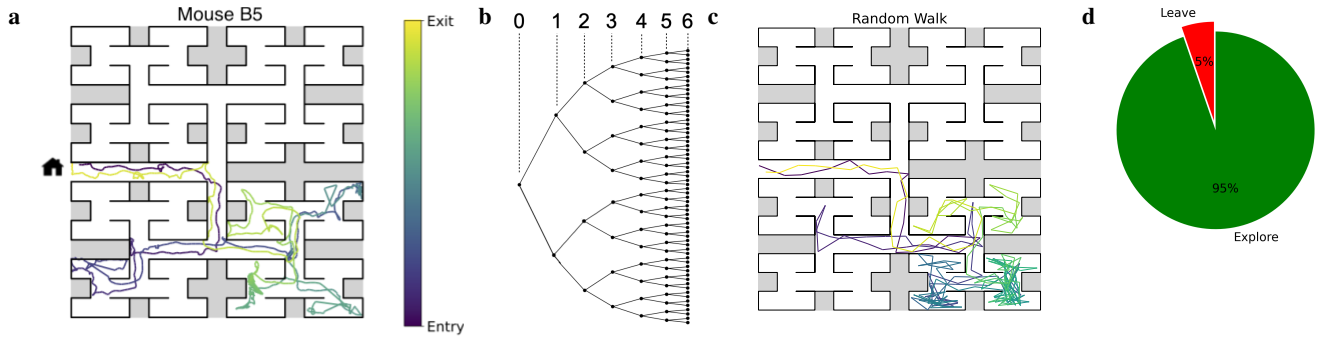

Figure 1: **The exploratory behavior of animals exhibits rich structure.** (a) The maze environment and a sample trajectory of an animal. The maze is connected to a homecage with food. Mice could move freely between the homecage and the maze. The colorbar denotes time elapsed since entering the maze. (b) The maze is structured as a complete binary tree with 63 T-junctions at levels 0-5 and 64 leaf nodes at level 6. (c) A random walk agent simulated for the same number of time steps as the mouse in (a). Random walk agent often gets stuck in a small region of the maze. (d) Pie chart shows animals spend on average over 95% of the experiment time exploring the maze. 'Leave' indicates the portion when animals are directed towards the home cage. Figures for panels (a), (b) and (d) adapted from Rosenberg et al. (2021).

to circumvent the complexity of sequential decision-making, giving rise to stereotyped action sequences. Computationally, we express this hypothesis in terms of a temporally-extended ε-greedy algorithm, recently proposed as a general exploration framework in RL by Dabney, Ostrovski, and Barreto (2020). Temporally-extended ε-greedy uses temporal abstraction to yield efficient exploration in a range of RL settings. However, while Dabney et al. (2020) only compared this algorithm against perfect memory agents or neural networks, here we test its ability to explain naturalistic animal behavior.

Our results show that a temporally persistent ε-greedy agent captures the exploration efficiency and the turning biases observed in the mouse navigation. Despite its simplicity, it outperforms several other exploration algorithms, including the more complex and less parsimonious model proposed in the original paper (Rosenberg et al., 2021). These results suggest that once the animals have chosen a sequence of actions to travel in a certain direction, they do not make decisions at intermediate turns. As such, a behavioral action policy that specifies longer spatial and temporal dependencies will serve more suitably in capturing animal actions as opposed to one made of fixed local rules. Further, we also show that the mice exhibit superdiffusive movement, similar to doing Lévy walks, within the maze and are optimizing for search efficiency. Our work makes a novel contribution to the field of cognitive science by providing a parsimonious characterization of the exploratory behavior of animals in a complex maze.

Approach

Mouse maze dataset

This study consists of a re-analysis of an existing dataset, which we briefly describe here (Rosenberg et al., 2021). In this study, mice freely navigate in a maze. The maze is an enclosed complex labyrinth connected to a home cage via a short tunnel (Figure 1a). The logical structure of the maze is a binary tree, with 6 levels of branches, from home to 64 leaf nodes. The levels are numbered 0, 1, …, 6 where level 0 is the central point of the maze and at level 6 are the leaf or end nodes (Figure 1b). The animal is initially kept in the home cage and, at time 0, the tunnel opens and the animal is free to travel between the cage and the maze. The maze is constructed with maximal symmetry around the center point. There are a total of 10 mice subjects that performed the experiment, and for each subject, six keypoints (nose, tail, and 4 limbs) were recorded continuously. We use the nose keypoint trajectories as the behavior data in our analysis. One out of the 10 subjects did not travel beyond first few cells; this animal was excluded from all subsequent analysis. While there is a second set of experiments done on an another group of 10 mice that are rewarded with water inside the maze, we chose to only focus on the unrewarded animals as our primary aim in this study was to understand the animals' exploration strategies in the absence of external motivators. See the original study for more details on maze construction.

Models of exploration in mouse maze

We formalize the problem of exploration in the current maze as a Markov Decision Process (MDP). The set of states constitutes all the 63 nodes at T-junctions, 64 end nodes and the home node. At each of the 63 T-junctions in the maze, there are 3 actions available to go-left (L), go-right (R) or go-back (B). On the binary tree, the action L and R take the agent one level deeper into the maze where the action B takes the agent up to the parent level. At each of the 64 end nodes, the only action is to go back and at Home, the only action is to go to node 0. The transition probability matrix is entirely deterministic and assumed to be known. There is 0 reward throughout the maze.

In this task, animals may follow a variety of exploration

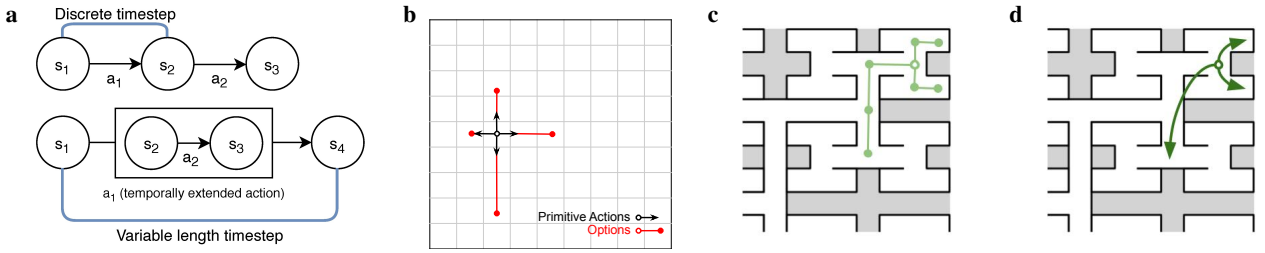

Figure 2: **Algorithmic implementation of εz-greedy in the maze task.** (a) The semi-MDP framework (bottom) allows the length between timesteps to be variable in comparison to standard MDP (top) and as such can support temporally extended actions. (b) The set of primitive actions (black) and few possible temporally-extended actions, or options, of variable length timesteps (red) shown for an open grid world environment. Open circle denotes the state where an option is initiated and the solid circle denotes its termination. (c) In the given maze, if modeled as MDP, the agent will only make progress by one node before having to make a decision again at every intermediate turn in the maze. (d) In εz-greedy model, which is based on the semi-MDP framework, the agent first chooses a direction at random and samples an entire action-sequence and executes it in one go, abstracting away the decisions at intermediate timesteps.

strategies, each associated with a distinct computational cost and efficiency level. The efficiency of an agent is heavily dependent on its ability to maintain memories of past explorations. On one hand, an agent can have perfect memory which can enable it to store a complete mental map of the world and perform an efficient systematic search (e.g., breadth-first search, depth-first search, etc.). In the other extreme case, an agent that has zero memory should exhibit completely random behavior. In between these two extremes, lie a wide range of exploration strategies with limited memory and computational demands. Accordingly, we next describe a few families of algorithms that we explored as plausible descriptions of the observed exploratory behavior by the mice in the maze.

Temporally-persistent ε-greedy exploration A common strategy used in RL to promote exploration in sequential environments is ε-greedy. However, in a reward-sparse or reward-free environment, relying on an ε-greedy strategy can be very inefficient (Dabney et al., 2020). In ε-greedy, the probability of being able to move away from one part of the environment to another reduces exponentially with the number of steps required. To tackle this, Dabney et al. (2020) recently proposed a temporally-extended version of ε-greedy as a general exploration framework. Rather than sampling an action at every time step, this algorithm instead samples a sequence of actions and executes this "composite" action (Fig. 2a-2d). These composite actions, also known as options in the semi-Markov Decision Process (semi-MDP) literature, abstract away the intermediate steps and allow flexible behavioral policies (Sutton, Precup, & Singh, 1999). The temporally-extended ε-greedy strategy requires choosing an exploration probability ε and an appropriate set of options O. Then, as with vanilla ε-greedy, it samples an option w with probability ε or follows the current policy with probability $1-\varepsilon$. For purely exploratory settings, we set $\varepsilon = 1$, so that the agent always samples an option, eliminating the need to specify a baseline policy and a learning algorithm. We make this assumption not only for parsimony but also because there can be no reward-based learning of a policy in a reward-free environment.

We adopt the spatial version of temporally-extended ε-greedy for our problem, called εz-greedy (Dabney et al., 2020). εz-greedy constructs an option w_{an} that takes the same action a for n time steps and terminates. The complete set of options O is made up of all such "action-repeats", for all combinations of valid actions and durations where the duration n is sampled from some distribution z. These "action-repeats" allow an agent to persist in one direction and not get stuck in a local region, in contrast to a vanilla ε-greedy agent. To test εz-greedy on our data, we construct an appropriate set of options that encode a similar sense of directional persistence in our maze environment. For the duration distribution, we use the heavy-tailed Lévy distribution $z(n) \sim n^{-\mu}$ with $\mu = 2$. Being heavy-tailed, it samples a lot of short steps and spends time in one region but also has a non-zero probability to switch to a different region when a large step size is sampled. Such heavy-tailed distributions have been observed in many animal foraging studies in ecology (Viswanathan et al., 1999, 2011).

Biased Walk Rosenberg et al. (2021) proposed a Biased Walk model in their paper to capture the maze exploration that uses four parameters estimated from the animal data. These 4 parameters define the probability of an action at a junction depending on the previous action taken. In a way, Biased Walk is also introducing correlations between consecutive steps of animals. However, it is important to note that these parameters are specific to the given data and the environment, and as such, their generalizability may be limited. Refer to Rosenberg et al. (2021) for more details.

Uncertainty-based exploration In RL, Bayesian Q-learning extends the traditional Q-learning algorithm to model uncertainty in the estimated values of state-action pairs

(Dearden, Friedman, & Russell, 1998). Instead of having a fixed Q-value for each state-action pair, Bayesian Q-learning maintains a probability distribution over possible Q-values. This distribution captures the uncertainty associated with the estimated value of each action in a given state which is then used to guide exploration. Actions with higher uncertainty are then greedily chosen during the exploration. We implement an agent that uses the standard deviation of a normal distribution around a fixed value of 0 as uncertainty measure for each Q-value. As the agent gains more experience, the uncertainty in the Q-values decreases and this normal distribution gets narrower over time.

Results

We simulate the εz-greedy exploration agent as well as a random walk, an optimal walker based on Depth-First Search, Biased Walk and a Bayesian uncertainty agent. We simulate each agent for equal number of time steps ($T = 25000$) and assess the performance of these agents by employing diverse metrics derived from many aspects of the behavior data. The Biased Walk simulations are specific to each animal and are simulated using the parameters computed from its trajectory data (Rosenberg et al., 2021). The εz-greedy uses no parameters apart from the Lévy distribution for sampling durations and is purely generative.

Efficient movement in the maze with εz-greedy model of exploration. We sought to first compare how close the εz-greedy agent is to the animals at covering the spatial extent of the maze. We use the definition of exploration efficiency from the original study as the total number of nodes visited N_{half} required to survey half the end nodes, and define $E = 32/N_{half}$. The optimal agent with perfect memory visits the end nodes systematically without any repeats, resulting in an efficiency of $E = 1.0$. A random agent with no memory repeats a node before having visited all of them results in an efficiency of $E = 0.23$ when simulated. The exploration efficiencies observed for animals lie in the middle of the two, with an average of $E = 0.39 \pm 0.03$. The εz-greedy model gives an efficiency of $E = 0.35$ and accounts for 91% of the variance observed in the animals' efficiencies. The Biased Walk shows an average efficiency of $E = 0.33 \pm 0.03$ and accounts for 85% of the variance observed. Figure 4a plots for two animals the number of distinct end nodes found as a function of the total number of end nodes visited, as a measure of exploration efficiency over multiple window sizes. A sample trajectory simulated for each agent is shown in Figure 3. Even though the Biased Walk closely approximates the exploration efficiency values of animals achieved by the z-greedy model, the two models demonstrate distinct movement behaviors in the maze (Figure 3). Notably, the εz-greedy agent seeks to cover a much larger extent of the space within the same timeframe compared to the Biased Walk. The uncertainty-based agent exhibits a very systematic behavior and performs very efficiently; however, at the expense of requiring intensive computations and keeping track of probabilities.

εz-greedy recovers the turning biases of animals. We next asked if the εz-greedy agent follows similar rules of exploration as the animals do. Rosenberg et al. (2021) found strong biases by animals at decision points that were remarkably consistent across animals. Animals exhibited a strong preference to go forward at T-junctions (P_{SF}, P_{BF}) and alternate at turns left and right (P_{SA}), as well as a mild preference to turn into stem (P_{BS}). Their Biased Walk model is based on this set of 4 biases to govern decisions at each turn during exploration. We calculate the same set of probabilities using the simulated trajectory data of the εz-greedy agent. The εz-greedy model recovers all the four turning biases within $\sim 90\%$ of animals' values (Figure 4b). εz-greedy does show a higher forward bias (P_{SF}) indicating a slightly higher directional persistence in our model than animals. Rosenberg et al. (2021) had speculated on the presence of these consistent biases in animals in their paper, questioning if such rules are genetic. However, we show that just adhering to the general principle of directional persistence in an environment is sufficient to replicate these biases.

εz-greedy exhibits outwards tendency similar to animals during exploration. Rosenberg et al. (2021) made a keen observation in their study that animals showed a strong preference for certain end nodes during their exploration of the maze. To quantify this, they measured the fraction of visits to the nodes at the periphery of the maze (outer nodes) compared to the visits to the innermost end nodes of the maze (Rosenberg et al., 2021). The outer nodes were favored by a factor of 2.3 ± 0.55 to the innermost 16 end nodes for various animals. The εz-greedy agent by design prefers traveling long distances in a certain direction. The outer nodes in the maze are reachable by alternating straight paths and the forward options favor those alternating paths as the agent enters a region, therefore, εz-greedy exhibits a similar outgoing tendency as animals. The εz-greedy agent shows an outward preference of 2.28, which is in range of the most animals whereas Biased Walk shows an outward preference of 1.98 ± 0.19 (Figure 4c). The random walk, optimal or uncertainty agents do not favor any nodes to visit and hence exhibit no such preference.

εz-greedy reduces the average uncertainty in decisions at a junction. To assess the quality of the various models objectively, Rosenberg et al. (2021) attempted to separate out the predictable component from the intrinsic randomness in an animal's decision. They defined cross-entropy between a model's predictions and the animal's observed actions at each of the 63 T-junctions as a measure of remaining uncertainty about an animal's decision. Rosenberg et al. (2021) developed a Markov-chain model that used the current node and preceding k nodes to predict the animal's next action at a junction. Using cross-validation, the authors found the best k to be $\sim$3 for most animals and the minimum cross-entropies ranged from $1.23 - 1.37$ bits/action on the test data. Note that the Markov chain model has a parameter associated with each k-length sequence of nodes leading to a total of $63 \cdot 3^k$ param-

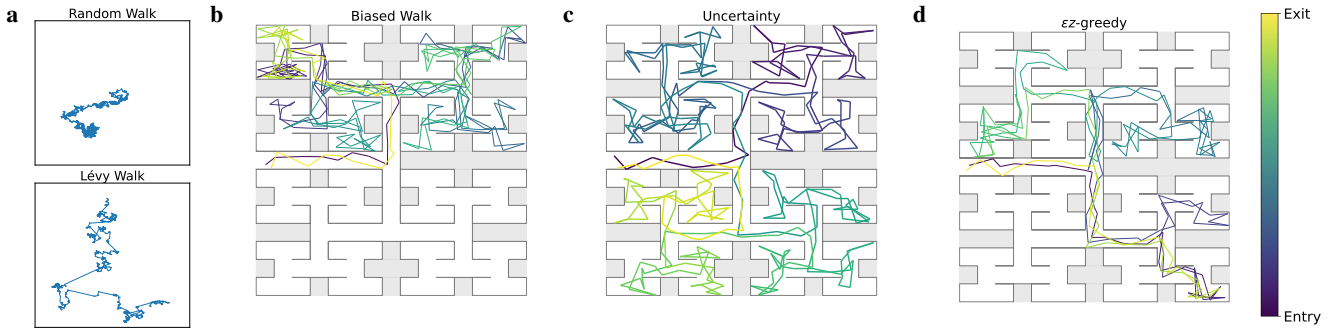

Figure 3: **Simulations.** (a). A random walk (top) and Lévy walk (bottom) simulated in an open field for 1000 steps. A sample episode for (b) Biased Walk, (c) Uncertainty-based agent, and (d) εz-greedy agent simulated for 100 steps in the maze.

eters. Given the exhaustive nature of this model and limited data availability, we treat this cross-entropy as a soft upper bound on the true source entropy of the animal behavior. A dependence on history size of $k = 3$ indicates that the animal's choice behavior is influenced by its current location and $\sim$3 locations preceding it. This is in line with our hypothesis of the presence of long-range correlations in the movement of animals in the maze.

We were interested in seeing how close the εz-greedy gets to the above sophisticated Markov-chain model in reducing the intrinsic uncertainty. We computed the cross-entropies on the simulated data of εz-greedy for various history depths. The εz-greedy produces the minimum cross-entropy of 1.35 bits at a history depth of preceding 3 nodes (Figure 4d). Biased Walk, simulated using each animal's bias parameters, produces the cross-entropy of 1.47 ± 0.02 bits/action and shows a dependence on only 2 preceding nodes. In hindsight, this is expected as the Biased Walk specifies correlations between any two consecutive steps, whereas an εz-greedy agent has the capacity to introduce correlations at both short and long time scales in action sequences. Down from an uncertainty of $\log_2 3 = 1.59$ bits for a random agent to 1.35 for εz-greedy, this indicates a reduction of over 15% in the average uncertainty over decisions by incorporating temporal persistence in actions.

Discussion

Operating in a sequential decision-making environment requires cognitive skills such as planning and remembering across spatial and temporal scales that are not required in most binary decision tasks (Mattar & Lengyel, 2022). Many neuroscience studies have also hinted at the presence of compressed representations of motor action sequences such as skills and habits in the striatum (Graybiel & Grafton, 2015). This is especially true when cognitive resources are scarce and there is a structure in the environment, then such compressed representations help free up working memory by efficiently encoding a chunk of the sequence together, memorizing it and executing it (Lai, Huang, & Gershman, 2022). In our work, we found that a temporally-persistent ε-greedy agent, where an agent samples and executes a whole action sequence, captures the exploration behavior of animals in the maze very well. This is the first study that documents the role of spatially-compressed (or temporally-abstracted) actions in the navigation behavior of real animals in a complex environment giving rise to the biases and efficiency in exploration we observe.

Further, our model's main strength lies in its interpretability over other models of maze behavior. The exploration movement patterns of humans and animals in open environments are known to be superdiffusive in nature and resemble Lévy walks (Viswanathan et al., 2011). Lévy walks exhibit an intermittent behavior where it spends some time searching in one region before relocating to another (Lomholt, Tal, Metzler, & Joseph, 2008). The success of εz-greedy model implies that mice exhibit superdiffusive movement within the maze and are optimizing for search efficiency. This further suggests that 1) the animals alternate between relocating and search phases in their exploration in the maze (Figure 4e), and 2) once the animals have chosen to go in a certain direction, they do not make decisions at intermediate turns but continue to persist in that direction. By segregating the learning process and the innate mechanical aspects of a behavior, models like εz-greedy serve additional purpose by aiding in the selection of the appropriate formulation of the action space and the behavior policies. We now know the mice take short and long-range actions, so an RL policy that considers a spatiotemporally flexible action space is going to be more effective than the one trying to learn only the action one step ahead. To our knowledge, majority of existing research on animal behavior has focused on exploration in open fields; therefore, our findings make a unique and valuable contribution to the field. The Biased Walk model, by introducing correlations at a single albeit dominant temporal scale, proves effective in approximating certain aspects of mice behavior but falls short in providing any general significance.

The effectiveness of εz-greedy model in explaining maze behavior highlights further aspects of mice's cognitive abilities. Being able to execute a temporally-extended option in this maze indicates that mice can sample and execute a "jump" in arbitrary directions, even when those directions appear to be obstructed by the presence of maze walls. This

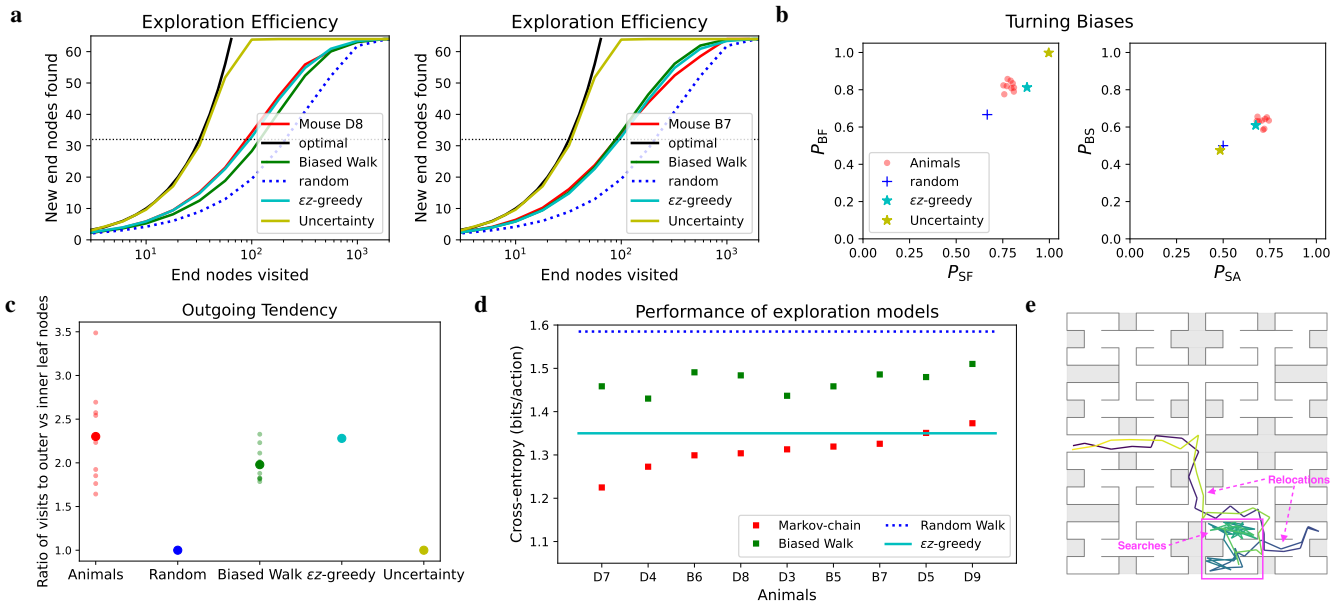

Figure 4: **A temporally-persistent ε-greedy agent captures the behavior well.** (a) Exploration efficiency plotted for: mouse D8 (red); εz-greedy agent (cyan), Biased Walk (green), Uncertainty-based agent (yellow); optimal agent (black); a random walk (blue). (Right) Same for another mouse B7. (b) Turning biases of animals compared with biases obtained from simulated trajectory data for εz-greedy, uncertainty agent and random walker. (c) Outer node preferences of individual animals and their corresponding Biased Walks, with mean values highlighted, and compared with εz-greedy agent, uncertainty agent and random walk. (d) Cross-entropy of the Markov-chain model with depth 3 for each animal compared with corresponding Biased Walks when predicting the decisions of the animal at T-junctions. Solid cyan line represents the cross-entropy for an εz-greedy agent. Dotted line represents the random walk agent with 1/3 probability of each action. (e) The interpretation of animals behavior in the maze as intermittent searches (alternating between relocating and searching in the corners).

could suggest that mice are not constrained to choosing where to go solely from what is currently within their field of vision; instead, they are able to choose any arbitrary direction within the maze and follow a direct path towards the chosen direction. This hints at mice's capability of intentional and non-visually guided movement in the maze, implying a degree of flexibility in their spatial decision-making. This is in contrast to the Biased Walk, which specifies the probabilities of only the adjacent nodes, impeding any potential for encoding long-term intention or multi-step planning. The ability to choose arbitrary directions might suggest a capacity for both allocentric and egocentric spatial awareness (Rinaldi et al., 2020; Vijayabaskaran & Cheng, 2022).

We also want to highlight that many ecological studies have found multiple search models to be effective at explaining the same animal movement data. Distinguishing among them is challenging as the discretization granularity of the animal trajectories can impact our conclusion of the underlying process (Palyulin, Chechkin, & Metzler, 2014; Bartumeus, da Luz, Viswanathan, & Catalan, 2005). It is therefore important to not fixate on the precise nature of options or the exact duration distribution used; but instead to focus on the relevance of the general idea of scale-free temporal persistence when modeling animal behavior in larger environments.

A limitation of our study is that we do not consider the learning aspects during exploration. Additional efforts could be directed towards characterizing their learning, internal states or intrinsic motivation of animals. Animal behavior is complex. Through our study, we emphasize it is important to acknowledge and account for this complexity if we were to understand the relationship between behavior and neural activity. Our results highlight that richer accounts are necessary even to explain apparently simple behaviors. Conducting additional studies would significantly enhance our ability to draw insights into animal exploration and planning strategies.

Acknowledgments

We are grateful to Rosenberg et al. (2021) for sharing their data in public. We thank Matthieu Le Cauchois and Kristopher Jensen for many discussions on the analysis and modeling of the behavior as well as feedback on the manuscript. We thank Homero Esmeraldo and Kokila Perera for many discussions on animal behavior in the beginning that led to inception of this work. We are grateful to the Departments of Cognitive Science and Computer Science at UC San Diego for providing resources and supporting interdisciplinary research.

References

Bartumeus, F., da Luz, M. G. E., Viswanathan, G. M., & Catalan, J. (2005). Animal search strategies: a quantitative

random-walk analysis. *Ecology*, *86*(11), 3078–3087.

Berman, M. G., Kardan, O., Kotabe, H. P., Nusbaum, H. C., & London, S. E. (2019). The promise of environmental neuroscience. *Nature human behaviour*, *3*(5), 414–417.

Costa, V. D., Mitz, A. R., & Averbeck, B. B. (2019). Subcortical substrates of explore-exploit decisions in primates. *Neuron*, *103*(3), 533–545.

Dabney, W., Ostrovski, G., & Barreto, A. (2020, June). Temporally-Extended {\epsilon}-Greedy Exploration. *arXiv:2006.01782 [cs, stat]*. (arXiv: 2006.01782)

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-Learning. , 8.

Gharbawie, O. A., Whishaw, P. A., & Whishaw, I. Q. (2004). The topography of three-dimensional exploration: a new quantification of vertical and horizontal exploration, postural support, and exploratory bouts in the cylinder test. *Behavioural brain research*, *151*(1-2), 125–135.

Gordon, G., Fonio, E., & Ahissar, E. (2014, September). Emergent Exploration via Novelty Management. *Journal of Neuroscience*, *34*(38), 12646–12661. doi: 10.1523/JNEUROSCI.1872-14.2014

Graybiel, A. M., & Grafton, S. T. (2015). The striatum: where skills and habits meet. *Cold Spring Harbor perspectives in biology*, *7*(8), a021691.

Lai, L., Huang, A. Z., & Gershman, S. J. (2022). Action chunking as policy compression.

Lomholt, M. A., Tal, K., Metzler, R., & Joseph, K. (2008). Lévy strategies in intermittent search processes are advantageous. *Proceedings of the National Academy of Sciences*, *105*(32), 11055–11059.

Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*, *110*(6), 914–934.

Nagy, M., Horicsányi, A., Kubinyi, E., Couzin, I. D., Vásárhelyi, G., Flack, A., & Vicsek, T. (2020). Synergistic benefits of group search in rats. *Current Biology*, *30*(23), 4733–4738.

Palyulin, V. V., Chechkin, A. V., & Metzler, R. (2014, February). Lévy flights do not always optimize random blind search for sparse targets. *Proceedings of the National Academy of Sciences*, *111*(8), 2931–2936. doi: 10.1073/pnas.1320424111

Rinaldi, A., De Leonibus, E., Cifra, A., Torromino, G., Minicocci, E., De Sanctis, E., … Mele, A. (2020). Flexible use of allocentric and egocentric spatial memories activates differential neural networks in mice. *Scientific reports*, *10*(1), 11338.

Rosenberg, M., Zhang, T., Perona, P., & Meister, M. (2021). Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *Elife*, *10*, e66175.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181–211.

Tchernichovski, O., Benjamini, Y., & Golani, I. (1998). The dynamics of long-term exploration in the rat. *Biological cybernetics*, *78*(6), 423–432.

Vallianatou, C.-A., Alonso, A., Aleman, A. Z., Genzel, L., & Stella, F. (2021, September). Learning-Induced Shifts in Mice Navigational Strategies Are Unveiled by a Minimal Behavioral Model of Spatial Exploration. *eneuro*, *8*(5), ENEURO.0553–20.2021. doi: 10.1523/ENEURO.0553-20.2021

Vijayabaskaran, S., & Cheng, S. (2022). Navigation task and action space drive the emergence of egocentric and allocentric spatial representations. *PLOS Computational Biology*, *18*(10), e1010320.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (1999, October). Optimizing the success of random searches. *Nature*, *401*(6756), 911–914. doi: 10.1038/44831

Viswanathan, G. M., Da Luz, M. G., Raposo, E. P., & Stanley, H. E. (2011). *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press.