Spatial subgoal learning in the mouse: behavioral and computational mechanisms

Philip Shamash

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
of
University College London.

Branco Laboratory
Sainsbury Wellcome Centre for Neural Circuits and Behaviour
University College London

September 2022

I, Philip Shamash, declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been acknowledged in the work.

Abstract

Here we aim to better understand how animals navigate structured environments. The prevailing wisdom is that they can select among two distinct approaches: querying a mental map of the environment or repeating previously successful trajectories to a goal. However, this dichotomy has been built around data from rodents trained to solve mazes, and it is unclear how it applies to more naturalistic scenarios such as self-motivated navigation in open environments with obstacles. In this project, we leveraged instinctive escape behavior in mice to investigate how rodents use a period of exploration to learn about goals and obstacles in an unfamiliar environment. In our most basic assay, mice explore an environment with a shelter and an obstacle for 5-20 minutes and then we present threat stimuli to trigger escapes to shelter. After 5-10 minutes of exploration, mice took inefficient paths to the shelter, often nearly running into the obstacle and then relying on visual and tactile cues to avoid it. Within twenty minutes, however, they spontaneously developed an efficient subgoal strategy, escaping directly to the obstacle edge before heading to the shelter. Mice escaped in this manner even if the obstacle was removed, suggesting that they had memorized a mental map of subgoals. Unlike typical models of map-based planning, however, we found that investigating the obstacle was not important for updating the map. Instead, learning resembled trajectory repetition: mice had to execute 'practice runs' toward an obstacle edge in order to memorize subgoals. To test this hypothesis directly, we developed a closed-loop neural manipulation, interrupting spontaneous practice runs by stimulating premotor cortex. This manipulation successfully prevented subgoal learning, whereas several control manipulations did not. We modelled these results using a panel of reinforcement learning approaches and found that mice behavior is best matched by systems that explore in a non-uniform manner and possess a high-level spatial representation of regions in the arena. We conclude that mice use practice runs to learn useful subgoals and integrate them into a hierarchical cognitive map of their surroundings. These results broaden our understanding of the cognitive toolkit that mammals use to acquire spatial knowledge.

Impact statement

Rodent spatial learning is a key model system for investigating how the brain stores information about the world. Results from this domain have informed a wide array of modern views in psychology, neurobiology, and artificial intelligence. In this project, we have developed an assay that provides an increase in the complexity and naturalism of spatial-learning experiments while also offering straightforward quantification methods. These innovations allowed us to make new discoveries about how mice build up 'mental maps' of their environment. Our findings have implications for the following four areas.

1) A key aim of neuroscience is to discover how the brain encodes maps or 'models' of the world. But without a thorough understanding of how animals actually use these maps, this project runs the risk of becoming highly hypothetical in nature. The results of our behavioral studies can help.

2) Machine learning agents require a great deal of experience to learn a new task, unlike animals. Our results describe how mice build of a mental map within minutes of entering a new environment for the first time ever. These results could inspire new artificial intelligence approaches for rapid learning in real-world environments.

3) In cognitive science, our results support the idea of enactive learning. This notion suggests that, to explain how we learn, it is crucial to describe the actions that we take to extract information from the world. This may apply to human learning as well, in particular human infants, who face a similar task of building an understanding of the world for the first time ever.

4) Finally, in biomedical research, rodent spatial memory assays are used to test neurological disease models and potential therapeutics. Those assays generally require an animal to be trained for many days prior to testing. Our assay involves only a 20-minute learning period with no pretraining, which could substantially speed up this process.

Acknowledgements

Thank you to my advisor Tiago Branco. I have been lucky to have had not just a supervisor but also a mentor, who always made himself available to discuss the latest Grand Unified Model of Escape and who took the time to make my work and my approach to science much better.

Thank you to the Branco lab for being an incredibly helpful and supportive group of people. In terms of direct contributions, thanks to: Ruben do Vale for originally teaching me how to do behavioral experiments (which is kind of the whole thesis); Kostas Bestios for programming data acquisition software; Federico Claudi for programming the backbone of my first data analysis pipeline; Nabhojit Banerjee for developing the food-seeking assay; Dario Campagner for training on surgery and developing that assay with Nabho; Sarah Olesen for doing most of the work on those food-seeking experiments; and Yiota Iordanidou for helping with behavioral experiments and histology preparation.

Thank you to Sebastian Lee for collaborating on the reinforcement learning section, including most of the programming of the simulations. Thank you to Sarah Elnozahy for facilitating this collaboration.

Thank you to my non-researcher colleagues such as FabLabs, the NRF, and Karen Fergus, for making things much easier.

Thank you to those who took the time to give feedback on my papers or presentations, including: Catarina Albergaria, Caswell Barry, Tim Behrens, Dora Biro, Neil Burgess, Dario Campagner, Federico Claudi, Claudia Clopath, William Dorell, Tom Mrsic-Flogel, Loren Frank, Alex Fratzl, Jesse Geerts, Tom George, Chris Hall, Yoh Isogai, Ted Moscovitz, Daniel Regester, Andrew Saxe, Kim Stachenfeld, Marcus Stephenson-Jones, Yu Lin Tan, Ruben do Vale, and Peter Vincent.

Thank you to my colleagues who deserve some thanks but don't quite fit into any of those categories: Mitra Javadzadeh, Simon Thompson, Ivan Voitov.

Thanks to my parents and my sisters for getting me here and for emanating support all the while. Thanks to Zeynab Blondin Diop for making sure my PhD was not the most important thing I got out of my time in London.

And thank you to Ilyes Khemakhem for this fine LaTeXtemplate.

Contents

Abstract 5

Impact statement 7

Acknowledgements 9

Contents 11

List of Figures 13

List of Tables 15

Chapter 1: Introduction 19
- 1.1 The place and response strategies for spatial learning . . . 19
- 1.2 Multi-step spatial navigation . . . 25
- 1.3 Escape behavior . . . 31
- 1.4 Our questions . . . 33

Chapter 2: Subgoal memorization 35
- 2.1 Mice quickly learn efficient escape routes . . . 35
- 2.2 A spatial memory strategy for navigating the obstacle . . . 37
- 2.3 Spatial learning here consists of memorizing subgoal locations . . . 40
- 2.4 Subgoal memorization also supports food-seeking routes . . . 49
- 2.5 Interim discussion . . . 53

Chapter 3: Action-driven mapping 57
- 3.1 Closed-loop activation of premotor cortex blocks edge-vector runs . . . 57
- 3.2 Interrupting edge-vector runs abolishes subgoal learning . . . 59
- 3.3 Subgoal-escape start points are determined by spatial rules . . . 64
- 3.4 Interim discussion . . . 68

Chapter 4: Reinforcement learning models of escape behavior 73
- 4.1 Introduction to RL models of navigation . . . 73
- 4.2 Modelling navigation in the obstacle-removal experiment . . . 75
- 4.3 Interim discussion . . . 82

Chapter 5: General discussion **87**
5.1 Conceptual overview 87
5.2 Methodological innovations 87
5.3 Key limitations 89
5.4 Future directions 90

Chapter 6: Methods **93**
6.1 Animals 93
6.2 Behavioral Assays 94
6.3 Neural manipulations 99
6.4 Analysis 101
6.5 Reinforcement learning simulations 105

Bibliography **113**

List of Figures

Fig. 1.1 The T-maze experiment 20

Fig. 2.1 Mice rapidly learn efficient escape trajectories past obstacles 36
Fig. 2.2 Escapes in the presence of an obstacle: extended results 37
Fig. 2.3 Platforms with the wall obstacle and hole obstacle for ch. 2 38
Fig. 2.4 Escapes with an obstacle in complete darkness 39
Fig. 2.5 Mice escape around obstacles even after they are removed 40
Fig. 2.6 Habitual, egocentric movements do not explain the spatial memory . 42
Fig. 2.7 Homing runs, turn angles, and heading directions: extended results 43
Fig. 2.8 Mice memorize previously targeted subgoal locations 45
Fig. 2.9 Mice memorize previously targeted subgoal locations: extended results 46
Fig. 2.10 Mice memorize previously targeted subgoal locations 48
Fig. 2.11 Edge-directed movements in different environments 49
Fig. 2.12 Edge vectors persist long after obstacle removal 50
Fig. 2.13 Training mice to approach and lick a spout in response to a tone . . 51
Fig. 2.14 Obstacle experience affects food-seeking but not exploratory paths . 52

Fig. 3.1 A closed-loop neural manipulation interrupts edge-vector runs . . . 58
Fig. 3.2 Optogenetic stimulation of right premotor cortex 60
Fig. 3.3 Behavioral platforms for ch. 3 61
Fig. 3.4 (No) effect of optogenetic stimulation on exploration 62
Fig. 3.6 Blocking edge-vector runs after allowing two runs 64
Fig. 3.7 Blocking edge-to-shelter runs 64
Fig. 3.8 Obstacle-removal escapes with a modified threat zone 65
Fig. 3.9 Subgoal escapes do not need to start near practice runs 66
Fig. 3.10 Subgoal-escape start points are determined by spatial rules 67

Fig. 4.1 The core reinforcement learning models we selected 74
Fig. 4.2 Reinforcement learning simulation protocol 75
Fig. 4.3 Reinforcement learning model results and interpretation 77
Fig. 4.4 Reinforcement learning models: extended results 79
Fig. 4.5 Hierarchical agent learning speed and behavior 81

List of Tables

Tab. 6.1 All experiments in chapter 2 94
Tab. 6.2 All experiments in chapter 3 95
Tab. 6.3 Hyper-parameters 110
Tab. 6.4 Training steps needed for the RL models to learn escape routes . . 111

It was all familiar; this turning, that stile, that cut across the fields. Hours he would spend thus, with his pipe, of an evening, thinking up and down and in and out of the old familiar lanes and commons, which were all stuck about with the history of that campaign there, the life of this statesman here, with poems and with anecdotes, with figures too...but at length the lane, the field, the common, the fruitful nut-tree and the flowering hedge led him on to that further turn of the road where he dismounted always, tied his horse to a tree, and proceeded on foot alone.

Virginia Woolf, *To the Lighthouse*

Chapter 1

Introduction

1.1 The place and response strategies for spatial learning

The idea that animals might use mental models or maps of the external world started to gain prominence in the 1930s with the work of Edward Tolman (O'Keefe and Nadel, 1978). Tolman and his colleagues at Berkeley showed that rats could solve mazes by devising routes that they had never performed before, through the power of 'cognitive maps' and mental computation (Tolman, 1948). A second camp disputed these findings. The stimulus-response group, centered at Yale, instead believed that learning could always be described in some sense as a process of chaining together series of well practiced actions - essentially a complex form of trial and error or trial and success (Hull, 1934).

For a time, the T-maze was at the center of this dispute (O'Keefe and Nadel, 1978). The T-maze is a simple spatial learning protocol in which an animal is trained to make a single left or right decision (right in our example) in order to get a food reward (Fig. 1.1a). After training in this setup over many days, the animal is then tested with a probe trial (Fig. 1.1b); The rat is placed at the opposite end, and the experimenter records which side it turns to to search for food. If the rat turns right, then it is supporting Hull's 'response' or 'action reinforcement' hypothesis: it is simply repeating the same rightward turning movement that it used during the training phase. However, if the rat turns left, then it is demonstrating Tolman's 'place,' or 'cognitive map' hypothesis. In this case, it has inferred a new action that leads to the previously rewarded location in space.

Which strategy did the rats use? Tulving and Madigan, 1970 summarize the results at the time as follows:

> Place-learning organisms, guided by cognitive maps in their head, successfully negotiated obstacle courses to food at Berkeley, while their response-learning

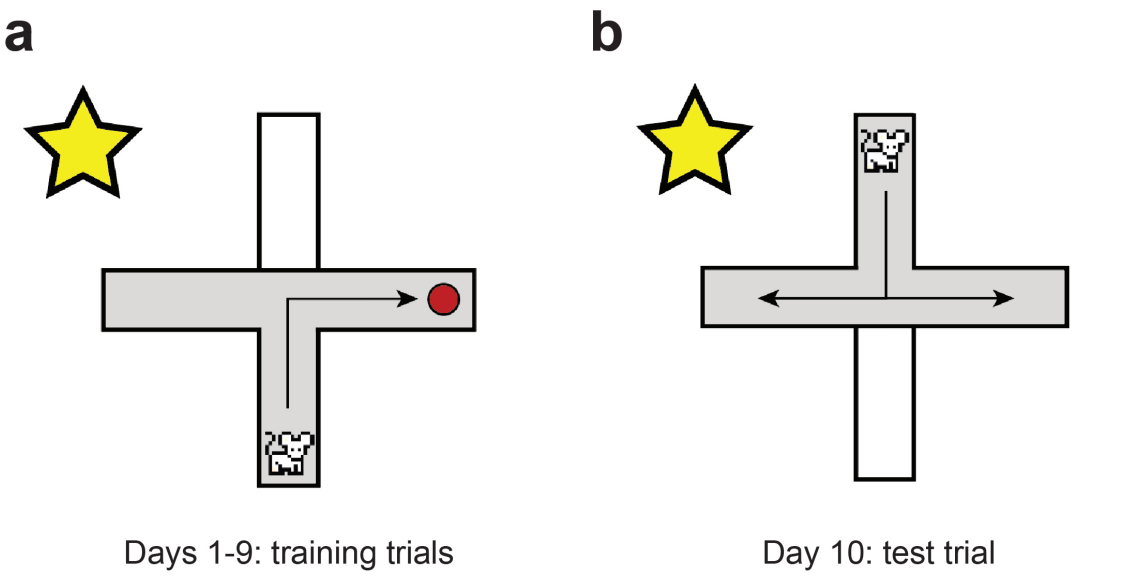

Figure 1.1: *The T-maze experiment*

(a) During the training phase, the rat or mouse typically allowed one food-seeking trial per day. The yellow star represents a salient, static cue that is visible to the animal, allowing it to determine its location in the environment. **(b)** During the testing phase, there is no reward and the animal's behavior is measured.

counterparts, propelled by habits and drives, performed similar feats at Yale.

Thus, the debate carried on. It was eventually resolved in the 1950s through the insight that both strategies are viable options; the result depends on the parameters of the training protocol (Restle, 1957). For example, having more salient cues in the experimental room helps the animal determine its location and favors place learning. Training the animal for more days, on the other hand, favors response learning.

This dual-system compromise cemented even further with the concomitant developments of reinforcement learning theory and better experimental paradigms to test place-based learning. In the 1970s, the Rescorla-Wagner model of incremental learning based on receiving rewards (Rescorla, 1972) emerged as a successful, quantitative model of animal behaviors such as Pavlovian conditioning (Siegel and Allan, 1996). Meanwhile, the Morris water maze definitively demonstrated that animals are able to navigate to a goal location using a sense of space (Morris, 1981). In this assay, a platform is submerged in a cylindrical tank filled with opaque water, and a rat must swim around until finding the platform. After several days of experience, the rat is able to swim directly to the invisible platform location regardless of where it was initially placed in the tank, on the basis distal visual cues positioned around the room by the experimenter. Moreover, if the goal platform is displaced, rats can infer routes to that new location after only a single trial (Hamilton et al., 2004). With a greater level of experimental control than the T-maze, these results showed that

rats can infer the action that will get them to a goal location, without needing to repeatedly practice that movement.

It was around this period that a series of striking discoveries were made, forever fusing the fields of rodent spatial navigation and systems neuroscience. First, the spiking activity of neurons in the hippocampus was found to correlate exactly with a rat's position within its environment (O'Keefe and Dostrovsky, 1971). Each of these 'place cells' only becomes active when the rat passes through a certain location in space (e.g. the northwest corner of its cage), offering concrete evidence that rats mentally keep track of their location in space. Second, it was found that the place and response strategies are neurally dissociable. Disrupting the hippocampus impaired place learning, causing animals to use response learning in the T-maze and rendering them unable to solve the Morris water maze (Packard et al., 1989; Packard and McGaugh, 1992; Packard and McGaugh, 1996). Disrupting the basal ganglia disrupted action reinforcement, causing rats to stick with the place strategy in the T-maze and not impairing learning in the Morris water maze. The implication of these findings is that place and response learning represent truly distinct neuro-cognitive modules in the brain.

In this thesis, we will end up questioning this dominant two-system view and specifically its capacity to explain animal navigation in more naturalistic scenarios. Before that, however, we must further unpack this dichotomy. First, we will describe several sub-strategies that are all labelled as 'response' or 'place' based. Second, we will home in on some key, implicit differences between place and response learning. Finally, we will mention alternative strategies for navigation that fit into neither category.

Sub-strategies

Response strategies and place-based navigation each have two primary sub-strategies. Response learning can take either a praxic or taxon approach (Redish et al., 1999), while internal maps can be either metric or topological (Trullier et al., 1997).

Praxic navigation

The first response strategy, praxic navigation, involves executing a fixed motor program. An example of this would be: walking forward 6 steps, turning left 90 degrees, and then walking forward again. A notable example of this strategy comes from rats with hippocampal lesions attempting to learn the Morris water maze. These lesioned rats were only able to navigate to the hidden platform if their starting

location and the goal location were the same each day, suggesting that they had memorized a particular sequence of movements (Eichenbaum et al., 1990).

Taxon navigation

The other response strategy is taxon navigation. Here, animals navigate toward a sensory cue that signals the presence of the goal. This cue could be a sight, sound or odor - anything that the animal can, through a sensory-motor loop, directly approach. While praxic navigation is always learned through practice, taxon navigation can be either learned or instinctive. In the Morris water maze, animals can only use this strategy if the goal platform is modified to have a visible cue poking out of the water (Packard and McGaugh, 1992). Taxon and praxic strategies can be chained together in sequences, to generate complex multi-step routes (Eichenbaum et al., 1990).

Topological mapping

Just like response learning, the mapping strategy comes in two styles. First, animals can build up a topological map: a representation of which locations are adjacent to which other locations (Trullier et al., 1997). This kind of map can be expressed in mathematical terms as a graph, with nodes representing a previously observed location and edges representing adjacency between nodes. Topological maps could be used to solve the Morris water maze, by selecting the series of locations that occur in a straight line between the animals and the goal. However, this strategy is most relevant for assessing complicated routes. In particular, it can be used to determine whether a candidate path can connect to the goal and, if so, how long the path is (i.e., how many adjacent locations need to be traversed before arriving at the goal). This capacity was famously demonstrated in Tolman and Honzik, 1930's detour task. There, rats could select among three increasingly long routes to a reward (paths 1-3). After choosing the shortest route (path 1), they encountered an unexpected barrier either blocking only path 1 or blocking both path 1 and path 2. If the barrier blocked only path 1, upon returning to the starting point the rats would select path 2; however, with the second barrier, they would select path 3. This indicates that the rats maintained a representation of the fact that path 2 was still connected to the goal with the first barrier but not with the second.

Metric mapping

The second type of cognitive map is a metric map. In a metric map, locations are related by distance and direction rather than adjacency. For example, locations could be represented as a set of (x,y) coordinates on a flat, 2D Euclidean surface.

This kind of representation is powerful as it allows for shortcuts across previously unvisited territory - the key to solving the Morris water maze. Chapuis, 1987 further demonstrated this strategy in dogs by leading the dogs along two paths (point A to point B and point A to point C) and then testing their ability to navigate from point B to point C. Even though the B-to-C and C-to-B routes had never been experienced, the dogs could navigate directly from point B to point C. This indicates that they knew the distance and direction between those two locations.

How the place and response strategies differ

Cognitive mapping and action reinforcement differ in many ways, but the three that are most relevant for us are their *reference frames, learning style, and learning speed*. Basal gangliar reinforcement learning is associated with egocentric, practice-driven, gradual learning, while hippocampal mapping is associated with allocentric, planning-based, rapid learning.

Egocentric vs. allocentric reference frames

The place and response strategies differ in the type of reference frame that the animal uses. An egocentric reference frame is defined in relation to the actor's body. Turning right, walking six steps forward, and foveating a visual landmark are all egocentric actions. Allocentric reference frames, on the other hand, are defined in relation to the layout of the outside world. Moving toward the east side of the room or toward the side of the room opposite from the giant yellow star are allocentric actions. Thus, remembering where the reward was located within the experimental room is an allocentric strategy. Note that this is deemed an allocentric strategy even though it must eventually turn into an egocentric motor command (e.g. turn left).

Practice vs. planning

A third distinction is how the animal goes about learning the route. The response strategy requires practice. This means that the learned action (e.g. turning right at the junction) has been previously enacted and then followed by positive reinforcement. These practice actions are typically motivated rather than random or meandering movements (O'Keefe and Nadel, 1978). With planning, on the other hand, practice runs are not necessary. By merely observing (and memorizing) the structure of the environment, e.g. while walking from the reward to the start of the maze, the animal can infer a novel route, e.g. how to get from the start of the maze to the reward. Tolman and Honzik, 1930 famously demonstrated the potency of this 'latent'

learning style, by showing that rats that spent time in a maze without any reward were later able to navigate to a reward in that environment just as well as rats that had practiced running through the maze to get the reward.

Incremental vs. all-or-none learning

Finally, these strategies differ in the number of observations/actions needed to learn a route. A response behavior is generally reinforced incrementally over tens to hundreds of practice trials. Even the simplest action reinforcement, such as mice learning to poke a button with their nose to release a food pellet, takes tens of learning trials (Baron and Meltzer, 2001). These behaviors are therefore also inflexible and can persist long after they cease to result in positive reinforcement. Integrating a new location into an animal's internal map of its environment, on the other hand, can happen after visiting that location for the very first time (Bittner et al., 2017). Similarly, with the cognitive map approach, a change the environment only needs to be observed once in order for the animal to update its route (Hsiao, 1929; O'Keefe and Nadel, 1978).

Revisiting the place-response dichotomy in light of these differences

Because these three properties are correlated, it is parsimonious to chunk them into two separate neurocognitive systems (Geerts et al., 2020). However, there is no theoretical necessity for allocentric reference frames, planning, and rapid learning to all go hand in hand. For example, Daw et al., 2005 showed that practicing sophisticated actions can produce the kind of flexible behavior normally associated with planning. Along these lines, it would also be plausible to discover a behavior that uses an allocentric reference frame but is learned through practice (e.g. practicing running to the west side of the room) or a learning mechanism that is both practice-based and all-or-none (e.g. learning an action after a single practice runs).

Alternative strategies

Animals have more navigational strategies at their disposal than the place and response strategies. Two additional categories of navigational behavior are search strategies and path integration.

Search strategies

If the animal believes that a goal is nearby but does not know exactly where to go or what action to take, the remaining option is to engage in a search strategy. With

random search, animals repeatedly select a movement and distance to travel, from some random distribution. This strategy has been largely studied in the context of foraging behavior. There, random search allows animals to discover the maximal amount of food patches in the minimum amount of time (Bovet and Benhamou, 1988; Viswanathan et al., 2000). In more restricted environments with a single goal, animals can employ a systematic search. For instance, if a rat knows that the Morris water maze platform is somewhere near the perimeter of the tank, it can swim in a circular pattern around the perimeter until stumbling upon the goal (Janus, 2004). Finally, if there are multiple identical cues but only some of them lead to reward, the animal can employ a serial search (Patil et al., 2009). This entails investigating each cue until the rewarding one is found.

Path integration

Animals can keep track of all the distances and angles that they have moved by since visiting a reward and use this information to navigate back to it (Etienne and Jeffery, 2004). This strategy is called 'path integration' or 'dead reckoning.' Note that it does not involve retracing steps on the outbound path to get back to the starting point; that would actually be most consistent with the topological mapping strategy. Instead, animals compute a vector (generally both distance and direction) toward the goal, which can be used to navigate directly there. Path integration is an ancient strategy that occurs across the animal kingdom, from insects to humans (Müller and Wehner, 1988; Etienne and Jeffery, 2004). In tasks calling for direct routes to a goal, this strategy produces similar behavior to the metric mapping strategy, since both strategies allow animals to target goals across previously unvisited territory. Path integration is generally described as capable of maintaining a vector to one goal at a time, so it cannot on its own explain how animals are able to compute multi-step routes to a goal. However, it is useful for *constructing* a metric map of the environment, by furnishing the animal with the distance and direction between pairs of recently visited landmarks.

1.2 Multi-step spatial navigation

Moving directly toward a nearby goal is a much simpler problem than devising multi-step routes past obstacles to get to a goal that is out of sight. Both the response and mapping strategies must be substantively extended in order to account for navigation in structured environments. For the response strategy, directly approaching a cue associated with reward or learning a single action does not work anymore. Instead,

multiple actions or cues must be chained together into a route. Similarly, looking up the distance and direction to the goal with the metric mapping strategy no longer suffices. In this section we examine how animals navigate to goals in structured environments, in practice and in theory.

Navigation in natural environments

Animals are capable of stunning feats of navigation over vast distances (Able, 1980). Notably, migratory insects (Reppert and Roode, 2018), birds (Chernetsov et al., 2008) and mammals (Horton et al., 2011) are all able to find their way across thousands of kilometers to transit between their summer and winter sites. Capitalizing on the ease with which animals navigate in the wild, experiments in natural environments have revealed an array of strategies for tackling long, complicated journeys. One approach, typical of migratory species, is to use an innate solar or stellar compass to guide the overall direction of movement and then to deal with individual obstacles as they are encountered (Able, 1980). On shorter spatial scales, a similar mechanism can be used with path integration (Huber and Knaden, 2015) or a metric cognitive map (Tsoar et al., 2011) guiding the direction of travel rather than the sun or stars.

Another common approach to navigating natural environments is to develop stereotyped routes between foraging and home sites. This approach is widespread, also appearing across insects, birds and mammals. Visually guided route memorization has been demonstrated in both desert ants (Kohler and Wehner, 2005; Cheng et al., 2009) and homing pigeons (Biro et al., 2004). Over the course of 20 training runs from a site about 10 m (in ants) or 10 km (in pigeons) from the animals' home, both species learned to consistently follow idiosyncratic homebound routes guided by a series of visual landmarks along the way. In ants it has been further shown that these routes are unidirectional, i.e. ants cannot infer how to get from the nest to the feeding site after learning a route from the feeding site to the nest (Wehner et al., 2006). This suggests a response-style strategy rather than a map. Route stereotypy also occurs in species with a poorly developed sense sight, such as the water shrew. The water shrew instead employs a praxic strategy for route learning, memorizing a precise sequence of movements to take them around their environment. Lorenz, 1949 describes this behavior in a terrarium built to observe a family of water shrews:

> Once the shrew is well settled in its path-habits it is as strictly bound to them as a railway engine to its tracks...the shrews, running along the wall, were accustomed to jump on and off the stones which lay right in their path. If I moved the stones out of the runway...the shrews would jump right up into the air in the place where the stone should have been; they came down with a

> jarring bump...For this animal the geometric axiom that a straight line is the shortest distance between two points simply does not hold good. To them, the shortest line is always the accustomed path.

Rodents also exhibit stereotyped foraging routes in natural habitats (Thompson, 1982; Benhamou, 1991). Thompson, 1982, for example, found that rats visited the same feeding sites each night and always in the same order, typically returning to their den in between each sequences. One mouse species has even evolved a tendency to place small objects along its paths to mark familiar paths back to its refuge (Stopka and Macdonald, 2003). Still, the roles of visual guidance, action reinforcement, path integration, and cognitive mapping in each of these behaviors remain unknown. To pinpoint the cognitive mechanisms of navigation in structured environments, ethological studies should thus be supplemented with rigorous laboratory work, which can provide a level of experimental control unattainable in natural settings.

Navigation in mazes

A key method to investigate multi-step navigation in the lab has been to train rats and mice on complex mazes. The most common is the multiple T maze, in which the animal faces a series of three-way junction at which they can turn left or right (Sharma et al., 2010). Each decision will lead them either to an empty dead end, to the next junction, or at the end to a dead end with a reward. Animals learn to navigate to reward without making any 'errors' (i.e. entering a dead end) over the course of about 1-5 trials per day for about one week (Tolman and Honzik, 1930; Schmitzer-Torbert and Redish, 2002; Sharma et al., 2010). Using this setup, scores of factors have been found to modulate multi-step spatial learning, from the number of days since the previous session (Sharma et al., 2010), to modulation of cholinergic signaling (Krejcova et al., 2004), to the composer of music played during development [1] (Rauscher et al., 1998; Aoun et al., 2005).

In behavioral neuroscience, the most celebrated insight from the multiple T maze is Tolman and Honzik, 1930's aforementioned latent learning experiment. They found that if they allowed rats to explore the maze without any reward for several days before the first proper training session, learning was dramatically accelerated. The prevailing interpretation is that rats learn about the structure of the maze during exploration, and that they can henceforth plan out the series of left-right decisions needed to get to the goal (Behrens et al., 2018). However, in our view this has not really been shown. For one, all that rats need to learn in this environment

[1] Rats that listened to Mozart outperformed the silence, white-noise, Beethoven and Philip Glass groups.

is to avoid each individual dead end; multi-step planning is not required to solve a multiple T maze. Moreover, learning is also accelerated if rats are trained in one maze and then get tested in a new maze with a different structure (i.e. a different sequence of dead ends; Schmitzer-Torbert and Redish, 2002). Thus, the latent learning in Tolman's experiment is not necessarily about the exact structure of the maze but could instead reflect learning about general laboratory maze setup (Dashiell, 1920). Overall, multi-step maze learning has served as a useful testbed of the factors influencing complex spatial learning, but it has not revealed much about the strategies that rodents use to learn about the structure of their environment.

Navigation in the presence of an obstacle

A key limitation in maze-learning paradigms is the way in which behavior is boiled down to a scalar quantity: the number of errors (i.e. dead-end entries) before finding the reward. The low dimensionality of behavior here makes it difficult to identify the exact reasons why a rat might have made an error. One way to investigate multi-step navigation in a more open-ended setting is to interpose a barrier between the subject and its goal and to analyze paths to the goal (Kabadayi et al., 2018). Initially, transparent or wire obstacles were used, in order to test whether animals could inhibit the direct visual guidance strategy and instead plan a detour around the obstacle (in dogs: Hobhouse, 1901; in chickens: Kohler, 1925; in chimpanzees: Thorndike, 1911; in toads: Collett, 1982; in octopuses: Wells, 1967; in human infants: Lockman and Adams, 2001).

By using opaque obstacles, researchers can further investigate how animals are able to memorize and calculate routes to a goal that is out of sight. If the barrier was added within a few seconds of when the animal attempts to reach the goal, then animals can simply maintain in working memory the direction to the goal. This ability has been demonstrated in frogs (Ingle, 1990) and in two-day-old chicks (Regolin et al., 1995). Over longer timescales, more robust spatial memory approaches must be used. Fiddler crabs, for example, navigate to their burrow using path integration combined with simple, sensory-guided obstacle circumnavigation (Layne et al., 2003). Desert ants initially use a similar strategy to navigate from a feeding site to a nest that is blocked by a large barrier; over time, they then learn to directly visually target the obstacle edge and then to move toward the nest using the (visually informed) axis of the obstacle as a cue (Collett et al., 2001). These strategies, relying on path integration and visual guidance, are usually effective but fail to intelligently select the overall most efficient route to the goal. Gerbils (Ellard and Eller, 2009), cats (Poucet et al., 1983) and dogs (Chapuis et al., 1983) have all been found to use

spatial memory to run to the side of the obstacle that allows for the shortest overall path to their goal, taking into account their starting position and the location of the goal. However, the nature of this spatial memory (e.g. response learning vs. mapping) and how it was learned was not investigated.

Overall, we find navigation with an obstacle to be a promising approach for looking at how mice learn to navigate a structured environment. No prior work has thoroughly dissected how spatial memory is used to navigate around an obstacle. Several approaches might be taken. For one, mice might move toward the goal using path integration and then use vision to avoid running into the obstacle. Second, mice could practice and memorize the sequence of actions or cues that had previously gotten them from the starting point to the goal (the response strategy). Third, they could build an internal map of the environment and use this map to calculate a route past the obstacle (the cognitive-map strategy). This third approach remains the most mysterious, so in our results we will focus on identifying and describing a regime where mice might use a mapping strategy to navigate an obstacle.

Computational models of multi-step mapping

While experimental evidence on how animals map out and navigate cluttered environments is sparse, theorists have been able to generate a wide variety of plausible solutions to this problem *in silico*. Following our classification of mapping strategies in the previous section, these fall into several categories: metric maps, topological maps, and combination approaches. These are all perfectly able to generate efficient multi-step routes to a goal, but they differ in the representations and route-searching algorithms that they employ to do so.

Burgess et al., 1994 demonstrate how a metric map could suffice to navigate to a goal that is hidden behind an obstacle. In this model, an agent uses simulated neurons from the hippocampal formation that track the distance and direction to a desired goal location. By adding an 'obstacle-to-avoid' feature to the internal map of locations, the animal can get to its goal while deviating around the obstacle.

Topological maps are naturally suited to multi-step routes, since they are intrinsically able to lead the animal along a sequence of adjacent locations. They are also the standard approach for model-based reinforcement learning models of navigation (see ch. 4). For example, Spiers and Gilbert, 2015 suggest a model in which animals build up a representation of adjacency relationship between locations in the environment (i.e. a graph) and then implement a tree-search algorithm to find the shortest route to a goal. This algorithm simulates a set of candidate routes starting at the animal's position and branching out across adjacent nodes until

reaching the goal.

Metric and topological mapping can also be used in parallel. Edvardsen et al., 2020 implement a system in which a metric map sets the initial heading direction, and if the agent gets stuck at an obstacle, a topological map comes online to generate a new heading direction. Specifically, animals start by heading directly toward their goal using a metric map inspired by a spatial cell type called 'grid cells'. If this strategy leads to a dead end, the animal uses a topological-map tree search to identify a new subgoal location at one of the obstacle's edges. The animal then uses its metric map to compute a direct shortcut to the subgoal, and this process repeats.

Finally, (Solway et al., 2014) describe a model that can combine the advantages of both topological and metric maps. In their model, a topological map is used to identify key points in the environment such as dead ends (i.e. a node with only one neighboring node), or doorways and obstacle edges (nodes that are adjacent to two otherwise unconnected regions). These key points are then connected together in a metric map representing the distances and directions between them. The agent can search through this high-level map and calculate which key point(s) it should target as a subgoal on the way to its ultimate goal. This algorithm thus capitalizes on topological maps' capacity to represent connectivity structure as well as metric maps' ability to represent distances and to generate direct shortcuts. Despite this strategy so far only being described in humans (Solway et al., 2014) and monkeys (Teichroeb and Smeltzer, 2018), this model will end up shedding more light on our mouse-behavior results than the other three.

Adding a dash of naturalism to laboratory experiments

Overall, it is known that animals are capable of impressive feats of navigation in the wild; rodent spatial learning has served as a key model system for neuroscience and psychology for about a century; and a wide array of plausible computational models of multi-step mapping have sprung from this field. Nonetheless, it remains mysterious how rodents go about spontaneously and rapidly building up spatial knowledge in new environments. The strategies and dichotomies we have gone over have largely been informed by repeatedly placing rodents in constrained mazes until they learn to navigate to a food reward and then removing them each time they reach the reward (Tolman and Honzik, 1930; O'Keefe and Nadel, 1978). Moreover, rodent navigation studies have focused on the cues that animals use to pinpoint locations, rather than the actions that they exploit during exploration (Restle, 1957; Morris, 1981; Cheng et al., 2013).

In a natural setting, however, spatial learning can occur within minutes rather

than days and is carried out via complex, internally generated exploration patterns. Mice explore in a highly structured manner, punctuating investigatory bouts along boundaries with rapid lunges to a familiar, enclosed space or a visually salient object (Crowcroft, 1966). Given the mismatch between nature and experiment, it is unclear how well previous classifications of navigation strategies map onto the learning procedures that animals use for naturalistic navigation. Thus, in addition to obstacle navigation, we sought a behavioral protocol that exploits spontaneous motivation to learn about the environment rather than a session-based training protocol. Escape behavior in mice furnished this additional degree of naturalism.

1.3 Escape behavior

The mouse is a cautious creature. Released into a new environment, it immediately searches for crevices to hide in, tests out routes to safety, and inspects any objects that might obstruct its path. If any indication of a predator should appear, it can use what it has leaned to quickly find its way to safety. Here, we have imported this behavior - escape to a shelter in an obstacle-laden environment - to the laboratory.

Diverse animals, including fishes, lizards, crabs, birds, and rodents, respond to threats by escaping to a familiar shelter (Cooper and Blumstein, 2015). While escape responses are often caused by innately threatening stimuli, animals select where, when and how to escape using sophisticated cognitive processes (Evans et al., 2019). For example, escape routes are modulated by the presence of nearby conspecifics (Dill and Ydenberg, 1987; Mateo, 1996), the type of threat stimulus (Blanchard and Blanchard, 1988; Reimers and Eftestøl, 2012; Yilmaz and Meister, 2013), and the layout of the local environment (Lagos et al., 2009; Zani et al., 2009; Vale et al., 2017).

For prey species such as mice, it is particularly important to find fast and efficient routes to shelter, because it reduces exposure to potential predators (Lima and Dill, 1990). This is a challenging task: natural environments are complex, and wild animals must compute multi-step routes taking into account uneven terrain, obstacles, and dynamically changing environments. Ethological studies of wild rodents have emphasized the roles of locating salient landmarks (Drickamer and Stuart, 1984; McMillan and Kaufman, 1995) and adhering to previously used routes (Thompson, 1982; Benhamou, 1991) in overcoming these challenges. In addition, rodents are known to spontaneously shuttle back and forth between the outside and the 'home' (Crowcroft, 1966), which could help them to rapidly learn escape routes in case a predator appears. However, all these observational studies are limited in their ability

to identify the precise actions, cues, and learning rules that animals use to compute escape routes.

Studying the homing behavior of rodents in the lab offers a powerful, complementary window into spatial behavior (Evans et al., 2019). The standard assay for measuring navigation strategies in the context of rodent escape behavior is the Barnes Maze assay (Barnes, 1979). There, rodents are placed in an open-field arena with an underground shelter and are presented with an ongoing aversive stimulus such as a bright light. Rodents locate the hidden shelter and, over multiple sessions, they go from searching all around the platform to navigating directly to shelter using spatial memory. Work from our laboratory has provided three updates to this protocol (Vale et al., 2017; Vale et al., 2018): the learning period and escape testing all occur within a single session; the mouse is never removed the arena until the experiment is over; and sudden-onset threat stimuli are used to evoke robust, shelter-directed escape paths. Thanks to these innovations, our assay can capitalize on mice's ability to identify and memorize a shelter location within minutes (Vale et al., 2017) and their tendency to respond to threatening auditory or visual stimuli by immediately running directly to a shelter (Yilmaz and Meister, 2013). It is thus a promising model for studying how animals learn and execute complex trajectories to a goal within the time constraints compatible with survival in natural settings.

So far, rodent escape behavior has mostly been used to study goal-directed spatial navigation in the context of open, obstacle-free environments. These studies have revealed that mice have access to all the same navigational strategies as they do during reward learning, such as orienting themselves based on distant visual landmarks (Alyan and Jander, 1994; Harrison et al., 2006) or path integration (Maaswinkel and Whishaw, 1999; Vale et al., 2017). A natural next step is to use escape behavior to better understand how mice cope with structured environments where a one-step homing vector no longer suffices. Rodent escape in the presence of obstacles has been studied before, with gerbils: Ellard and Eller, 2009 showed that if the direct path to shelter is blocked on one side by a barrier, gerbils can use spatial memory to reach the hidden shelter after a brief period of exploration. Thus we know that rodent escape behavior offers not only reliable, stimulus-locked trajectories and rapid learning within a single session but also a reliance on sophisticated multi-step spatial reasoning. The cognitive mechanisms supporting rapid learning of multi-step escape routes, however, remain unknown.

1.4 Our questions

In a scenario with free exploration and no pre-training, how do animals learn to navigate a structured environment? Specifically, what happens when a lab mouse leaves its home cage for the first time and is exposed to an obstacle and a shelter? Will it be able to escape to the shelter despite the presence of an obstacle? If so, does it use a cognitive map strategy to infer novel routes without any practice, or will it rely on habitually reinforced practice routes? Or - most likely - will these classical frameworks break down upon adding an additional degree of naturalism to the experiment?

As we will see, the answer to this last question is yes: mice combine the action-reinforcement and map-build systems in a their strategy for learning multi-step routes past an obstacle. Thus, an additional question will be: what are the computational principles underlying this new hybrid strategy?

Chapter 2

Subgoal memorization

2.1 Mice quickly learn efficient escape routes

As a baseline condition for investigating how mice learn escape trajectories, we placed naïve animals in a circular, open-field platform with a shelter and overhead lighting. After a brief exploration period during which mice spontaneously located the shelter, we exposed them to a loud, overhead crashing sound while they were in a pre-defined threat zone (Fig. 2.1a). This reliably elicited rapid escapes directed at the shelter along a straight 'homing vector' (N=23 escapes, 10 mice; escape-to-shelter within 12 sec: 93%, compared to 12% in a no-stimulus control; Fig. 2.1a-b, Fig. 2.2a), similar to previous results (Vale et al., 2017).

We then repeated this experiment in a separate group of mice, with a wall positioned between the threat zone and the shelter (N=66 escapes, 24 mice; Fig. 2.3a). This wall was white against a black background, and all mice approached and walked along it during the exploration period (median time within 5 cm of the obstacle: 37 sec, IQR: 29 - 45 seconds; Fig. 2.2b). To quantify escape trajectories in relation to the obstacle, we computed a target score: escapes aimed at the shelter get a score of zero; escapes targeting the obstacle edge get 1.0; and escapes aimed beyond the obstacle edge get scores >1.0 (Fig. 2.1a). Escapes are classified as "edge vectors" if their score surpasses the 95th percentile of escape scores in the open field (0.65) and are otherwise classified as "homing vectors". Upon the first threat presentation, the majority of the mice (57%) executed homing-vector escapes (Fig. 2.1a-b; permutation test on the proportion of edge vectors, trial 1 vs. trials 2-3: p=0.0004 (***)). Replacing the wall obstacle with an unprotective hole obstacle did not reduce this proportion (N = 23 escapes, 8 mice; Fig. 2.2c-d, Fig. 2.3b); thus, homing-vector escapes cannot be accounted for by the safety provided by running along a wall and are likely directed at the shelter location.

Over the course of three threat presentation trials (17±4 minutes into the session,

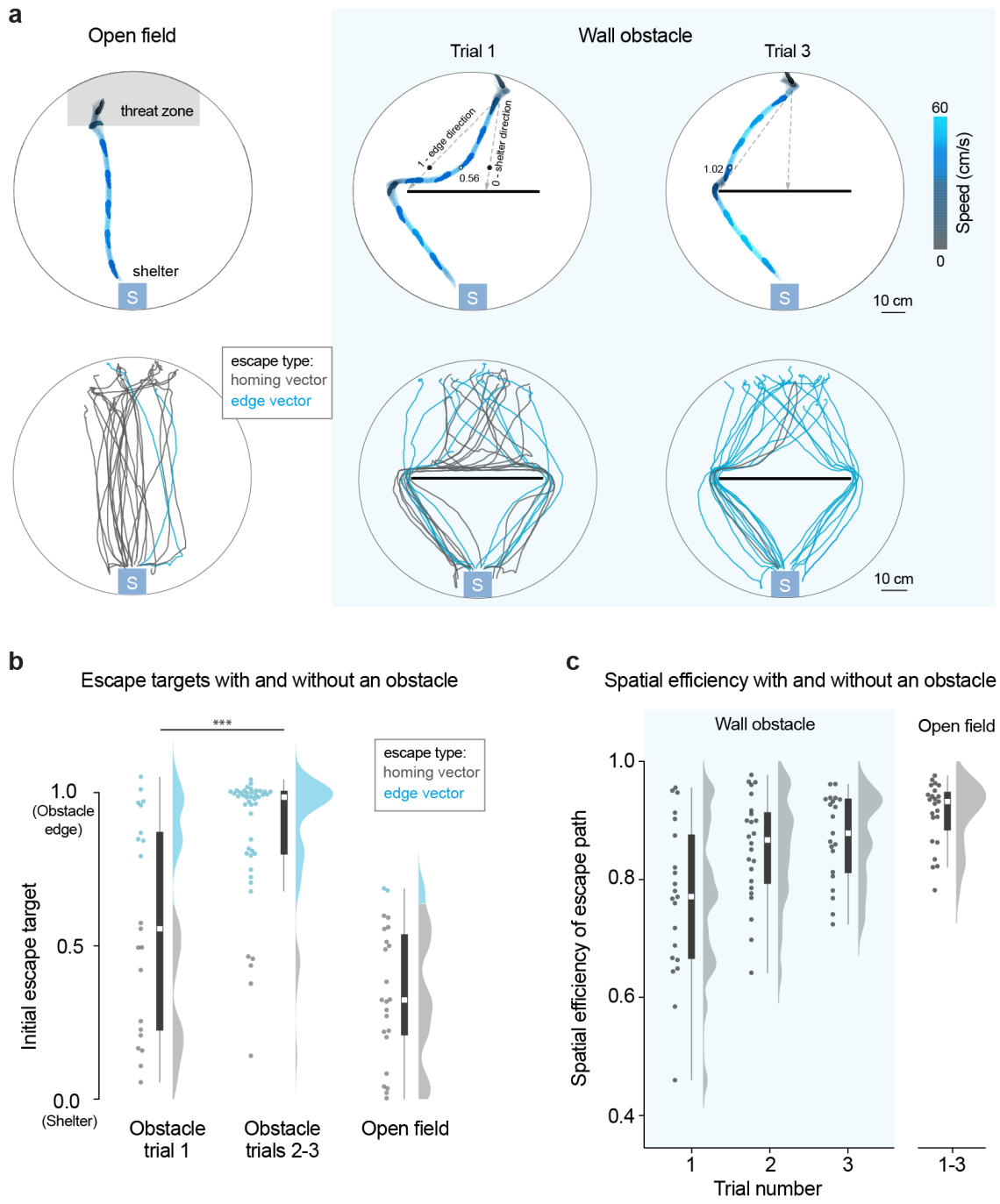

Figure 2.1: *Mice rapidly learn efficient escape trajectories past obstacles*

(a) Single escape trials colored by speed (top) and all trajectories color-coded by trajectory type. **(b)** Summary of initial escape targets. Each dot represents one escape. **(c)** Each dot represents one escape. White squares show the median, thick lines show the IQR, and thin lines show the range excluding outliers. Distributions are kernel density estimates.

mean±std), mice performed escapes that were increasingly spatially efficient (ratio of the shortest possible path to the actual escape path: median for trial 1 = 0.77;

for trial 3 = 0.87; F(2, 30)=7.2, p=.003, repeated measures ANOVA on trials 1-3; Fig. 2.1b). By this point, almost all trajectories were aimed directly at the obstacle edge (90% edge vectors; median target score = 0.98). Thus, while inefficient homing responses initially dominated, mice acquired rapid and streamlined routes to shelter over the course of 20 minutes and three escape trials.

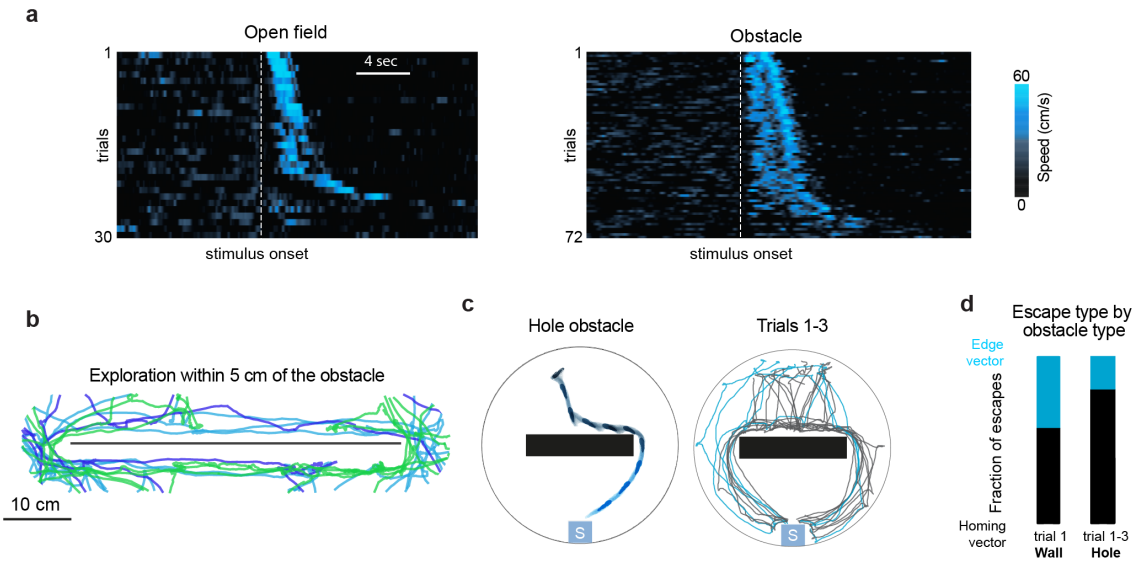

Figure 2.2: *Escapes in the presence of an obstacle: extended results*

(a) Speed profile for all threat stimulation trials escape trials (3 trials per mouse). Trials are sorted by shelter arrival time. **(b)** Exploration trajectories around the obstacle area (each color represents the movements of one mouse prior to trial 1). 3 randomly selected sessions are displayed. All mice approached and explored the area within 5 cm of the obstacle prior to the first escape trial. **(c)** Example trial and escape trajectories for experiment with an unprotective hole obstacle instead of the wall obstacle. The black rectangle represents the hole obstacle. **(d)** Summary of escape trajectories with the wall and hole obstacles.

2.2 A spatial memory strategy for navigating the obstacle

Visual cues not necessary to navigate past an obstacle

We next investigated whether mice require visual input to locate and run toward the obstacle edge. We repeated the obstacle experiment from Figure 1, but now in complete darkness (Fig. 2.4b-d). Mice now executed fewer edge-vector escapes (% edge-vector escapes on trials 1-3: 33% with the lights off vs. 74% with the lights on, p=0.002, permutation test; N = 33 escapes, 14 mice with the obstacle in the dark).

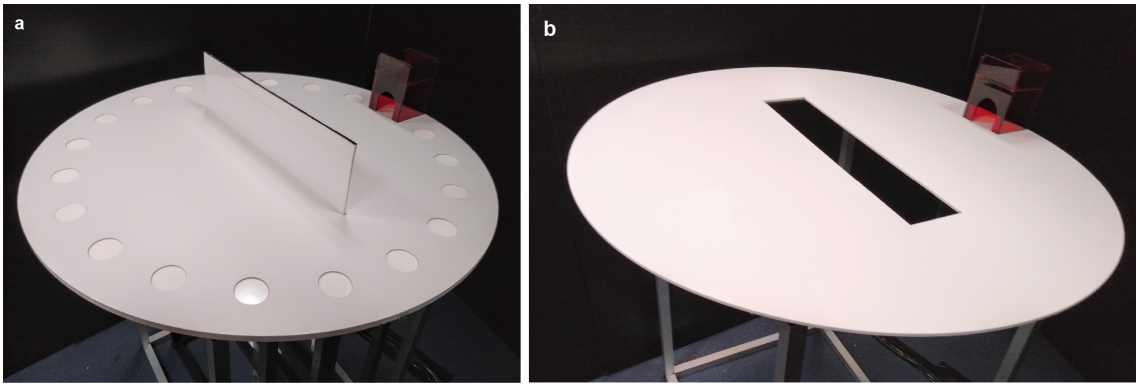

Figure 2.3: *Platforms with the wall obstacle and hole obstacle for ch. 2*

(a) Platform with the wall obstacle. The platform is 92 cm in diameter, and the wall obstacle is 50 cm long x 12.5 cm tall. The shelter is made from red acrylic that is opaque to the mouse but transparent to (infra)red light. **(b)** Platform with the hole obstacle. The hole obstacle is 50 cm long x 10 cm wide.

This proportion of edge vectors was not significantly different than chance, i.e. the open field condition (p=0.2, comparison with the 22% edge-vector escapes in the dark without an obstacle; N = 41 escapes, 14 mice). However, after 20 minutes with three escape trials in the light, mice were able to execute mostly edge vectors in the dark (55% edge-vector escapes vs. 22% in the open field, p=0.002, permutation test; Fig. 2.4c-d; N = 33 escapes, 14 mice). Thus, for naïve mice with limited experience, visual cues are required for efficient obstacle avoidance. However, immediately after experiencing a 20-minute behavioral session, streamlined escapes can occur even in complete darkness.

The obstacle removal experiment

We thus considered that learning efficient escapes might entail developing a memory of the obstacle edge location, making perception of the obstacle unnecessary. To further test this hypothesis, after the animals explored the environment with the obstacle for 20 minutes and with three escape trials, we removed the obstacle at the moment of threat onset ("acute obstacle removal"). Although the obstacle disappeared before the initial orientation movement could be completed, all animals escaped along the edge vector and did not turn toward the shelter until they passed the location where the obstacle edge used to be (median target = 0.98; N=8 escapes, 8 mice; edge-vector proportion compared to open field: $p=1 \times 10^{-5}$, permutation test; Fig. 2.5a-b). Next, we examined how persistent this memory-based strategy is. In a "chronic obstacle removal" experiment (CORE), we allowed mice to explore after this acute obstacle removal trial (for 9±5 minutes, mean±std), during which

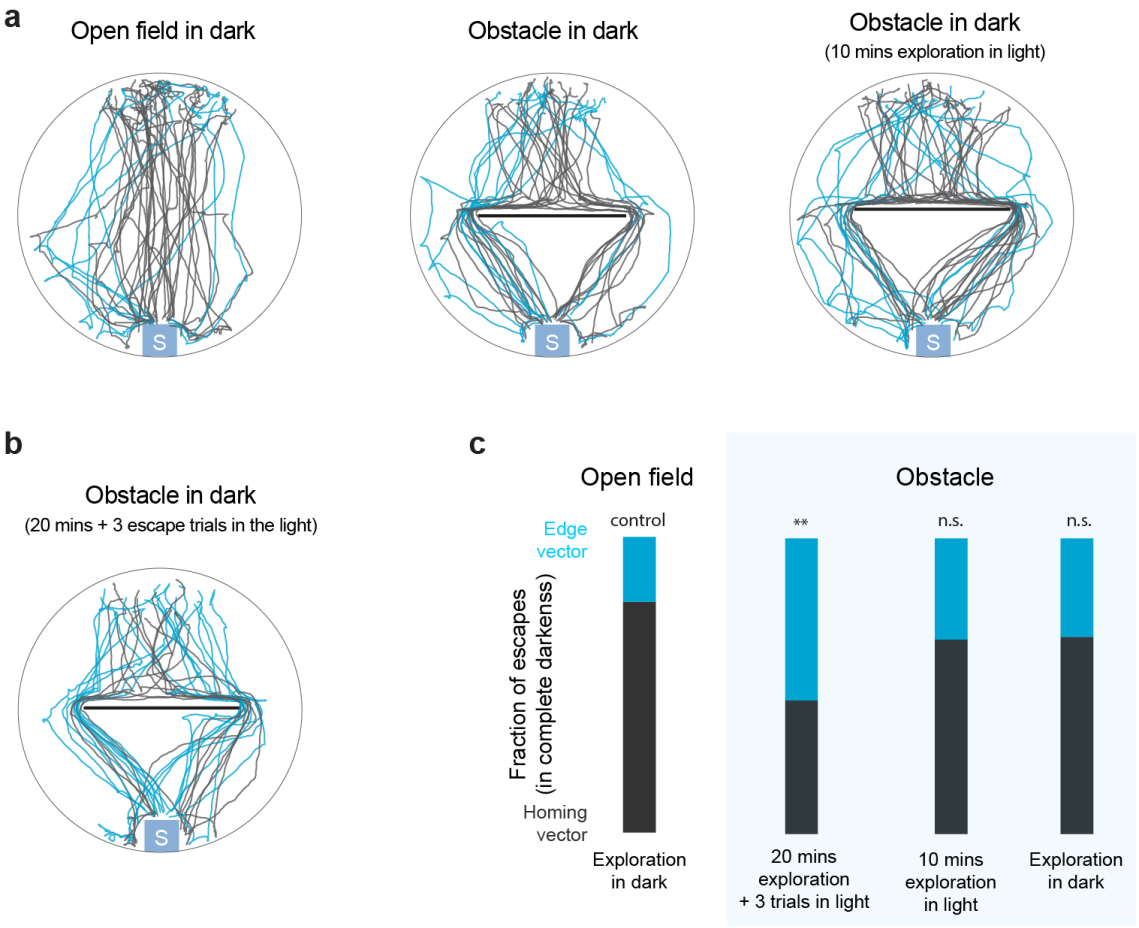

Figure 2.4: *Escapes with an obstacle in complete darkness*

(a) Escape trajectories for experiments in which naïve mice escape to the shelter in complete darkness. **(b)** Mice with 20 minutes of experience in the light, including three escape trials. **(c)** Summary of escape trajectories in complete darkness.

time 100% of mice visited the now empty center of the platform (Fig. 2.5c). 44% of the subsequent escapes were still directed at the location where the obstacle edge used to be (N=18 escapes, 8 mice; more than the 9% edge-vector rate in the open field: p=0.02, permutation test; Fig. 2.5a-b), while the remaining 56% mice reverted to the homing-vector response.

This spatial memory for edge-vector escapes could in principle be learned during escapes trials or through spontaneous exploratory behavior. To distinguish between these possibilities, we repeated the CORE with zero baseline threat stimuli during the 20-minute exploration period (CORE-ZB). As in the previous experiment, we then removed the obstacle and allowed the mice to explore the newly unobstructed environment (for 5±4 minutes, mean±std). Threat presentation after this period resulted in mostly edge-vector responses (57% edge-vector escapes; N = 23 escapes, 10 mice; more edge vectors than in the open field: p=0.004, permutation test;

Fig. 2.5a-b). Thus, within 20 minutes in a novel environment, mice spontaneously develop a persistent spatial memory for efficient, multi-step escapes. The rest of the thesis will focus on unraveling the behavioral and computational mechanisms of this memory in the CORE-ZB.

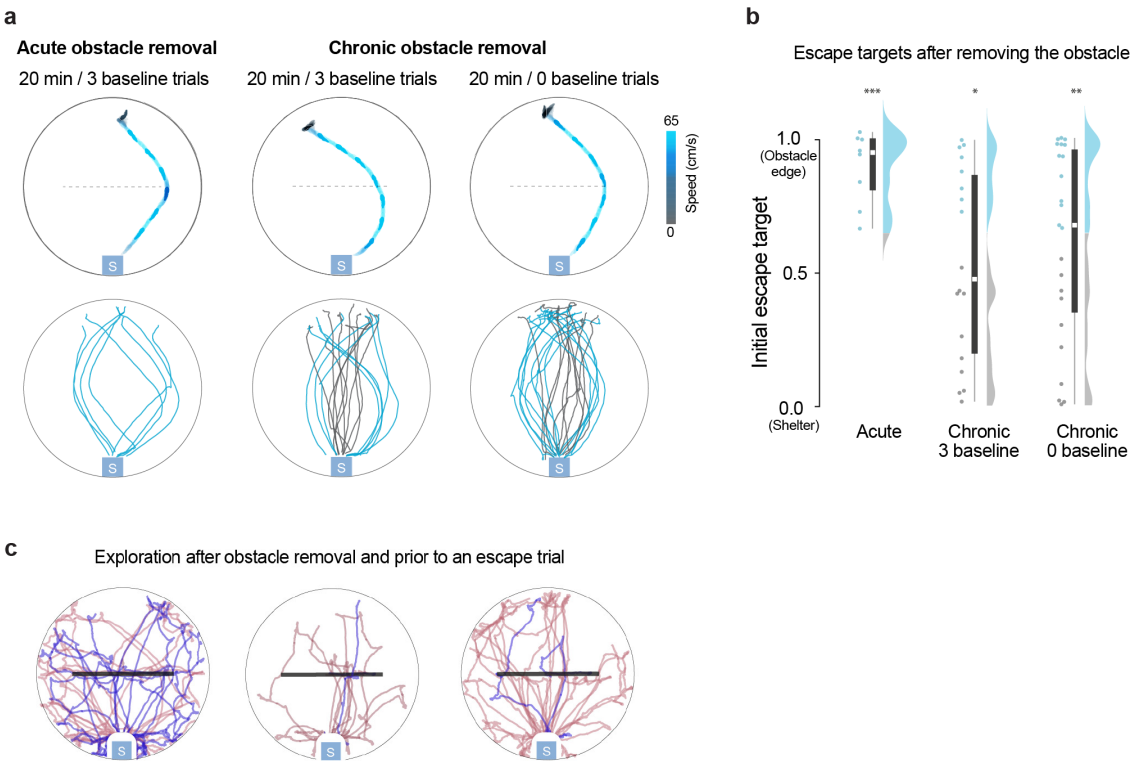

Figure 2.5: *Mice escape around obstacles even after they are removed*

(a) Examples of edge-vector escape trials (top) and all trajectories (bottom) after removing the obstacle. Subheaders: experience in the environment prior to removing the obstacle. Dotted line: where the obstacle used to be. **(b)** Summary data for initial escape targets. **(c)** Black bar: area where the obstacle used to be (black bar). Each colored trace represents the movements of one mouse after the obstacle was removed and prior to an escape trial. Six mice were randomly selected for visualization.

2.3 Spatial learning here consists of memorizing subgoal locations

We first aimed to characterize this spatial-memory strategy and how it is learned. We evaluated three possible strategies: habitual learning of turn angles, sampling the environment to build a cognitive map, and memorizing subgoals encountered during practice homings. We evaluated each possibility by analyzing the relationship

between escapes and spontaneous behavior during exploration, primarily in the chronic obstacle removal experiment with zero baseline trials (CORE-ZB). For each analytical finding, we then performed further experiments to validate the analysis.

Habitual, egocentric movements do not explain the spatial memory

First, we tested whether mice learn egocentric movements from the threat zone to the obstacle edge, similar to the habitual response strategy in mazes (Restle, 1957). We extracted all spontaneous homing runs, or 'practice runs,' defined as sustained turn-and-run movements from the threat area toward the shelter during the CORE-ZB's exploration period (median [IQR] number of runs = 7 [6,7]; time from their end point until reaching the shelter: 11 [5, 19] sec; Fig. 2.6a; Fig. 2.7a). We then computed each run's starting position and orientation, and the angle turned during its initial turn-and-run segment (difference in heading direction from the homing initiation point to the point where the mouse has travelled 15 cm; Fig. 2.6a).

Homing runs were sparse, and their initial positions and body orientations were highly variable. It was unlikely for any escape's starting conditions to closely match a previous homing: only 22% of escapes were preceded by a run with starting points within 10 cm distance and 30°body orientation (Fig. 2.7a). Despite this lack of stereotypy, we attempted to account for the memory-guided edge-vector escapes observed in the CORE-ZB using the assumption that mice repeat turn angles from previous homing runs. First, we validated a method to predict escape targets based on homing-run turn angles. We put mice on a modified platform with two narrow corridors, ensuring that homings and escapes were stereotyped (N=30 escapes, 10 mice; Fig. 2.7b). Here, we could precisely predict escape targets using the mouse's starting point and its history of previous turn-and-run movements (R^2 of the prediction = 0.97 using the homing run with the most similar turn angle; R^2 = 0.65 using the homing run with the closest initial position; R^2 = 0.58 using the homing run with the closest initial body orientation; Fig. 2.6b; Fig. 2.7c). In the CORE-ZB, however, repeating turning movements did not explain any of the variance in post-removal escape targets (R^2 of the prediction = 9×10^{-4} (most similar turn angle); R^2 = 0.04 (closest initial position); $R^2 = 9 \times 10^{-6}$ (closest initial body orientation); compared to R^2 = 0.05 for a randomly generated prediction, averaged over 1000 random seeds; Fig. 2.6b; Fig. 2.7d).

This analysis suggests that memory-guided edge-vector escapes are not based on habitually repeating egocentric actions. We performed an additional experiment to test this finding: testing whether memory-guided escapes are sensitive to the

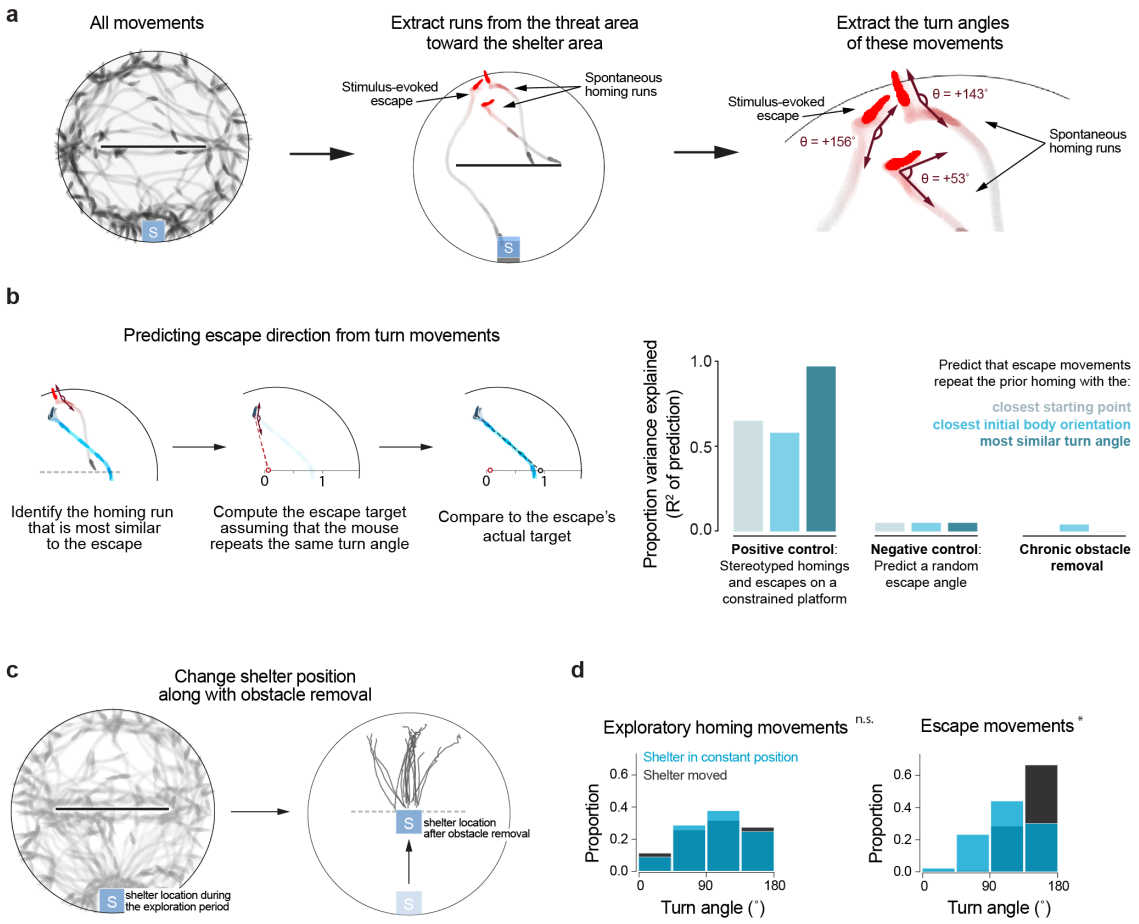

Figure 2.6: *Habitual, egocentric movements do not explain the spatial memory*

(a) Examples of homing practice runs extracted from all movement data. Turn angles are defined positive for rightward turns and negative for leftward turns. **(b)** Left: method for predicting the escape target based on the turn angles from previous homing runs. Right: the y-axis measures the amount of variance in escape targets across trials that can be explained by mice repeating turn angles from previous homing runs. **(c)** An experiment like the CORE-ZB, but in which the shelter is moved following the exploration period. **(d)** Distribution of turn angles from homing movements and from escapes. Black shows the new CORE with the moved shelter, and blue shows the two original COREs.

shelter location. This would not be expected from a habitually repeated action (Fig. 2.6c-d; N=18 escapes, 10 mice). After a 20-minute exploration period just like in the CORE-ZB, we moved the shelter to the middle of the platform. As expected from a goal-directed or geometry-dependent process, escape turns differed from the two original COREs, with zero escapes targeting the obstacle edge location (p=0.5, chi-square test on binned homing-run turn angles; p=0.03, chi-square test on escape turn angles; Fig. 2.6d). This experiment demonstrates a dissociation between egocentric turning movements and memory-guided escapes: identical exploratory

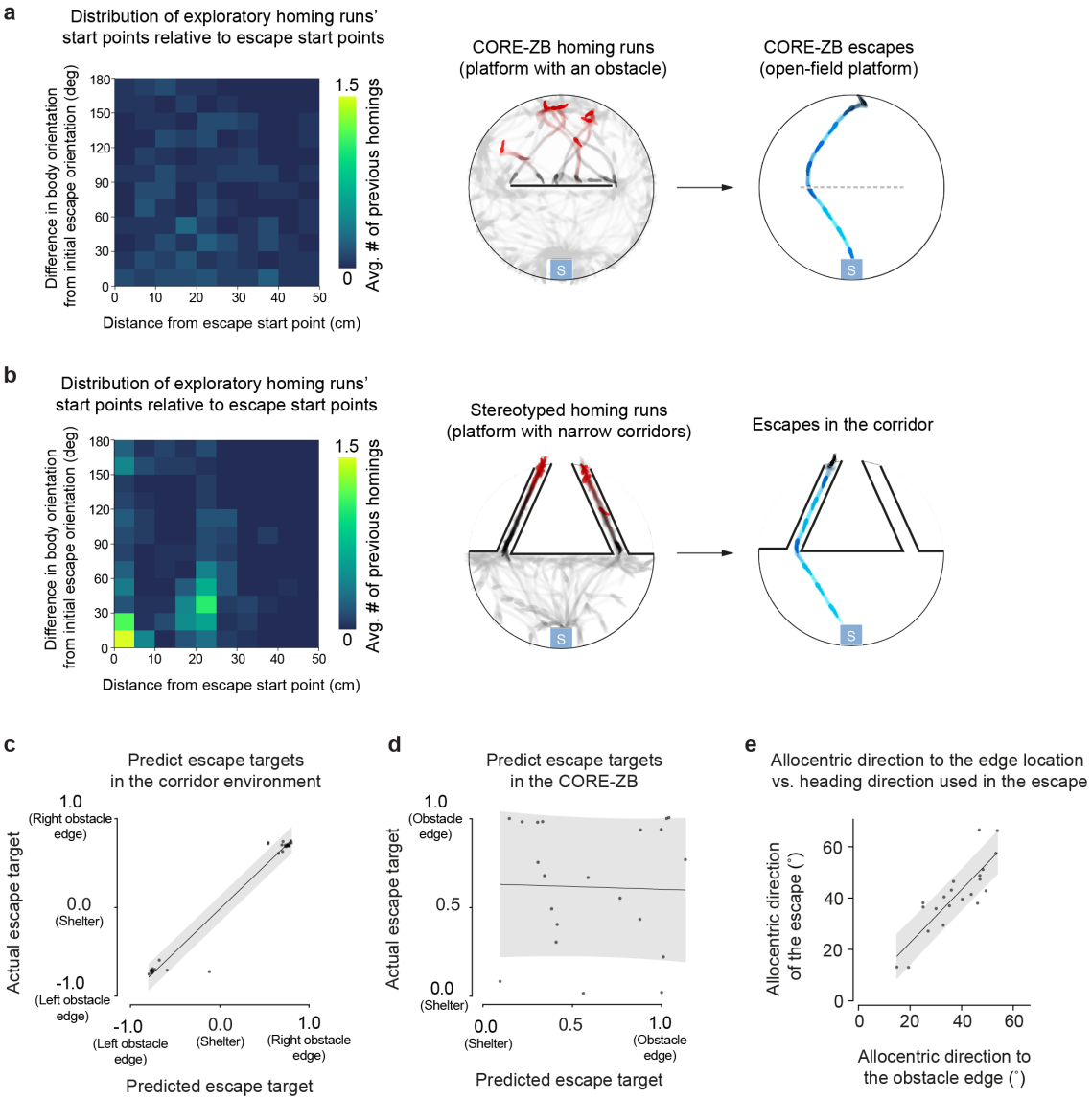

Figure 2.7: *Homing runs, turn angles, and heading directions: extended results*

(a) Left: histogram of each practice homing runs' initial condition. Each bin reflects proximity in both the position (x) and body orientation (y) of the homing's starting point. Right: example of homing runs extracted from exploration and a subsequent escape. **(b)** Same as panel a, but for the experiment with narrow corridors. **(c)** Escape targets are predicted using the homing run with the most similar turn angle to the escape turn angle. Correlation coefficient r=0.98; p=2x10-22. **(d)** Same analysis as in panel c, but with escapes and homings from the CORE-ZB. Correlation coefficient r= -0.03; p=0.9. **(e)** The escape direction (y) is measured when the mouse is 15 cm away from the escape initiation point. The vector from the center of the platform to the shelter (pointing south) is set as 0°. Homing-vector escapes are not included. Lines show the linear regression fit. Shaded area shows the prediction interval within 1 standard deviation.

movements can lead to distinct escape movements.

Observing the obstacle does not explain the spatial memory

The goal-directed nature of these escapes suggests that the obstacle edges become subgoal locations, i.e. allocentric locations targeted as a waypoint en route to the ultimate goal. An alternative possibility, however, is that mice target the edge by learning allocentric heading directions. For example, edge-vector escapes could be generated by consistently running in the southwest or southeast direction, relative to the north-south axis connecting the shelter and the threat zone. Analysis of our data indicates that mice instead target allocentric locations. Following obstacle removal, escape heading directions follow whichever direction is required to reach the edge location (correlation between the heading direction to the edge and the heading direction taken in the escape: r=0.85, p=1.2x10-6; Fig. 2.7e). This corroborates the results above, suggesting that mice learn subgoal *locations* at the obstacle edge.

We next investigated the learning process that generates these subgoals during the spontaneous exploration period. We found two variables in the CORE-ZB with high, positive correlations to subgoal-targeting behavior: the total distance of exploratory movement on the threat side of the platform (correlation with escape targets: r=0.72, p=1×10^{-4}; Fig. 2.9a) and the number of homing runs from the threat area that directly targeted the obstacle edge (within 10 cm; correlation with escape targets: r=0.75, p=5x10-5; Fig. 2.8a-c). Two primary interpretations of these correlations are possible. The first is that routes are computed directly from a 'cognitive map': investigating the obstructed area updates the mouse's internal map, which is reflected behaviorally in the mouse's use of subgoals. If this were true, we would predict that: 1) investigating relevant features like the obstacle or its edge will also correlate with the subgoal memory; and 2) after obstacle removal, investigating the region where the obstacle used to be will suppress edge-vector escapes. Neither prediction matched the data. The amount of exploration near the obstacle or the obstacle edge was not correlated to subsequent escape target scores (correlation of escape targets with distance moved around the obstacle: r= -0.09, p=0.7; around the obstacle edge: r=0.06, p=0.8; Fig. 2.9a). Furthermore, after obstacle removal, mice that densely sampled the empty center of the arena more did not execute different escape trajectories from mice that explored very little (correlation with distance moved around where the obstacle used to be: r= -0.12, p=0.6; with total post-removal exploration distance: r= -0.17, p=0.4; Fig. 2.9a).

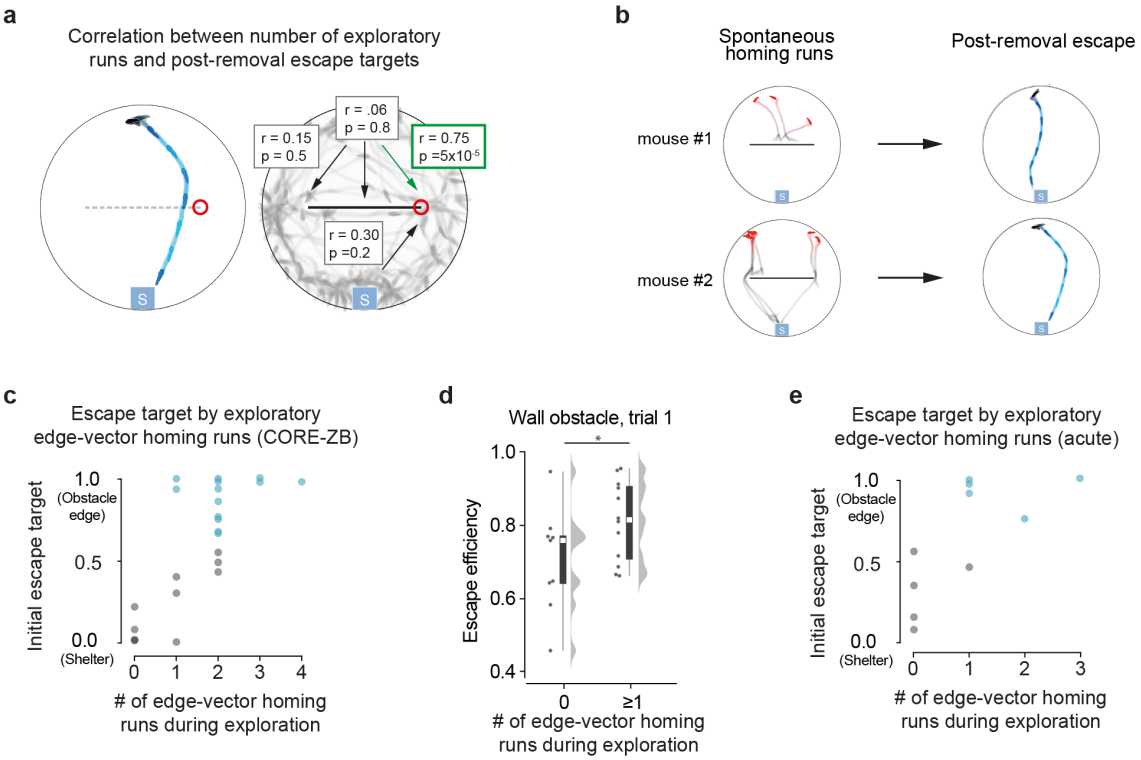

Figure 2.8: *Mice memorize previously targeted subgoal locations*

(a) Correlation of different running movements with the escape target score, in the CORE-ZB. Movements toward the same edge targeted in the escape are shown here as the right edge; movements toward the opposite edge are shown as the left. Significant correlations have green outlines. **(b)** Homing run history for two mice in the CORE-ZB, and subsequent escape trajectories. **(c)** Escape targets plotted against the number of spontaneous edge-vector homing runs during exploration in the CORE-ZB (movements toward the same edge targeted during the escape). **(d)** Spatial efficiency of escapes on the first trial in the presence of an obstacle. **(e)** Escape targets plotted against the number of spontaneous edge-vector homing runs during exploration, for acute obstacle removal on the first trial.

Practice runs can explain the spatial memory

A second possibility is that learning occurs during the 'practice' edge-vector homing runs. In this case, we would predict that: 1) subgoals do not form in mice with zero edge-vector homings; and 2) the correlation with spontaneous homing runs would be specific to the edge targeted during escape (i.e., left vs. right) and to the direction taken during escapes (i.e., from the threat side to the shelter side). Both predictions were confirmed by the data. Every edge-vector escape following obstacle removal was preceded by at least one homing run targeting that same edge (100% of post-removal edge-vector escapes have $\geq$1 prior edge-vector run; greater than chance: p=0.02, permutation test; Fig. 2.8c). Second, escape targets in the CORE-ZB were not significantly correlated with homing runs from the threat area to the opposite

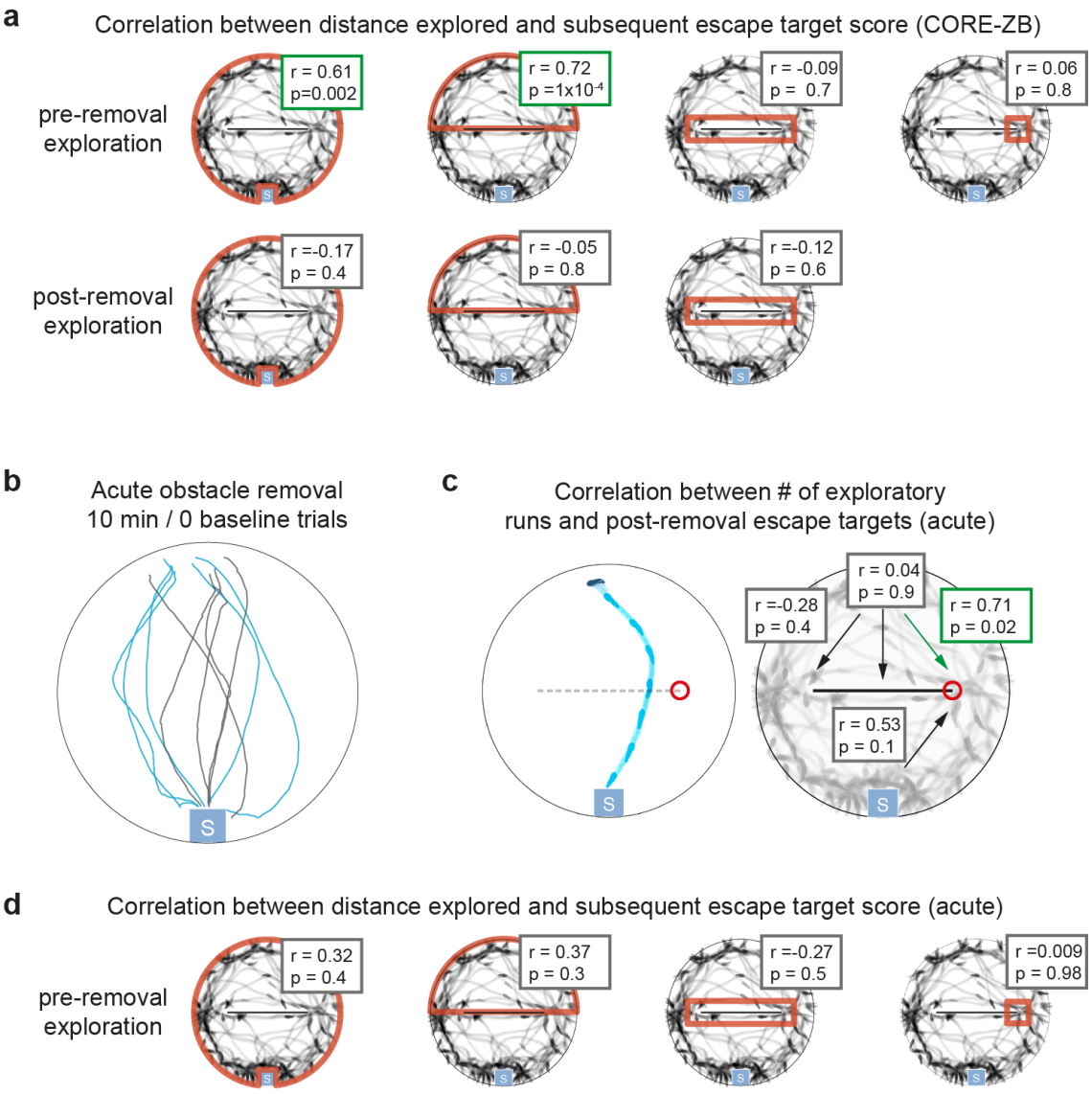

Figure 2.9: *Mice memorize previously targeted subgoal locations: extended results*

(a) Correlation between escape targets in the CORE-ZB and the amount of exploration in different sections of the platform. Red outlines indicate the section of the platform in which the distance explored is measured. For exploration near the obstacle edge, only the edge that was targeted during the escape (i.e., left vs. right) is considered. Boxes show the correlation coefficients and respective p-values; significant correlations have green outlines. **(b)** Escapes from an experiment acutely removing the obstacle on the first trial, after 10 minutes of exploration. **(c)** Correlation of different running movements to escape target score, in the trial-1 acute removal experiment. **(d)** Correlation between escape targets in the trial-1 acute removal experiment and the amount of exploration in different sections of the platform. Note that there is no post-removal exploration in this experiment.

edge (r=0.15, p=0.5), with homing-vector runs from the threat area to the middle of the obstacle (r=0.06, p=0.8), or with runs from the shelter area to the same obstacle

edge (r=0.30, p=0.2; Fig. 2.8a).

Our analysis of the CORE-ZB suggests that executing edge-vector homings, rather than sampling the environment, could be the rate-limiting step in spontaneously learning subgoals. To further test this hypothesis, we first examined whether spontaneous homings explain escape routes in the obstacle-present condition. On the first trial with an obstacle, mice with prior edge-vector homings performed more efficient escapes than mice with none (median spatial efficiency with zero runs = 0.76; with one run = 0.82; p=0.04, permutation test; Fig. 2.8d; same data from Figure 1). As expected, this significant effect did not also occur for runs to the obstacle edge not used in the escape, runs from the shelter to the obstacle edge, or runs toward the center of the obstacle. Thus, subgoal memorization does appear to play an adaptive role when perception of the obstacle is still available.

Next, we examined the acute obstacle removal experiment. We could not apply correlational analysis to the previous acute obstacle removal since this dataset had 100% edge-vector responses and 100% prior edge-vector homings. Thus, we performed a new experiment, removing the obstacle acutely on the first trial, 10±1 minutes into the session, mean±std. Here, 50% of escapes took edge-vector paths (N = 10 escapes, 10 mice; Fig. 2.9e). Among the variables examined – exploration in different parts of the platform and various running movements – only the number of runs from the threat area to the edge used in the escape was significantly correlated with escape targets (r=0.71, p=0.02; Fig. 2.8e; Fig. 2.9f-g). Furthermore, 100% of edge-vector escapes were preceded by at least one edge-vector homing (greater than chance: p=0.02, permutation test; Fig. 2.8e).

Next, we tested the practice-homing hypothesis with two new experiments. First, we repeated the CORE-ZB but without a shelter during the exploration period (N=24 escapes, 10 mice; Fig. 2.10a). This gives the mouse opportunity to observe the platform and obstacle, but without performing homings. After 20 minutes, we added the shelter and removed the obstacle as soon as the mouse entered the shelter (median [IQR] time to enter shelter: 84 [39, 154] sec). Subsequent escapes did not exhibit the subgoal memory (13% edge-vector escapes; not more edge vectors than in the open field: p=0.4, permutation test; Fig. 2.10a). Second, we repeated the CORE-ZB with an extra barrier blocking off the threat side of the platform during the exploration period (N=25 escapes, 10 mice; Fig. 2.10b). This prevents long-range homings while allowing investigation of the obstacle. Only 1/10 mice targeted the edge location with scores close to 1.0, and post-removal escapes did not significantly differ from the open-field control (20% edge-vector escapes; not more edge vectors than in the open field: p=0.2, permutation test; Fig. 2.10a-c). Both experiments thus demonstrate a dissociation between investigating the obstacle and memorizing

subgoals, and further support the hypothesis that subgoal locations are learned through practicing homings.

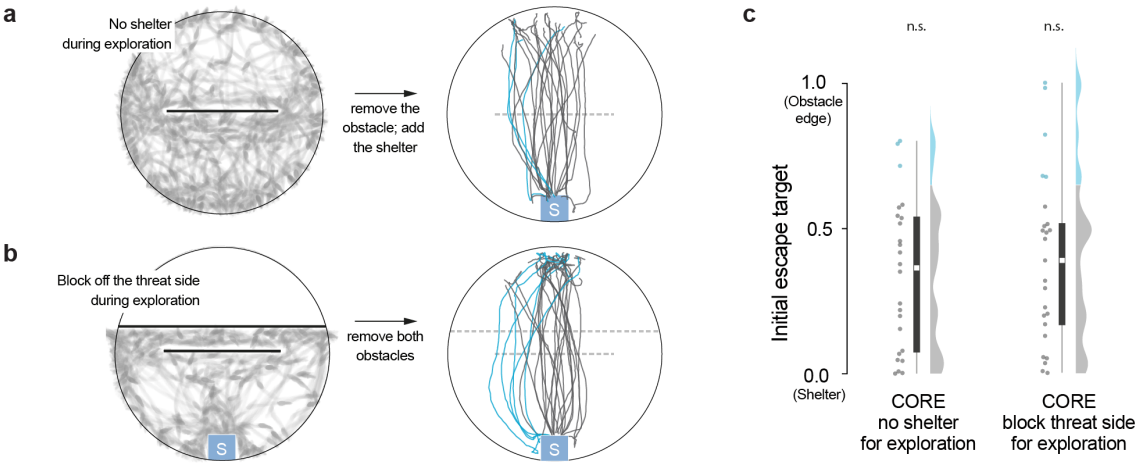

Figure 2.10: *Mice memorize previously targeted subgoal locations*

(a) Chronic obstacle removal experiment without a shelter present during the 20-min exploration period, and all subsequent escapes. **(b)** Chronic obstacle removal experiment with an extra barrier blocking the threat area during the 20-min exploration period, and all subsequent escapes. **(c)** Summary of escape targets in the two modified CORE-ZB experiments.

Edge-vector runs are instinctive exploratory actions

It remains unclear what prompts spontaneous edge-vector runs in the first place. One possibility is that during practice runs, a cognitive map is used to compute efficient routes to shelter; once this happens, subgoals are tagged for later use during escapes. Another possibility is that mice are innately predisposed to run to salient obstacle edges. Our data support the latter option. Spontaneous edge-directed movement occurs most during the first few minutes of the session and occur equally with or without a shelter in the environment (Fig. 2.11a-b). When the obstacle is a hole instead of a wall (Fig. 2.3), edge-directed movement occurs with the same, low frequency as in the open field (computed using the location where the obstacle edges would be if the obstacle were present; Fig. 2.11b). Correspondingly, it takes twice as long for mice to perform predominantly edge-vector escapes in the presence of a hole obstacle (20% edge-vectors escapes on trial 2-3, 67% edge-vector escape on trial 6-7; Fig. 2.2d).

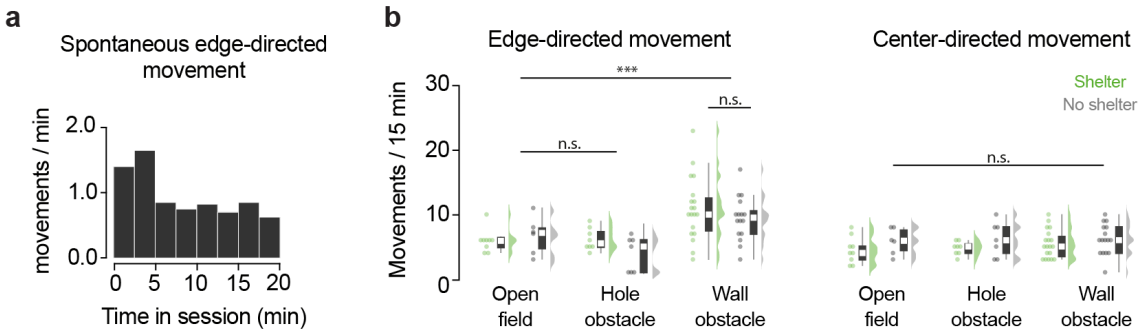

Figure 2.11: *Edge-directed movements in different environments*

(a) Frequency of spontaneous movements toward the obstacle edges by time in session (all sessions with the wall obstacle) . **(b)** Frequency of edge-directed movements for different conditions. Statistical test: permutation test on number of edge-directed movements. Each dot is one session. White squares show the median, thick lines show the IQR, and thin lines show the range excluding outliers.

2.4 Subgoal memorization also supports food-seeking routes

While subgoal memorization enhances spatial efficiency in a static environment, it can also generate unnecessarily roundabout routes past an obstacle that no longer exists. In fact, edge-vector escapes can persist over at least 20 minutes and 7 trials following obstacle removal (Fig. 2.11a-b). We considered that subgoal memorization may be specific to escape behavior, as mice might sacrifice flexibility for the sake of quickly reacting to imminent threats. To test this, we performed an obstacle removal experiment in the context of a less urgent, reward-based task (open field control: N=32 reward runs, 6 mice; obstacle removal: N=34 reward runs, 6 mice).

First, we trained food-deprived mice to approach and lick a reward port in response to a 10-kHz tone, which indicated the availability of condensed milk at the port. This took place across 5 sessions, in an operant conditioning box (Fig. 2.13a-c). Next, we transported this task to the platforms previously used for escape behavior. The shelter was replaced by the reward port, and the threat stimulus was replaced by the 10-kHz tone. To start the session, mice were given 20 minutes in the open-field or obstructed environment. This included 1 food-approach trial per minute with start points throughout the platform, to facilitate transferring the task to this new environment. After this point, mice successfully ran to the reward port during tone presentation on 85% of trials starting in the trigger zone (the same region as the threat zone), but with slower reaction times than escape (median [IQR] time to start running toward the goal for food-seeking = 1.5 [0.7, 3.5] sec; for threat response = 0.6 [0.4, 1.2] sec; p=0.005, permutation test).

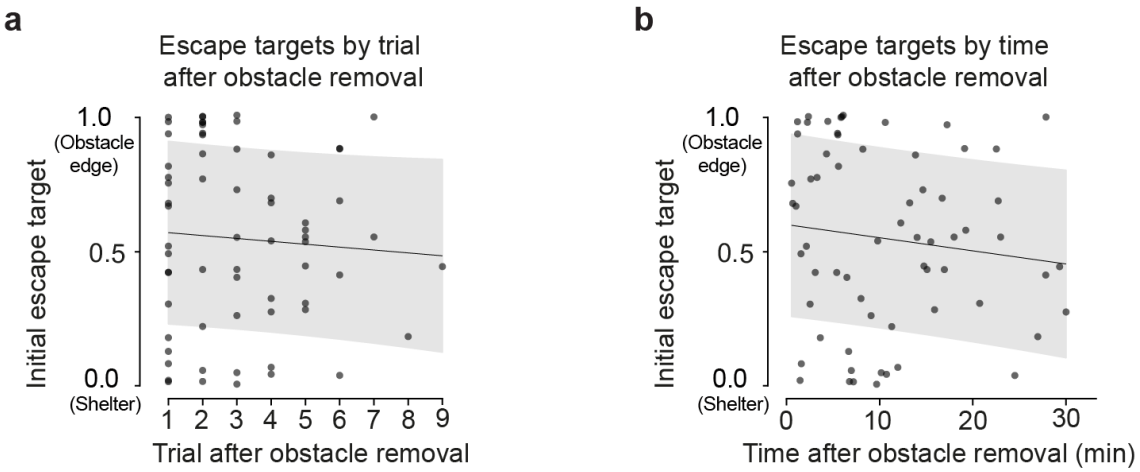

Figure 2.12: *Edge vectors persist long after obstacle removal*

(a) Escape targets vs. trial number in the chronic obstacle removal experiments (now including the minority of mice that performed >3 trials). Only successful escape trials are counted toward the trial number. Correlation coefficient r=-0.07, p=0.6. **(b)** Escape targets vs. time. Correlation coefficient r=-0.12, p=0.3. Lines show the linear regression fit, and the shaded area shows the prediction interval within 1 standard deviation.

At this point, we repeated the CORE with the food-seeking behavior. We removed the obstacle and triggered food-approach trials (trials occurred 5±3 minutes after obstacle removal, mean±std). Similar to escape routes, a large proportion of paths in the obstacle-removal condition initially targeted the obstacle edge location (53% edge vectors; p=0.006 compared to 12% in the open field, permutation test; Fig. 2.14a-b).

Finally, we tested whether experience with the obstacle induces a non-specific increase in edge-directed movement, as this could explain the apparent use of subgoal memorization across two distinct tasks. We compared spontaneous movements from the ends of the platform toward the center and obstacle edge locations. Exploration following obstacle removal were not enriched in edge-directed movements (number of edge-directed movements per 15 min: median after obstacle removal = 4; in the open field = 6; with the obstacle present =12; p=4x10-5, permutation test on open field vs. obstacle; p=0.7, permutation test on open field vs. obstacle removed; Fig. 2.14c-d). Subgoal memorization therefore reflects a strategy for goal-directed navigation rather than a general bias in how mice move around their environment following experience with an obstacle.

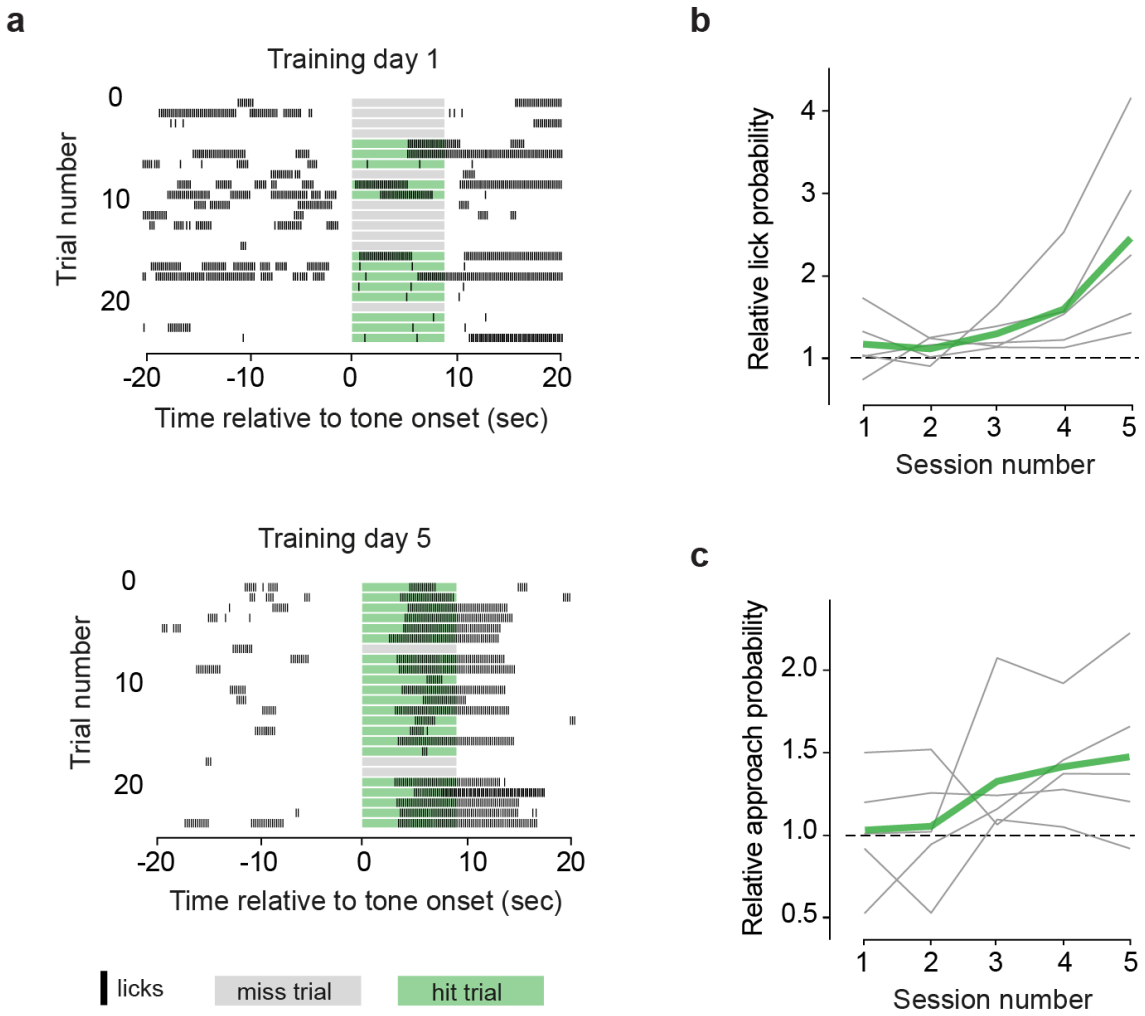

Figure 2.13: *Training mice to approach and lick a spout in response to a tone*

(a) Lick raster plots for an example mouse during the first (top) and the last training day (bottom). **(b)** Summary data for lick probability during training. Relative lick probability is the average probability of licking the spout within a 4.5-second window during the stimulus, divided by the lick probability during the 20 seconds before or after the stimulus. Relative lick probability on day 5 is > 1 (p=0.002, permutation test) but not day 1 (p=0.09). **(c)** Summary data for reward-port approach probability during training. Relative approach probability is the average probability of moving from the back of the conditioning box to the side where the spout is located in response to the tone, divided by the approach probability at other random time points during the session. Relative approach probability on day 5 is $>$ 1 (p=0.02) but not on day 1 (p=0.41). Gray lines are individual mice and green line is the mean.

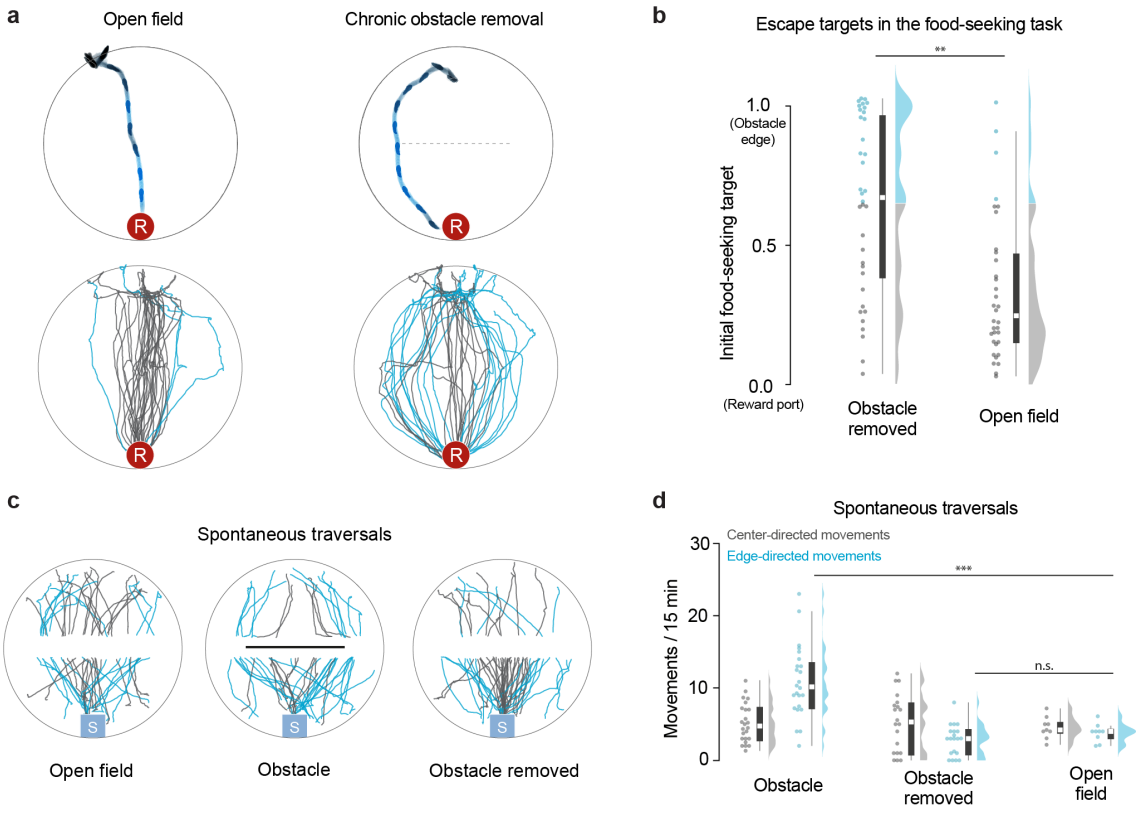

Figure 2.14: *Obstacle experience affects food-seeking but not exploratory paths*

(a) Food-approach paths. An example trial is shown on top, and all paths are shown below. The red circle with 'R' represents the reward location, i.e. the metal spout with milk. **(b)** Summary data for food-seeking paths, computed the same way as escape targets. **(c)** Paths across the platform during spontaneous exploration in the escape experiments. All paths go from the ends of the platform toward the center. Conditions with more sessions are randomly downsampled so that the same number of paths is displayed for each condition. **(d)** The number of spontaneous center-directed and edge-directed movements during exploration.

2.5 Interim discussion

During their first few minutes in an obstructed environment, mice escaped to shelter by relying on their memory of the shelter location and their innate ability to negotiate barriers using vision and touch. These escape routes were spatially inefficient; they resembled obstacle avoidance in animals with lower cognitive capacities, such as toads, crabs, and ant colonies (Collett, 1982; Layne et al., 2003; McCreery et al., 2016). Over a single 20-minute session, however, mice began to exploit their aptitude for spatial memory. They increasingly targeted the obstacle edge directly and could do so even in complete darkness or after the obstacle had been removed. We found that this capacity relied on memorizing allocentric subgoal locations rather than egocentric turning movements, and our data further suggested that mice identified and memorized subgoals during spontaneous homing runs.

Previous work has shown that rodents use spatial memory to navigate to shelter in an open field (Etienne et al., 1985; Alyan and Jander, 1994; Vale et al., 2017). In such a simple environment, however, escape routes can be implemented by path integrating self-motion cues to keep track of a single vector to the shelter location - a one-step, egocentric process. With obstacles in the environment, a more advanced strategy is needed. Previous results in gerbils escaping in an obstructed environment suggested that spatial memory was employed to reach the shelter (Ellard and Eller, 2009), but their navigational strategy was unknown. Our results show that mice use subgoals in an allocentric reference frame. Several observations support this view. First, mice can accurately target the edge location minutes after the obstacle or the lights have been removed, which is not well explained by pure path integration. Second, escapes involved immediately orienting and running toward a subgoal 50 cm away, which is not consistent with following odor trails or gradients (Wallace et al., 2002; Liu et al., 2020). Finally, repeating stereotyped turning movements or allocentric heading directions did not explain memory-guided escape paths in our assay; instead, mice consistently targeted the edge location.

Traditional models of allocentric navigation involve three key elements: an internal map of the environment (located in the hippocampus and entorhinal cortex), a stored goal location, and a mental search for paths to the goal (Burgess et al., 1994; Spiers and Gilbert, 2015; Stachenfeld et al., 2017; Edvardsen et al., 2020). The limiting factor is the quality of the map. Finding efficient multi-step routes - be it through a tree-search algorithm (Spiers and Gilbert, 2015; Edvardsen et al., 2020), a map-partitioning algorithm (Stachenfeld et al., 2017), or warping around an 'obstacle-to-avoid' feature (Burgess et al., 1994) - can occur as soon as the map faithfully reflects the current environment. To build up this map, animals simply

have to investigate unfamiliar or altered parts of the environment. The amount of exploratory movement thus matters for spatial learning, but movements' intentions or targets do not (Cf. Schölkopf and Mallot, 1995). Our observations of escape routes in naïve mice do not support such views of allocentric learning. In our data, none of the following was sufficient to generate subgoals: 1) spending time exploring the obstacle; 2) running along the homing-vector path and then being blocked by the obstacle; 3) learning a subgoal at the other obstacle edge; 4) targeting the obstacle edge while running away from the shelter; 5) investigating the obstacle in the absence of a shelter; and 6) investigating the obstacle while the threat area was blocked off. Furthermore, investigating the formerly obstructed area following obstacle removal did not restore direct homing-vector responses.

The subgoal strategy does contain elements of classical map-based navigation: it is learned in all-or-none fashion and depends on a sense of allocentric space, i.e. a 'map'; however, it also includes a component similar to response learning, which entails inflexible routes based on previous goal-directed movements. Hybrid strategies - combining rapid learning, inflexible routes, and special 'learning movements' - have been discovered before, as in the orientation flights of wasps (Collett, 1995). They are also a key part of a strain of research in the cognitive sciences called sensorimotor enactivism (Ward et al., 2017), which asserts that an explanation of learning should include not only how we extract meaning from sensory data but also how our actions are used to control this stream of data (Chase and Simon, 1973; Petitto and Marentette, 1991; Mataric, 1992; Clark, 1999). For instance, Ballard et al., 1997 demonstrated a key role for saccadic eye movements in devising a solution to a physical construction task. However, orientation flights entrain one-step routes to a visual beacon, and studies of enactive learning strategies have largely been limited to human psychology. Here we find that, in a experimental setup ripe for systems neuroscience, mice use learning movements to entrain multi-step routes to an obstructed goal. Our working model is that mice instinctively execute visually guided movements toward a salient wall edge; if this movement gives the mouse direct access to a subsequent goal (e.g., the shelter), then its target is memorized as a subgoal location. We hypothesize that a rapid, all-or-none learning rule works on practice homings, but the evidence proferred in this chapter was largely correlative. The next chapter will explore experiments to test a causal role for practice runs.

Memorizing subgoals confers distinct survival advantages: it can drive escape routes with the optimality of map-based planning and the rapidity of instinctive responses. However, this strategy is less flexible than responding to sensation or updating maps. The steady persistence of 50% biphasic escapes for tens of minutes after removing the obstacle was longer than expected, and it remains unclear how

mice learn to reinstate the homing-vector response after obstacle removal. Responses to imminent predatory threats are known to favor quick reaction times at the expense of computational sophistication (Mobbs et al., 2020), and so this inflexible strategy could in principle be specific to defensive behavior. However, we found that it was also used in a less urgent food-seeking task. Thus, subgoals appear to be a general building block for quickly learning spatial locations important for survival.

Contributions

Panagiota Iordanidou helped me run many of the escape experiments. Dario Campagner and Nabhojit Banerjee designed the food reward training protocol. Sarah Olesen and I performed the food reward experiments together. The acute obstacle removal experiment was engineered by the Sainsbury Wellcome Center FabLabs and Ruben do Vale.

Chapter 3

Action-driven mapping

When mice are placed in an arena with a shelter and an obstacle, they spontaneously execute continuous runs targeting the obstacle edge. Our main aim in this chapter is to test the causal necessity of these runs in learning that the obstacle edge is a subgoal. Notably, our evidence for this hypothesis in the previous chapter was largely correlational. We therefore set out to design a manipulation that could prevent mice from executing spontaneous runs to an obstacle edge.

3.1 Closed-loop activation of premotor cortex blocks edge-vector runs

To prevent confounding effects, our run-blocking manipulation should avoid modifying the external environment, should not decrease the opportunities for the animal to observe its environment, and should not generate place aversion. We found that closed-loop stimulation of premotor cortex (M2) fit all these criteria. We injected channelrhodopsin in excitatory neurons in the right, anterior M2, and performed optogenetic stimulation via an implanted optic fiber (Fig. 3.1b, Fig. 3.2a). In line with previous reports (Gradinaru et al., 2007; Magno et al., 2019), stimulating M2 with a 2-sec, 20-Hz pulse wave caused a low-latency (<200 ms) deceleration, halting, and leftward turning motion (Fig. 3.2b). This stimulation protocol did not generate place aversion when tested in a two-chamber place-preference assay (Fig. 3.2d). We thus leveraged this approach to specifically interrupt edge-vector runs during spontaneous exploration. Using online video tracking, we set up a virtual "trip wire" in between the threat area and the left obstacle edge; whenever mice crossed this line while moving in the direction of the edge, a 2-sec pulse of light was automatically delivered (Fig. 3.1c). Up to three subsequent pulses were triggered manually if the mouse continued moving toward the edge. All other movements, including runs to

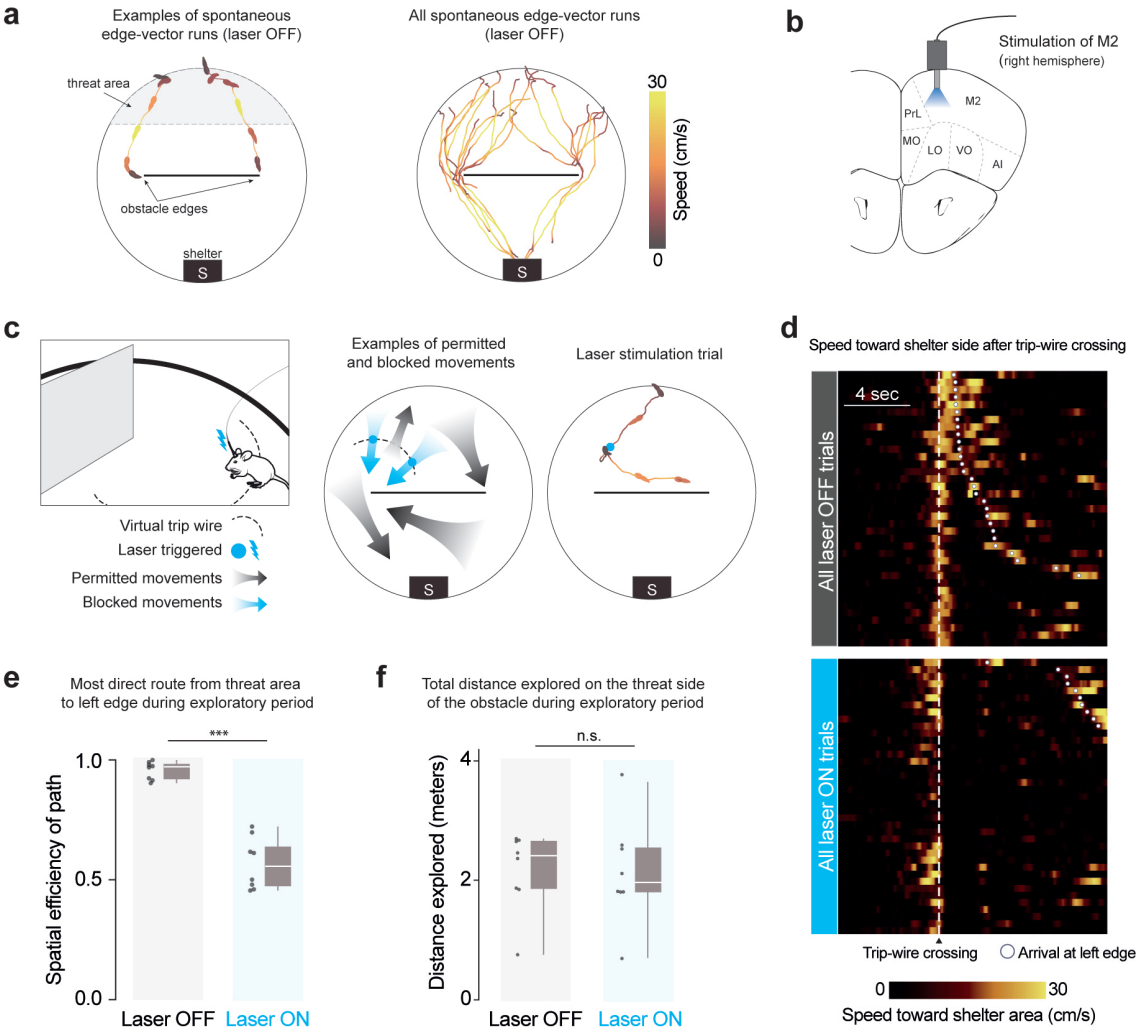

Figure 3.1: *A closed-loop neural manipulation interrupts edge-vector runs*

(a) Spontaneous edge-vector runs during the initial 20-minute exploration period. **(b)** Optic fibers were implanted in right premotor cortex. M2: supplementary motor cortex (premotor cortex), PrL: prelimbic cortex, MO/LO/VO: medial/lateral/ventral orbital cortex, AI: agranular insular cortex. **(c)** On crossing a virtual trip wire (dotted line) during exploration, mice automatically received a 2-sec, 20-Hz, 30-mW pulse of 473-nm light. In the example trial, the mouse was stimulated with two 2-sec pulses and then ran to the right side of the platform. **(d)** All trip-wire crossings, with and without laser stimulation, ordered by time of arrival to the left obstacle edge. Note that mice must be moving toward the shelter area (i.e., south) in order to trigger the trip wire. **(e)** White horizontal lines indicate median, black dots indicate mean, gray boxes indicate the first and third quartiles, and gray vertical lines indicate the range. Each dot represents one mouse/session. $p=5 \times 10^{-5}$, permutation test. **(f)** Distance explored on the threat half: p=0.5, permutation test.

the left edge along the obstacle or from the shelter, were not interrupted by laser stimulation.

We divided up injected and implanted animals into a laser-on (trip wire active)

and a control, laser-off group (trip wire inactive). Both groups of mice were allowed to explore a circular platform with a shelter and an obstacle for 20 minutes (n=8 mice/sessions; pictured in Fig. 3.3). During this time, all mice located the shelter and visited the entire platform, including the obstacle (Fig. 3.4a,c). In agreement with results from ch. 2, all mice in the laser-off group executed continuous turn-and-run movements from the threat area (Fig. 3.1a) toward the shelter area ('homing runs'; # per session: 6 [5, 8.25] (median [IQR]); Fig. 3.4b,d). These included at least one homing run that directly targeted an obstacle edge ('edge-vector runs'; # per session: 1.5 [1, 2.25] (median [IQR]); Fig. 3.1a, Fig. 3.4e). Mice in the laser-on group triggered 3.5 [2.75,6] (median [IQR]) laser stimulation trials, lasting 20 [16, 26] seconds in total and interrupting all potential edge-vector runs (Fig. 3.1d, Fig. 3.4b,e). While mice in the laser-off group executed nearly direct paths between the threat area and the left obstacle edge, the paths taken by mice in the stimulation group were twice as long, reflecting the inaccessibility of edge-vector runs (Fig. 3.1e; spatial efficiency here is defined the ratio of the straight-line path to the length of the path actually taken). Exploratory behavior in general, however, was not reduced. Mice in the stimulation condition explored the obstacle, the edge, the threat area and the entire arena as much as the control group (Fig. 3.1f, Fig. 3.4a,c).

3.2 Interrupting edge-vector runs abolishes subgoal learning

We next measured the impact of blocking edge-vector runs on subgoal learning. After the 20 min exploration period, we elicited escape behavior using a loud, unexpected crashing sound. Mice triggered an auditory threat stimulus automatically by entering the threat zone and staying there for 1.5 seconds. Escape routes were quantified using a target score and classified as targeting the obstacle edge ('edge vector') or the shelter ('homing vector') as in the previous chapter.

First, we acquired a negative-control distribution by letting a group of mice explore and escape in an open-field environment with no obstacle (n=8 mice; same viral injection and implantation procedure as above). As expected from previous work (Vale et al., 2017), mice generally responded to threats by turning and running directly along the homing vector (Fig. 3.5a). Second, we examined escapes in a positive-control condition known to generate subgoal learning. After the laser-off group explored the arena with the obstacle and shelter for 20 minutes, we removed the obstacle and triggered escapes (2-30 minutes later, IQR: 8-17 minutes). We found that 42% of escapes were directed toward the obstacle edge location, despite

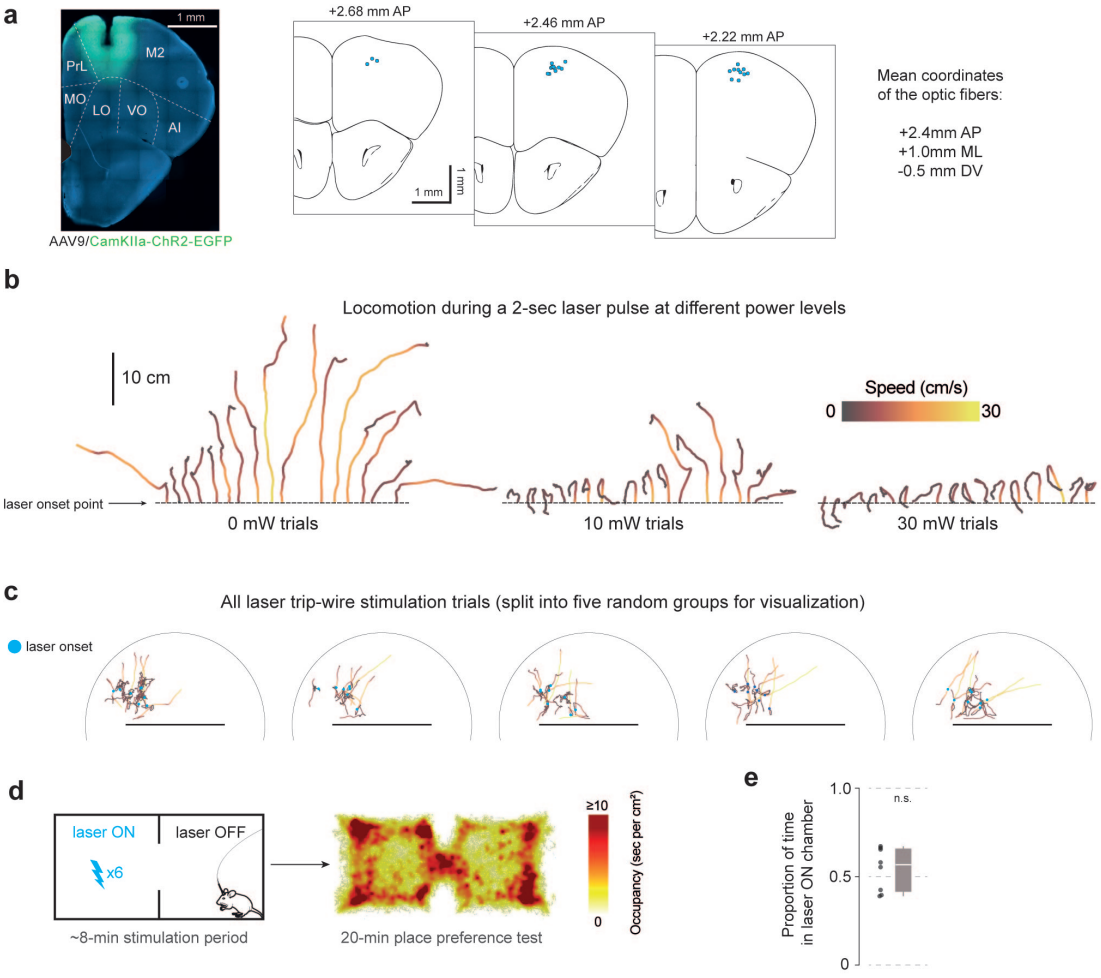

Figure 3.2: *Optogenetic stimulation of right premotor cortex*

(a) Left: example viral injection and optic fiber implantation site. M2: supplementary motor cortex (premotor cortex), PrL: prelimbic cortex, MO/LO/VO: medial/lateral/ventral orbital cortex, AI: agranular insular cortex. Right: Putative optic fiber tip locations are overlaid on brain-slice diagrams adapted from Paxinos and Franklin, 2019. AP and ML coordinates are relative to bregma, and DV coordinates are relative to the brain surface. **(b)** Locomotion following a 2-sec, 20-Hz, 30-mW pulse wave (duty cycle 50%) of 473-nm light in implanted mice. Each mouse received 4 trials at each laser power, sequentially interleaved. n = 4 mice. Lines ordered by the distance and direction of movement following laser onset. **(c)** Trajectories before and after laser stimulation, for the edge-vector blocking protocol. **(d)** Place preference assay. For the occupancy heatmap, stimulation is shown as if it were on the left side for all mice. **(e)** Occupancy in the stimulation chamber is not significantly below 50%. p = 0.7, Wilcoxon signed-rank test.

the obstacle being gone ('edge vectors'; 26 total escapes on the left side; more edge vectors than in the open field: p=0.003, permutation test; Fig. 3.5a-b). This result is consistent with ch.2, where we found that these edge-vector escapes reflect the memorization of a subgoal location.

Third, we tested the laser-on group, which explored with an obstacle and shelter

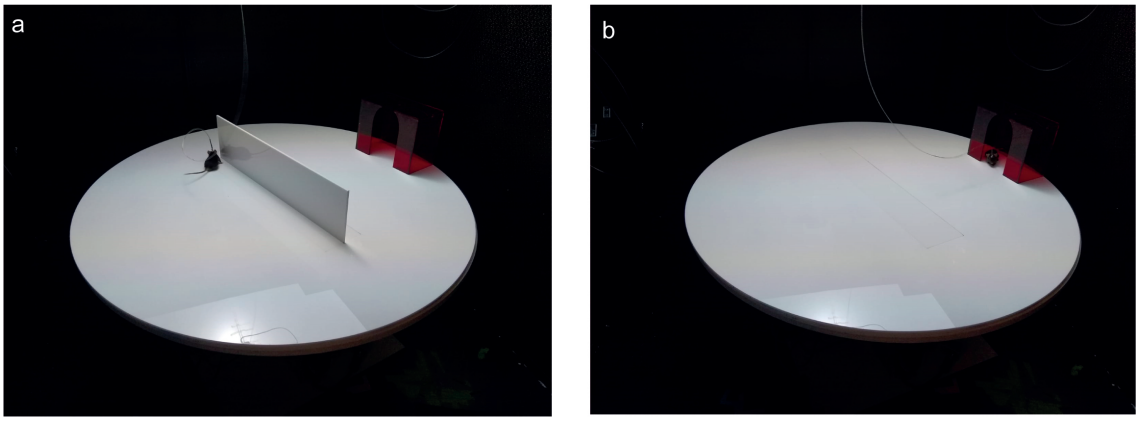

Figure 3.3: *Behavioral platforms for ch. 3*

(a) The platform with the wall obstacle. The platform is 92 cm in diameter, and the wall obstacle is 50 cm long x 12.5 cm tall. The shelter is 20 cm wide x 10 cm deep x 15 cm tall. It is made from red acrylic that is opaque to the mouse but transparent to red and infrared light. The mouse has just run to the right obstacle edge. **(b)** The platform with no obstacle. A central panel (50 cm wide x 10 cm wide) with the obstacle has been replaced, and a flat panel has been slotted in, in its place. The mouse is sitting in the shelter.

but had their exploratory edge-vector runs interrupted. After removing the obstacle, threat-evoked escape routes resembled the paths taken in the open-field condition rather than the subgoal-learning group (13% edge vectors; 23 escapes (left side); fewer edge vectors than in the laser-off condition: p=0.03, and not significantly more edge vectors than in the open field: p=0.2, permutation tests; Fig. 3.5a-b). Thus, interrupting spontaneous edge-vector runs abolished subgoal learning.

An alternative explanation could be that these mice did learn subgoals, but the stimulation during edge-vector runs taught them to avoid expressing edge-vector escapes. To address this possibility, we repeated the stimulation experiment (n=8 mice), this time allowing mice to perform two spontaneous trip-wire crossings without interruption. We then subjected them to the same edge-vector-blocking protocol as above, blocking about three runs (3 [1.75, 4.25] laser trials per session (median [IQR]) lasting 16 [5.5, 26.5] secs in total; Fig. 3.6a). Removing the obstacle and triggering escapes now revealed robust subgoal behavior (65% edge vectors; n=23 escapes (left side); more edge vectors than in the open field: $p=3 \times 10^{-4}$, and not significantly fewer edge vectors than the laser-off condition: p=.9, permutation tests). This shows that our manipulation does not reduce the use of subgoals once they are learned and therefore suggests that edge-vector runs are causally required for learning subgoals.

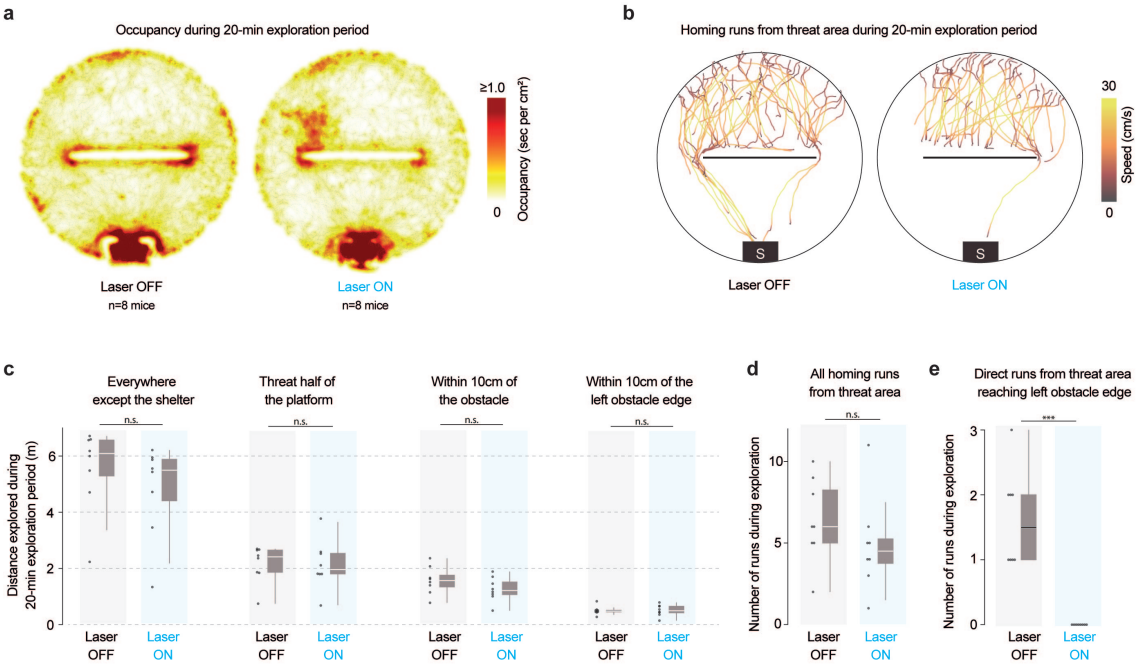

Figure 3.4: *(No) effect of optogenetic stimulation on exploration*

(a) Occupancy heatmaps are smoothed with a gaussian filter (σ = 1 cm). **(b)** Runs from all eight mice in each condition. Right: note that homing runs do not reach the left obstacle edge due to the closed-loop optogenetic stimulation. **(c)** Everywhere except the shelter: p = 0.2; threat half: p = 0.5; obstacle: p = 0.1, edge: p = 0.5, permutation tests. **(d)** Total number of homing runs (trajectories shown in panel b): p = 0.15. **(e)** Runs reaching the left edge: p = 3×10^{-5}, permutation test.

Blocking edge-to-shelter runs does not affect subgoals

Spontaneous edge-vector runs are often followed by an edge-to-shelter run. After completing an edge-vector run, mice in the laser-off condition reach the shelter within 2.5 [1.7,10] secs (median [IQR]), generally taking direct paths (spatial efficiency: .87 [.47, .95]; 1.0 corresponds to the direct path). We therefore considered whether edge-vector runs support subgoal learning because they are part of a sequence of actions that quickly brings the mouse from the threat zone to the shelter.

To test whether edge-to-shelter runs are important for learning, we repeated the stimulation experiment (n=8 mice), but with a new trip-wire location. Using 10-sec laser pulses, we stopped movements from the left obstacle edge toward the shelter (restricted to edge-to-shelter movements that occurred after having crossed the original trip wire, i.e. the second phase of a threat-area-to-edge-to-shelter run; 3 [2, 3.25] laser trials per session (median [IQR]) lasting 25 [20, 30] secs in total; Fig. 3.7a). Due to this manipulation, edge-vector runs on the left side were followed by long, slow paths to shelter (seconds to shelter: 29 [18, 55]; spatial efficiency: .28 [.13, .37]; slower than the laser-off condition: p=1×10^{-3}; less spatially efficient than

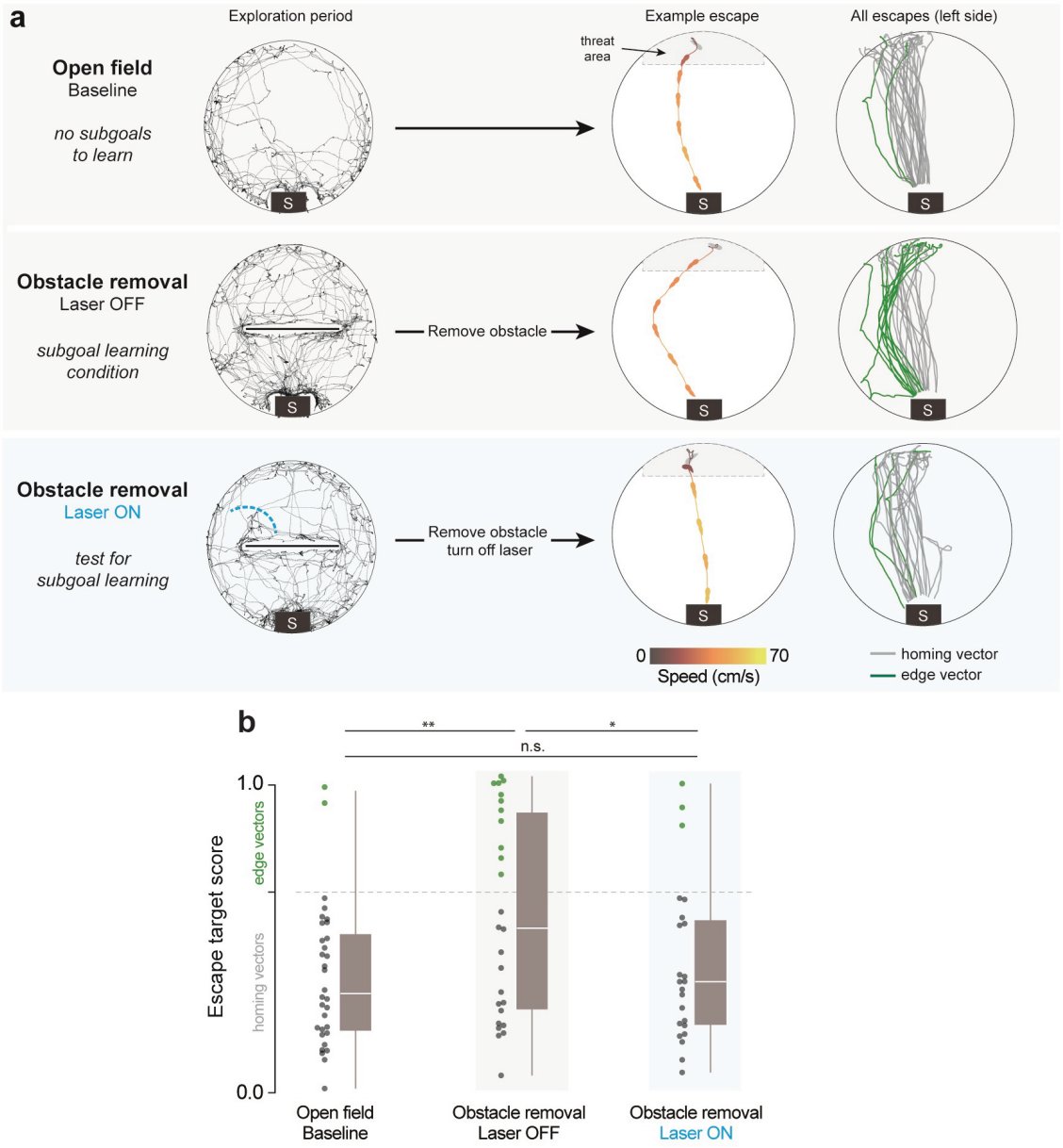

Figure 3.5: *Interrupting spontaneous edge-vector runs abolishes subgoal learning*

(a) Black traces show exploration during an example session. Lines and silhouette traces show escape routes from threat onset to shelter arrival. Analysis is limited analysis to escapes on the left side of the platform. n=8 mice per condition. **(b)** Summary of escape trajectory data.

the laser-off condition: $p=2 \times 10^{-3}$, permutation tests). Despite this effect, removing the obstacle and triggering escapes revealed robust subgoal behavior (55% edge vectors; n=23 escapes (left side); Fig. 3.7b-c; more edge vectors than in the open field: $p=1 \times 10^{-4}$, and not significantly fewer edge vectors than the laser-off condition: p=.8, permutation tests). Thus, for their causal role in subgoal learning, edge-vector runs do not need to be rapidly followed by the extrinsic reward of entering the shelter.

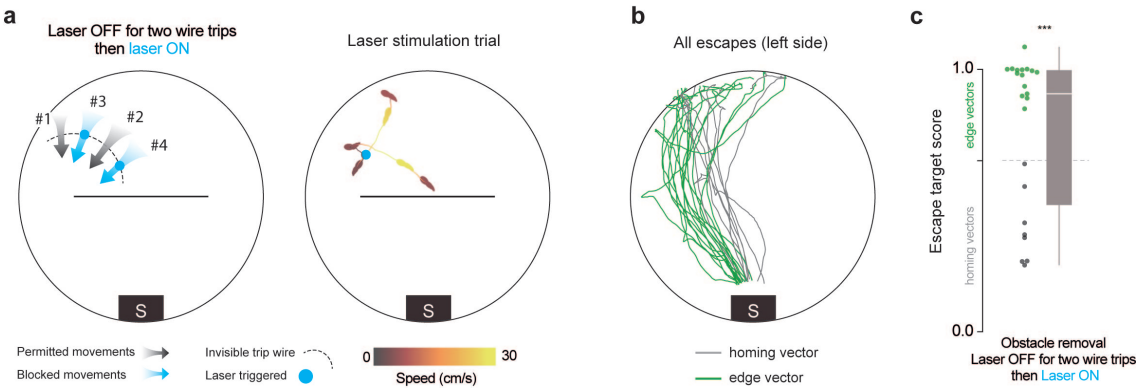

Figure 3.6: *Blocking edge-vector runs after allowing two runs*

(a) Schematic of stimulation blocking all but the first two edge-vector runs. The example shows four seconds after laser onset: the mouse was stimulated for two seconds, and then ran toward the center of the obstacle. **(b)** Escapes after obstacle removal. **(c)** Summary of escape trajectory data.

This result also supports the argument that optogenetic stimulation at the left edge does not teach the mice to avoid passing by that location during escapes.

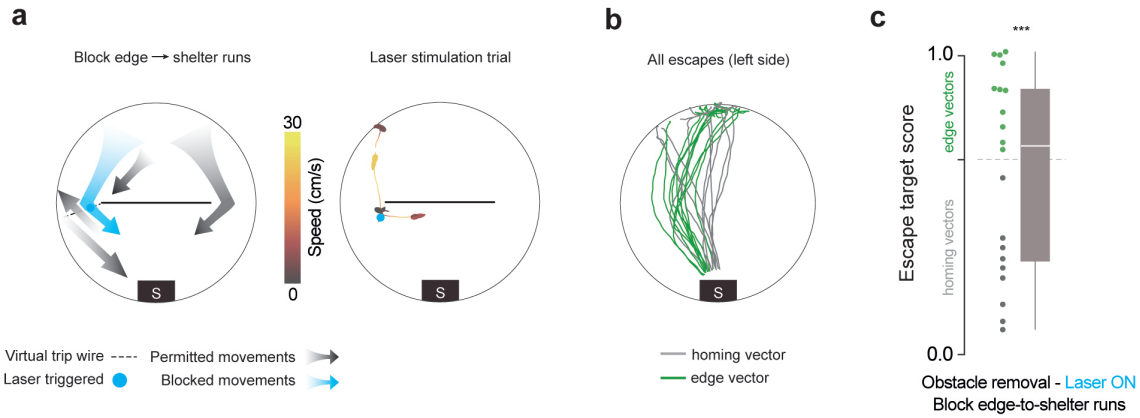

Figure 3.7: *Blocking edge-to-shelter runs*

(a) Blocking left-edge-to-shelter runs. In the example trial, the mouse was stimulated for ten seconds, and then ran toward the center of the platform. **(b)** Escapes after obstacle removal. **(c)** Summary of escape trajectory data.

3.3 Subgoal-escape start points are determined by spatial rules

The results from the previous experiment suggest that learning subgoals with edge-vector runs is not simply a matter of reinforcing actions that lead to the shelter. This fits with the finding in ch. 2 that these subgoals are stored as allocentric

locations rather than egocentric movements, and it raises the possibility that the learning process combines actions and spatial information. To explore this further, we investigated the rules governing the set of locations from which mice initiate memory-guided subgoal escapes - the 'initiation set' of subgoal escapes. We aimed to determine whether the initiation set is 1) spread indiscriminately throughout the environment; 2) restricted to the vicinity of previous edge-vector-run start positions; or 3) related to the spatial layout of the environment, independent of past actions. Option 1 would be expected if mice learned to execute edge-vector actions without taking into account their starting location; option 2 would be expected if mice learned to repeat edge-vector actions based on proximity to previous successful actions; and option 3 would be expected if mice selected the subgoal strategy through a map-based process. We first repeated the obstacle removal experiment but now elicited escapes from in front of the obstacle location, near to the shelter (n=8 mice with no laser stimulation, 28 escapes; Fig. 3.8a). From this starting point, mice did not escape by running toward a subgoal location but instead fled directly to shelter. This result suggests that the initiation set is spatially confined rather than indiscriminate.

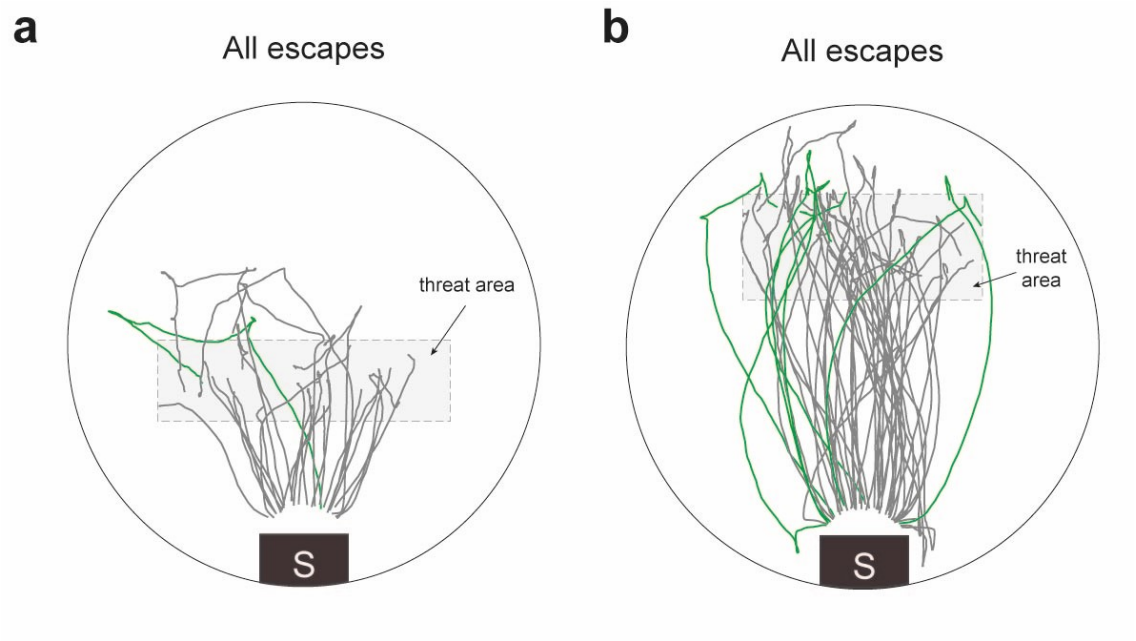

Figure 3.8: *Obstacle-removal escapes with a modified threat zone*

(a) Escapes triggered after obstacle removal, using the new threat zone (dotted lines). Only 1/28 escapes (the green trace) begins by moving toward the obstacle edge location; however, this appears to be a continuation of the pre-threat movement rather than a genuine subgoal. **(b)** Escapes triggered after obstacle removal, using another new threat zone (dotted lines).

Next, we tested whether the initiation set is confined to the area in which

spontaneous edge-vector homing runs had previously occurred. We modified our laser stimulation experiment with a new trip wire location, so that edge-vector runs were allowed from a section of the arena next to the threat zone, but were interrupted if they started within the threat zone (n=8 mice; 2 [1.75, 4] laser trials per session (median [IQR]) lasting 4 [6, 9] secs; Fig. 3.9a,b). As before, laser stimulation succeeded in blocking edge-vector runs from the threat zone. In this configuration, however, mice were still able to execute edge-vector runs starting from the area to the left of the threat zone (illustrated by the leftmost gray arrow in Fig. 3.9a). Removing the obstacle and triggering escapes in this cohort revealed robust subgoal behavior (63% edge vectors; n=19 escapes (left side); Fig. 3.9b-c; more edge vectors than in the open field: $p=6\times10^{-4}$, and not significantly fewer edge vectors than the laser-off condition: p=.8, permutation tests). Thus, the initiation set for subgoal escapes extends beyond the locations in which successful edge-vector runs have been initiated (Fig. 3.9b inset). This result also reaffirms that optogenetic stimulation does not teach mice to avoid paths that are blocked by laser stimulation during exploration.

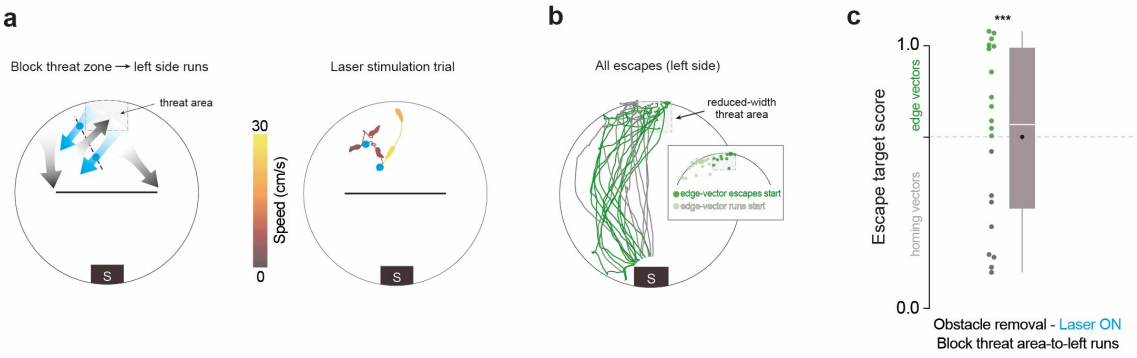

Figure 3.9: *Subgoal escapes do not need to start near practice runs*

(a) Blocking threat-zone-to-left-side runs with a new trip-wire location. The dotted gray line outlines the threat zone used in this experiment. In the example trial, there were two consecutive trip-wire crossings (2-sec stimulations), then the mouse moved back toward the threat zone. **(b)** Escapes after obstacle removal. Inset: All start locations for spontaneous edge-vector runs (light green) and subsequent edge-vector escapes (dark green). **(c)** Summary of escape trajectory data.

To more precisely examine the impact of spatial location on subgoal behavior, we repeated the obstacle removal experiment with a larger threat zone, located between the obstacle location and the original threat zone (n=8 mice, 53 escapes; no laser stimulation; Fig. 3.10a, Fig. 3.8d). By combining these escapes with the original threat zone data, we could test the relationship between the location of escape onset and the tendency to use a subgoal, using logistic regression (n=40

total sessions, 207 escapes; Fig. 3.10b-d). We found that being closer to previous edge-vector runs was not significantly related to the likelihood of executing edge-vector escapes (McFadden's pseudo-R^2=0.086; p=0.5, permutation test; Fig. 3.10g, Fig. 3.8c-d); in fact, the non-significant relationship tended toward greater distance from an edge-vector run predicting a higher likelihood of edge-vector escapes. In contrast, a number of spatial metrics were effective predictors of edge-vector escape probability (Fig. 3.10f-g, Fig. 3.8c-e). These include the distance from the obstacle (pseudo-R^2=0.28; p=0.007; values of 0.2-0.4 represent 'excellent fit' (McFadden, 1977)), the distance from the central axis of the platform (the axis perpendicular to the obstacle; pseudo-R^2=0.26; p=0.01), the distance from the shelter (pseudo-R^2=0.29; p=0.006), and the angle between the edge-vector and homing-vector paths (pseudo-R^2=0.29; p=0.006). Thus, the initiation set is defined in relation to the layout of the environment rather than proximity to previous successful actions.

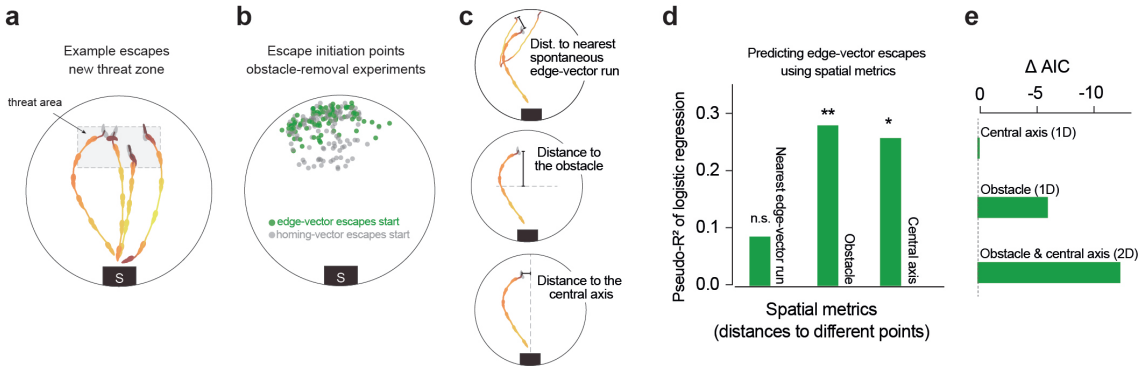

Figure 3.10: *Subgoal-escape start points are determined by spatial rules*

(a) Four example escapes triggered after obstacle removal, using the new threat zone. **(b)** Data from all obstacle-removal experiments are combined here (with the exception of the block-edge-vectors experiment). In this visualization, right-sided escapes are flipped horizontally in this visualization; thus, all the green dots can be seen as left-edge vectors. Each dot represents one escape. **(c)** Three spatial metrics used to predict the likelihood of executing an edge-vector escape. Silhouettes: an example escape. Black bar: the distance being measured. Thin orange lines: the mouse's history of practice edge-vector runs. **(d)** The strength of the relationship between each metric and the odds of executing edge-vector escapes, measured by McFadden's pseudo-R^2 and a permutation test using 10,000 random shuffles of the edge-vector/homing-vector labels. **(e)** Akaike Information Criterion (AIC) analysis on a logistic regression with different predictors. $\Delta AIC_i = AIC_i - AIC_{min}$ where AIC_{min} here is the AIC from the model with the single, distance-from-central-axis predictor.

We next analyzed whether a two-dimensional spatial-location predictor fit the data better than a one-dimensional predictor by applying the Akaike Information Criterion (AIC) analysis to the logistic regression model (Fig. 3.10h). If the initiation

set were fully explained by the mouse's perception of which side of the obstacle it is on, or of how close to the shelter it is, then combining multiple spatial predictors should not improve the model (i.e., AIC should increase). On the other hand, if mice use their 2D position within the environment to select whether to use a subgoal, then using 2D spatial information should improve the model (i.e., AIC should decrease). In line with this possibility, using only distance from the obstacle (i.e. distance along the y-axis) gave an AIC of 206.8, using distance from the central axis (i.e., distance along the x axis) gave an AIC of 212.9, and using both dimensions as input gave an AIC of 200.5. The AIC decrease of 6.3 from the model using obstacle distance only falls within the ΔAIC range of 4-7, indicating that the combined, 2D model has 'considerably' more support than either 1D model (Burnham and Anderson, 2004). We found similar results when we compared using distance from the shelter only to using both distance from the shelter and distance to the central axis (AIC decrease of 5.8). Thus, mice's selection between subgoal vs. direct routes is modulated by their two-dimensional starting position within the arena.

3.4 Interim discussion

When a mouse investigates a new environment, it does not act like a 'random agent'. Instead, its exploration consists of purposive, extended, sensorimotor actions. In this chapter, we have demonstrated that one such class of movements - running to an obstacle edge that grants direct access to a goal - plays a causal role in the process of gaining useful spatial information about the environment.

As discussed in ch. 2, typical spatial learning models rely on two steps to explain allocentric behavior: 1) constructing an internal map of space by observing how locations and obstructions in the environment are positioned relative to each other; 2) determining a goal location; and 3) using this map to derive a useful subgoal location, computed either at decision time or in advance during rest (Spiers and Gilbert, 2015; Edvardsen et al., 2020). This process is well suited for agents that learn by diffusing throughout their environment, be it randomly or with a bias toward unexplored territory (Schulz and Gershman, 2019). However, it does not account for the prevalence of goal- and object-oriented actions in natural exploratory patterns (Crowcroft, 1966; Schulz et al., 2017).

We thus explored a potential role for a fourth process: executing 'practice runs' to candidate subgoal locations during exploration. This idea follows from a strain of research in the cognitive sciences called sensorimotor enactivism (Ward et al., 2017), as mentioned in the previous chapter. Here, we made the innovation of combining

a key enactivist principle - the importance of intrinsically motivated actions for learning - with the causal perturbation techniques and spatial behaviors available in rodent neuroscience. Specifically, we used closed-loop optogenetic stimulation of M2 to interrupt edge-vector practice runs, and we found that this manipulation abolished subgoal escape routes.

It is important to note that this effect does not inform us about the role that M2 may play in computing subgoals. That question is not the point of this study. Notably, three of the M2 stimulation protocols spared edge-vector runs, and these manipulations did not impair learning. Thus, stimulating M2 does not intrinsically affect spatial learning. Only when M2 stimulation interrupted practice edge-vector runs did we see the effect. Our results therefore indicate that *the edge-vector actions themselves* are necessary for triggering subgoal memorization.

Ideally, the premotor-cortex stimulation works by causing the mouse to suddenly 'decide' to abort the edge-vector run. If it instead works by punishing edge-vector runs or by causing the mouse to experience an invisible barrier, our claim about practice runs' role in learning would be less certain. The three control experiments were designed to test this, and they showed a clear pattern in which only the manipulation that blocked all edge vectors blocked learning. Still, if there is a control experiment that we overlooked, our claim should be tempered. One way to gain more certainty would be to find a manipulation that specifically causes the mouse to decide to make an edge-vector practice run and to see if this accelerates subgoal learning.

Based on the specificity of the correlation between practice runs on one side of the obstacle and edge-vector escapes to that same side (ch. 2), we expect this manipulation to only affect subgoals located at the left obstacle edge. Thus, we restricted analysis to escapes that passed closer to that edge. While subgoal escapes on the other side were not significantly reduced, there was a trend in that direction. Future replications of this experiment would benefit from a larger dataset of escapes on both sides so that this question can be answered more definitively.

One interpretation of the need for practice runs in learning could be that subgoal behavior is a naturalistic form of operant conditioning. In this view, edge-vector runs are followed by reinforcement and then simply get repeated in response to threat. This framework could explain why edge-vector responses persist after obstacle removal: they are habits that have not yet been 'extinguished'. On the other hand, subgoal learning diverges from instrumental learning in two ways: it operates within an allocentric framework (seen as distinct from the response strategy (Restle, 1957; Packard et al., 1989; Doeller et al., 2008; Geerts et al., 2020)), and it only requires 1-2 practice runs (even simple instrumental training takes tens of learning trials (Baron and Meltzer, 2001)). More importantly, the set of locations from which mice

initiate subgoal escapes are defined by the mouse's spatial position relative to the obstacle and shelter, and not by their proximity to previous edge-vector runs. The concepts of action and reinforcement are therefore insufficient for explaining subgoal memorization; an internal map of space must also be invoked.

One of the key experiments demonstrating the specificity of the edge-vector manipulation was the experiment in which we interrupted edge-to-shelter runs. The lack of effect in this condition suggested that practice runs only need to arrive at the obstacle edge to trigger learning, and not all the way to the shelter. This raises the question of how edge-vector runs are associated with shelter such that subgoal learning can occur. There are a variety of possibilities. From a response-learning perspective, mice could perceive the shelter upon passing the obstacle edge and therefore experience reward. If the shelter is too distant to be seen or smelled, instrumental chaining (Hull, 1934; Gollub, 1977) could be used. This is a known phenomenon in which an action is rewarded for arriving at the starting point of a different action that is associated with reward. Thus, assuming that the edge-to-shelter run has previously been performed and rewarded, running to the obstacle edge would itself positively reinforce the edge-vector action. Alternatively, a mapping strategy is possible. In this case, mice could 'connect the dots' of their topological spatial map, inferring that since the threat zone and obstacle edge are connected and the obstacle edge and shelter are connected, that the obstacle edge could be a subgoal. Our intuition is that the true explanation will be a combination of all of these possibilities.

Finally, we showed that the decision of subgoal vs. homing-vector escapes is modulated by the animal's spatial position in a way that does not reflect proximity to previous practice runs. For one, we used an optogenetic manipulation to show that subgoal escapes are not reduced when they are required to start outside of the bounds of previous edge-vector runs. Second, an analysis of starting positions and edge-vector escape likelihood revealed a significant propensity to use subgoals when the mouse is further back from the obstacle location and farther from the arena's central vertical axis. There are several possible explanations of this pattern. First, it could reflect the outcome of a spatial cost-benefit analysis: the preferred subgoal-escape starting points are in the locations where the subgoal route is almost as short as the homing vector. Second, it could indicate that the memory-guided escape strategy is only used when the animal is so far away from the shelter or obstacle's center that the animals know that they cannot rely on local visual cues. One final possibility is that the mouse clusters its spatial map of the arena into regions with similar features. In that case, subgoal actions might generalize across the back perimeter region but not to the region right in front of the obstacle.

Contributions

Panagiota Iordanidou performed the histology processing to confirm the location of the injections and implantations in the brain.

Chapter 4

Reinforcement learning models of escape behavior

4.1 Introduction to RL models of navigation

The key concepts from experimental psychology and behavioral neuroscience that we have been using, such as 'habits' or a 'cognitive map,' do not exactly map onto the complex, spontaneous mouse behavior of subgoal escapes. Our next aim was therefore to model our results through the lens of a distinct perspective that could uncover some of the general computational principles of subgoal learning, or at least principles whose conceptual baggage is complementary to that of behavioral neuroscience.

We selected reinforcement learning (RL) as our modelling strategy. RL is a branch of machine learning that addresses how to make decisions in an environment in order to gain the maximum amount of future rewards (Sutton and Barto, 2018). To approach this problem, RL uses the formalism of a Markov decision process (MDP). An MDP expresses the agent-environment complex as a set of states, a set of possible actions, and a function defining what happens upon taking a particular action in a particular state. Two possible things can happen after each action: a new state can be reached and a reward can be given to the agent. States can be defined as physical locations within an environment, but they could also be physiological states such as hunger, abstract states such as proximity to a subgoal, or some combination of these. The job of the agent is to devise the action policy that will culminate in the most reward. If the states are physical locations, actions are movements between neighboring locations, and rewards are given when the agent reaches a goal location, then the RL problem becomes equivalent to a spatial navigation problem.

We chose to model spatial navigation using RL for three principal reasons. First, RL is in the zeitgeist of contemporary systems neuroscience and will therefore

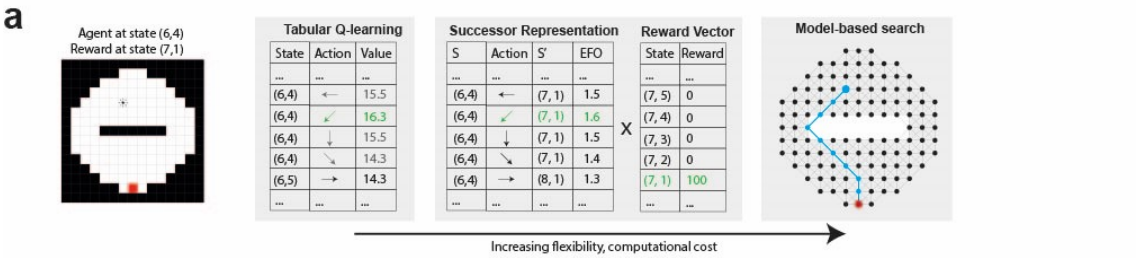

Figure 4.1: *The core reinforcement learning models we selected*

(a) Schematic of our core reinforcement learning models. Q-learning updates a table of values for each state-action pair. Successor Representation updates a matrix recording how much each state predicts future occupancy in each other state, as well as a table of how much reward is in each state. Our model-based agents updates a list of stats, a graph of connections between each state, and a label of the amount of reward in each state.

be accessible to a neuroscience audience. This makes sense since RL has, from its inception, been deeply tied to neural and behavioral models of trial-and-error learning in animals (Sutton and Barto, 2018). Second, RL has previously been used to explain findings from behavior and neural activity during rodent navigation (Spiers and Gilbert, 2015; Stachenfeld et al., 2017; De Cothi et al., 2022), giving us a starting point of specific algorithms to test. Third, and most important, with RL it is not necessary to hardcode a solution the route-planning problem. Instead, RL provides a spectrum of high-level algorithms for solving the generic problem of maximizing reward extracted from a given environment (Sutton and Barto, 2018). Therefore, we can set these models in motion on simulations of our behavioral experiments and discover unexpected computational insights.

Three main hypotheses have emerged from the small field of modelling navigation behavior with RL. The first is that the hippocampal map vs. basal-ganglial-habits dichotomy is equivalent to the model-based vs. model-free distinction in RL (Spiers and Gilbert, 2015). Model-based algorithms learn a representation of the structure of the environment (e.g. a map of which locations have obstacles, are empty, and/or contain a reward) and use it to calculate the best series of actions to get to a reward. These methods are highly data efficient but the planning procedure, e.g. searching the model for all possible routes to a goal, comes with considerable computational complexity and overhead. Model-free algorithms instead assign a single value to each possible action in each location, gradually updated throughout the agent's history of taking actions and receiving rewards; they then select routes by taking the highest-value action at each step. This approach makes the action selection process simple and straightforward but at the cost of slow updating.

The second hypothesis is that the model-based system here is implemented

through an algorithm called the Successor Representation (SR; Russek et al., 2017; Stachenfeld et al., 2017; De Cothi et al., 2022). The SR represents a compromise between model-free and model-based learning, with an updating speed and computational complexity somewhere in-between those two approaches. Since we implement the SR in the following section, it is described in more detail below and in the methods section.

Finally, it has been proposed that hierarchical reinforcement learning (HRL; Sutton et al., 1999) provides a better fit for behavioral (Tomov et al., 2020) and neural (Ribas-Fernandes et al., 2011) data than standard RL. In standard RL setups, agents select among low-level actions at each time-step (e.g. moving a small amount toward the left). In HRL, agents can also select and attribute value to high-level actions, groups of low-level actions that sequentially implement a subroutine (e.g. moving to the door connecting two rooms). Breaking down complex problems into a series of sub-problems can vastly reduce the complexity of learning and is well suited to transferring knowledge (i.e. how to perform a particular sub-task) to new tasks. HRL is thus particularly well suited to modelling subgoal behavior.

4.2 Modelling navigation in the obstacle-removal experiment

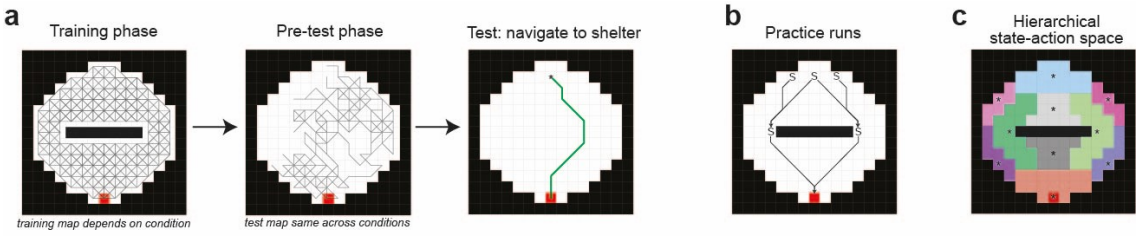

Figure 4.2: *Reinforcement learning models of mouse escape behavior*

(a) Gray traces represent paths taken during exploration/learning. The training map here is the map used in condition 1. Accessible states are white; blocked states are in black; rewarded states are red. Middle: a representative pre-test phase. Right: an example "escape" trajectory from the threat zone (asterisk) to the shelter (red square). **(b)** The training-phase exploration was a random walk punctuated with practice runs. Each 'S' represents a start point for the hard-coded action sequence and each arrow head represents the terminal state. The sequences were triggered with probability p=0.2 upon entering each start state. **(c)** We segmented the arena into regions, shown here, for the hierarchical state-space agent. Each colored region represents a distinct state. When executing the policy, the agent selects a neighboring high-level region to move to. It then proceeds from its current location to the central location indicated by the asterisk in that high-level region.

We used a panel of RL algorithms that have previously been used to model navigation (Russek et al., 2017; De Cothi et al., 2022) in a tractable grid-world environment based on our experimental setup. The three core algorithms we used were model-free tabular Q-learning, the Successor Representation (SR; Dayan, 1993), and (topological) model-based tree search (Fig. 4.1). The tabular Q-learning agent incrementally learns the value of each of the 944 state-action pairs (e.g., "go northwest from the shelter state"), based on its history of receiving rewards. The SR also computes state-action values, but it does so differently. It updates two separate representations: a spatial representation measuring which locations tend to follow each state-action pair and a reward representation. It then combines this information to compute the estimated value of each state-action pair. Third, the model-based agent does not update action values at all. Instead, it updates a graphical representation of the arena and searches through this graph to calculate optimal routes to the reward. This model is different from the other two algorithms in two big ways: it uses model-based search and it updates its model immediately after visiting a state. To disambiguate these difference, we added a model-based agent that updates its model gradually, taking into account its past 15 observations of each edge in the graph to decide if two adjacent states are connected or blocked by a barrier.

Similar to the experiments in mice, all simulations include a training map (e.g. the arena with an obstacle present) and a test map (e.g. the arena with the obstacle removed) and take place over three phases (Fig. 4.2a). First is the **training phase**, where the agent explores a training map for a duration long enough to learn a route from the threat zone to the shelter (Table 6.4). Importantly, this phase also includes stochastically generated practice-run sequences from the threat zone to the obstacle edge and from here to the shelter, to mimic the natural exploratory pattern observed in mice (Fig. 4.2b). The next phase is the **pre-test phase**, which takes place in the test map. In this phase, the agent starts in the shelter and executes a random-exploration movement policy until reaching the threat zone. Finally, there is a **test phase**: executing the learned policy in the test map, starting from the threat zone. The pre-test and test phases are repeated three times per seed (with a total of 100 seeds), similar to how each mouse performs several escape trials. We selected four particularly revealing behavioral and optogenetic experiments to model *in silico* using this procedure. All test maps have a shelter and no obstacle, so the only difference between the four experimental conditions is the training map.

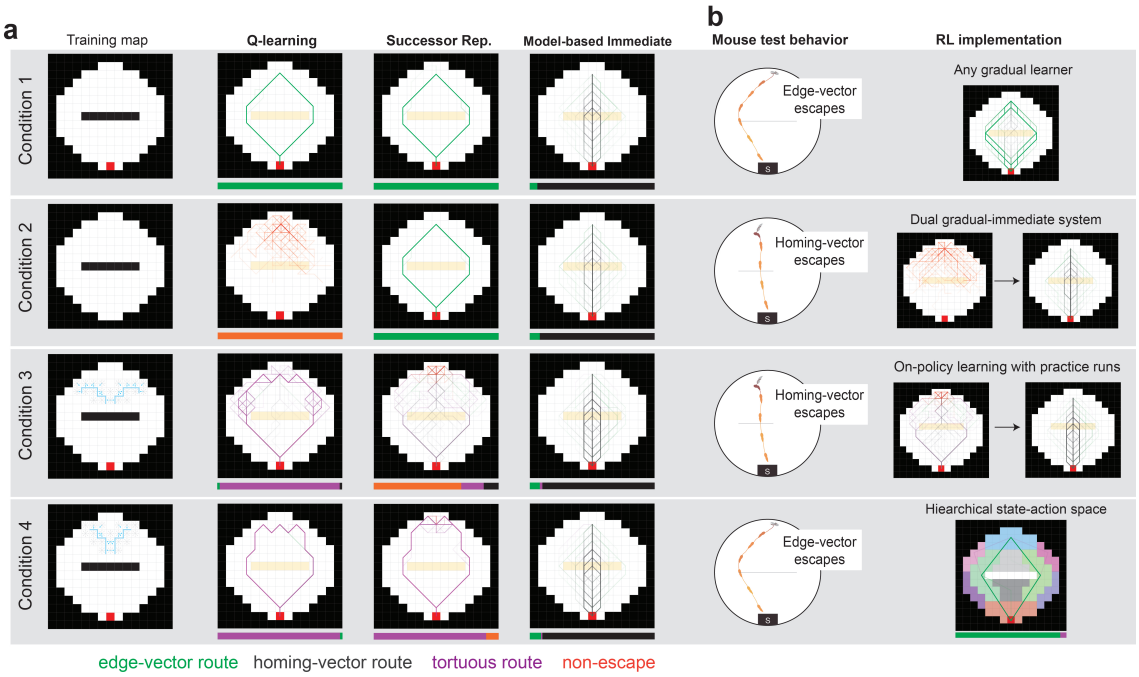

Figure 4.3: *Reinforcement learning models of mouse escape behavior*

(a) Escape runs from all three trials of all 100 random seeds in all four conditions. All trials are superimposed, with high transparency. Beneath each plot is a bar chart representing the proportion of each type of escape. In the training map of conditions 3-4, the one-way trip wire is represented by the blue line, and the blue arrows indicate all of the transitions that are blocked. **(b)** The qualitative mouse behavior in each condition and the type of RL agent that matches this behavior. In condition 1, the model-based (gradual) agent's behavior is shown. In condition 2, the Q-learning agent's failed trajectories and the model-based (immediate) agent's homing-vector escapes are shown. In condition 3, the SR agent's failed trajectories and the model-based (immediate) agent's homing-vector escapes are shown. In condition 4, the hierarchical-state-space Q-learning agent's trajectories are shown.

A dual reinforcement learning system matches mouse behavior after obstacle removal

The first training map corresponds to the basic obstacle removal experiment (Fig. 4.3a-b). Here, the training map has an obstacle and a shelter (i.e., a reward). After exploration in this condition, mice tended to execute edge-vector escape routes in the test phase. Similarly, the Q-learning, SR, and gradual model-based (MB-G; Fig. 4.4a) agents all exhibited persistent escape routes around the obstacle. The immediate-learning model-based (MB-I), on the other hand, was able to update its model during the test-map exploration and compute the new, fastest route to the shelter 94% of the time. The differentiating factor here is whether the agents update their policy immediately upon observing the changed environment (MB-I) or

incrementally/stochastically (all others). In the latter case, the pre-test exploration is too brief to learn the homing-vector path.

In the second condition, the training map has an obstacle but no shelter (Fig. 4.3a). Mice in this experiment (ch. 2) failed to learn edge-vector routes and instead escaped using homing vectors. The only agent to take homing vectors here - MB-I (92% homing vectors) - was the agent that did not persistently execute edge vectors in condition 1. The remaining agents differed in their behavior. The SR (100% edge vectors; Fig. 4.3a) and MB-G (93% edge vectors; Fig. 4.3b) agents learned edge vectors, thanks to their ability to separately solve spatial-learning problems even in the absence of reward. Q-learning came closer to the behavioral data: it failed to learn edge vectors (100% non-escape; Fig. 4.3a). This agent cannot learn without reward in the environment, so it was unable to come up with any escape route in this condition.

Overall, mice exhibit a pattern unlike any of these RL agents. Mice fail to immediately learn a homing-vector path in condition 1, but they do immediately learn the homing-vector path when they do not have a memorized policy in place (condition 2). For the RL models, this represents a paradox: the models that learn fast enough to run straight to shelter in condition 2 will also do so in condition 1. What *does* work here is a dual system that can switch between flexible and inflexible learners depending on the situation (Daw et al., 2005; Geerts et al., 2020). For example, an agent could contain both a Q-learning and an MB-I system. When the Q-learning model suggests an action with a positive value above some threshold, then the agent will take that action. If no such action is available, as in condition 2, then the immediate model-based system is invoked to find a novel route (Fig. 4.3b). This dual system matches mouse behavior on conditions 1-2.

Behavior with the full trip wire is matched with non-uniform exploration and on-policy learning

Next, we added the optogenetic trip wire to our modelling environment. In addition to the obstacle and shelter, the training map now contains one-way obstacles blocking paths from the threat area to the obstacle edge. The mice in this experiment again failed to learn edge-vector routes. We are thus looking for a gradual learning system that fails to learn viable escapes with the trip wire present, thereby triggering the backup immediate learner. For Q-learning (98% tortuous routes around both trip wire and obstacle; Fig. 4.3a) and MB-G (78% tortuous routes; 19% homing vectors; Fig. 4.4a), the trip wire simply adds an additional detour. These agents are perfectly able to learn tortuous routes around both the trip wire and the obstacle. This

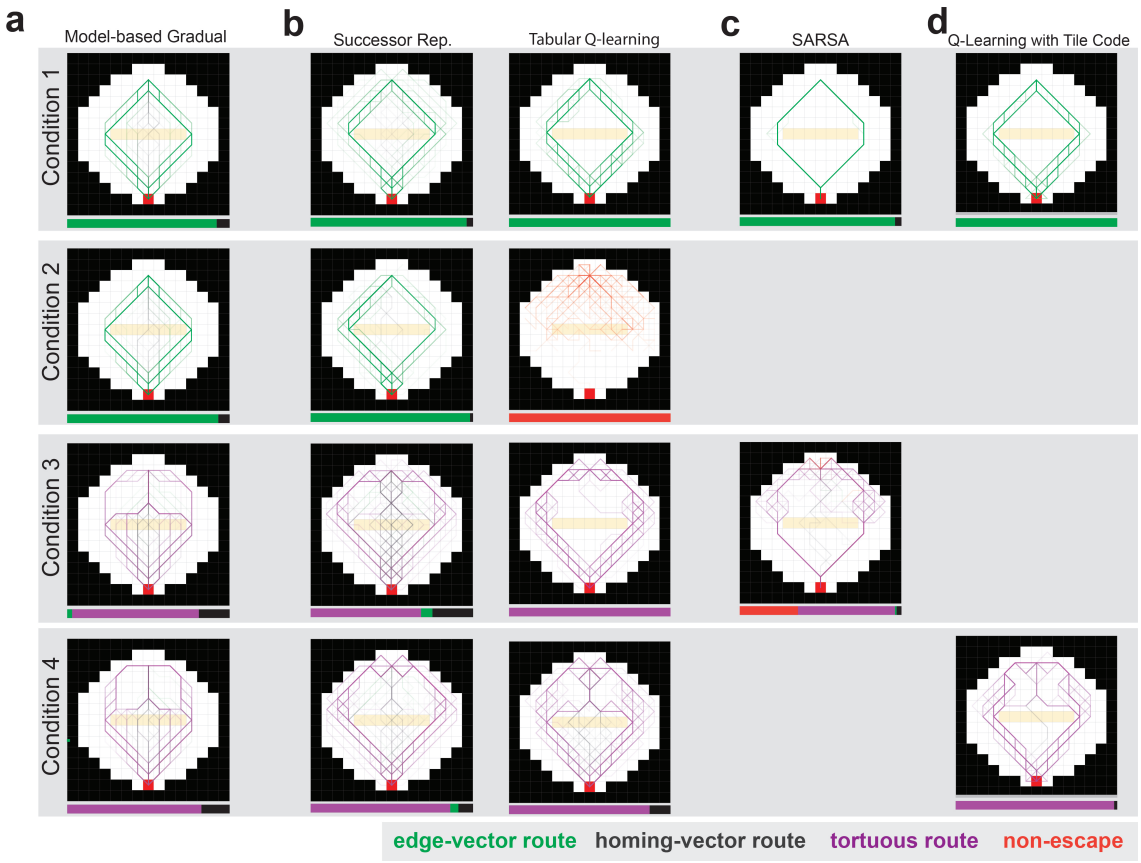

Figure 4.4: *Reinforcement learning models: extended results*

(a) Results for the gradual-learning model-based agent. **(b)** Results for the successor representation and Q learning, with random exploration (no practice runs). **(c)** Results for the SARSA agent. This agent is similar to tabular Q-learning, but it learns *on-policy* (the policy it converges to depends on its exploration policy). **(d)** Results for Q-learning with tile coding. Instead of a representing each state individually, tile-coding agent represents its approximate location with a set of n x n tiles.

appears to represent another paradox for the RL agents: models that memorize routes around the physical barrier will tend to do the same with the trip wire.

The SR agent (70% non-escape; Fig. 4.3a) learns routes around the obstacle in condition 1 but struggles with the trip wire here. This agent is able to overcome that paradox for two reasons. The first reason is the presence of practice runs in the training phase. With a fully random policy, the SR agent is able to learns routes in condition 3 just as quickly as in routes in condition 1 (Fig. 4.4b). Thus, it is the practice edge-vector runs that predispose this agent to learn edge-vector routes faster than other, arbitrary paths through space. The second reason is that, unlike Q-learning, our SR implementation is an *on-policy* learner (Sutton and Barto, 2018). This means that value that it attributes to an action depends on how much that

action actually led to the shelter during the training period. Since uninterrupted practice-run action sequences are possible only in condition 1, edge-vector actions are able to rack up high value faster than the meandering actions leading around the trip wire in condition 3. In line with this explanation, an on-policy variant of Q-learning (SARSA) with practice runs behaves similarly to the SR here: it also often fails to find routes in condition 3 but not condition 1 (Fig. 4.4c). Thus, the pattern of exploration we observe in mice - slow meandering exploration punctuated by rapid edge- and shelter- directed runs - can explain why an on-policy learner would learn edge-vector runs in condition 1 but fail to learn a route in condition 3.

Behavior with the partial trip wire is matched with state-action abstraction

Our final condition (condition 4) mimics the optogenetics experiment in Figure 4. This partial trip wire blocks edge-vector runs from the threat zone itself but not from other, nearby locations. Unlike in the previous condition, mice were perfectly able to learn direct edge-vector escapes here. The gradual-learning RL agents, on the other hand, all executed tortuous routes around both the trip wire and the obstacle (Q-learning: 98% tortuous routes; SR: 90% tortuous routes; MB-G: 81% tortuous routes, 17% homing vectors; Fig. 4.3a, Fig. 4.4a). To match mouse behavior on both condition 3 and 4, an RL agent would need to run through the line where the trip wire was during training, instead of taking a step-by-step route around it. In addition, it would have to infer the availability this direct edge-vector route based on nearby but non-identical practice runs during the training phase.

We reasoned that an agent with a coarse-grained state space could possess these features. We first tried implementing Q-learning with a coarse-grained state representation designed to promote spatial generalization (tile coding; Sutton, 1995). This agent's behavior, however, was not substantively different from tabular Q-learning (98% tortuous routes; Fig. 4.4d). Next, we tried a more targeted state-action abstraction protocol akin to hierarchical reinforcement learning. We divided the state space into groupings of grid squares (e.g. the shelter area, the left obstacle edge area) and the action space into vectors connecting those regions (Fig. 4.2c). (Note that we could have used a more sophisticated state-action abstraction scheme such as the options framework (Sutton et al., 1999) but found this to be the most direct solution to condition 4). This Hierarchical State Space (HSS) Q-learning agent explores using the same random walk policy on the full-resolution training map, but updates its controller only with respect to transitions between the high-level regions. As expected, this agent was able to learn edge-vector escapes even with

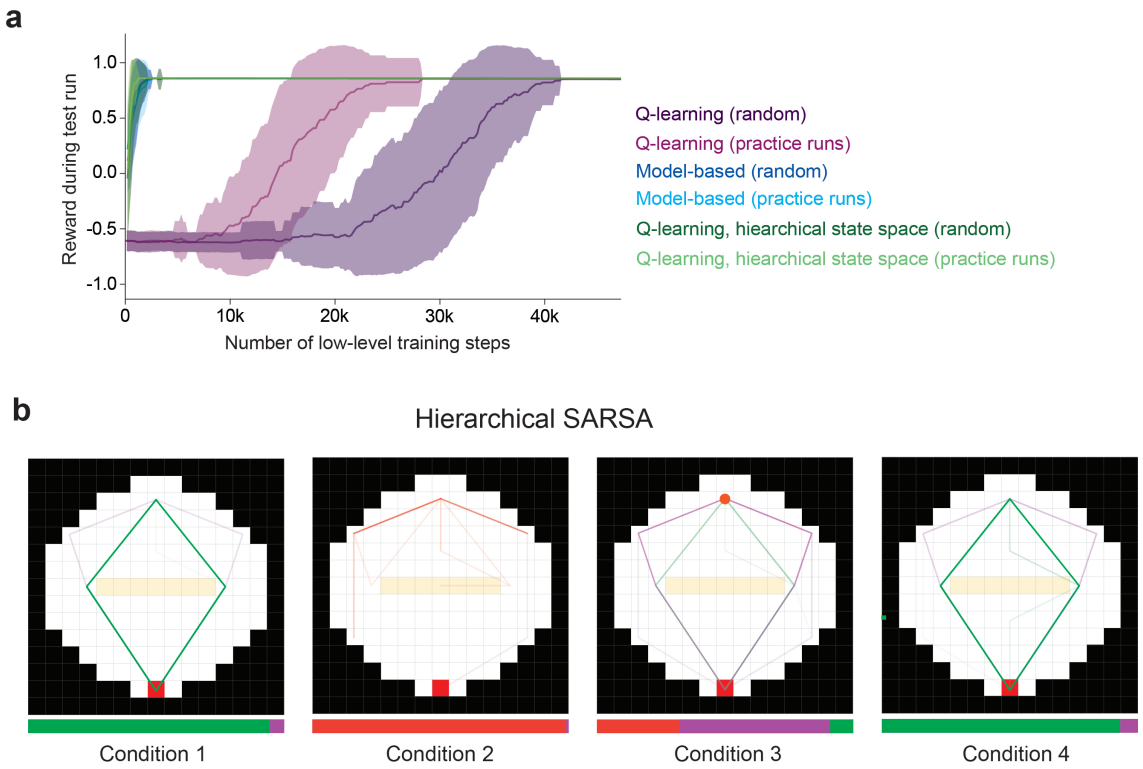

Figure 4.5: *Hierarchical agent learning speed and behavior*

(a) Learning curves plot the amount of reward received from a trial run in the test map (going from threat zone to shelter) over number of training steps. For the learning curves, there is no pre-test phase. The trial terminates when the agent reaches the shelter or at a maximum number of 50 steps. Since negative reward is given at each non-reward state, the minimum reward is approximately -0.5. **(b)** Results with the hierarchical state-space agent using the SARSA algorithm, with practice runs during exploration. Assuming a homing-vector agent takes over when this agent fails to come up with a route, this agent qualitatively matches mouse behavior on all four conditions: edge vectors in condition 1 and 4 and failure to learn a route in conditions 2 and 3. The orange dot in condition 3 indicates selecting an invalid action.

the partial trip wire in place (94% edge vectors; Fig. 4.3b, Fig. 4.5b). Notably, the HSS agent can learn a valuable 'threat area to obstacle edge area' action without ever having taken that action from the exact grid cell where the escape is triggered. These high-level actions also better match the smooth, biphasic escape trajectories we see in mice and generate a much faster learning profile (Supp. Fig. 4.5a). In addition, the regional state representation fits nicely with our finding that mice use a spatially defined 'subgoal initiation set' from ch. 3.

To summarize, the vanilla RL agents we tested were not effective at matching mouse behavior across more than one or two experimental conditions. However, the following principles did allow agents to match mouse behavior across multiple

conditions:

1. The agent includes a gradual-learning system (e.g. Q-learning, SR, MB-G; condition 1)

2. This system does not fully separate spatial and reward learning (e.g. Q-learning; condition 2)

3. The agent experiences non-uniform exploration, with rapid and direct practice runs toward the obstacle edges and shelter, and learns on-policy (condition 3).

4. This system abstracts over regions of space and the actions connecting those regions (e.g. HSS Q-learning; condition 4).

5. In addition to the gradual-learning system, the agent has an immediate-learning system (e.g. MB-I) that comes online when the gradual learner has no valuable action (condition 2-3).

Having defined these five key computational principles we then built an agent possessing all of these properties. This agent includes a gradual learning system that directly learns action values in an on-policy manner (i.e. the SARSA algorithm), within the high-level state-action space introduced above. The agent performs practice runs during exploration, and we assume that it switches to a default MB-I agent in conditions with high failure rates. This agent is able to qualitatively match mouse behavior on all four conditions, executing persistent edge vectors in conditions 1 and 4 and frequently failing to escape in conditions 2 and 3 (Supp. Fig. **??**b).

4.3 Interim discussion

To make the implications of our behavioral results more precise at a computational level, we performed reinforcement learning modelling of four key behavioral and optogenetic experimental conditions. These simulation experiments took place in a simple grid world environment in which the agent experienced its environment as a series of discretized location indices and could only select from a short list of possible actions. Upgrading the simulation environment to more realistically represent the mouse's experience is an interesting future direction (as in Banino et al., 2018). Here, we instead capitalized on the simplicity of the simulation in order to extract a set of basic principles underlying mouse behavior in the obstacle removal experiment.

First, we found that practically any model that updates gradually - be it model-free or model-based - can match the persistent edge-vector escape result. Our initial

assumption upon observing this behavioral phenomenon was that mice must be using a model-free response strategy. However, over time we realized that this assumption was derived not from any fundamental property of reinforcement learning but rather from the association in the literature between gradual updating and model-free action reinforcement (see introduction). Thus, this first result serves to demonstrate that gradually updated, model-based algorithms are *a priori* perfectly plausible.

Second, we found that mice exhibit differing levels of flexibility in different conditions and are thus best modelled through a dual-system agent. The agent we used included one system that updates a policy gradually and another that learns much more rapidly (at greater computational cost), which comes online whenever the first system fails to produce a valuable action. We assessed 'failure' here in a largely qualitative fashion, by noting when a large proportion of escape trajectories failed to reach the shelter. Other work offers more detailed models of how a dual-system agent could arbitrate between its two options. A promising example is to calculate the level of certainty in each system and select whichever system reports a higher likelihood of achieving a rewarding action (Daw et al., 2005).

To model the rapid learner, we simply used the immediate-learning model-based system (MB-I). However, there is no reason that the rapid learner needs to be a classical model-based system. One appealing alternative is the homing-vector instinct. In this case, the agent could have a hardwired policy of running directly toward a recently visited shelter. This system would produce the exact same result as the MB-I agent, namely homing vectors in condition 2 and 3. Moreover, it better corresponds to known navigation strategies in rodent escape behavior (Maaswinkel and Whishaw, 1999; Vale et al., 2017) and to our results on homing-vector escape responses with the obstacle (ch. 2).

The third condition modelled the laser trip wire. With unlimited uniform exploration, the RL models found valid but convoluted escape routes around the trip wire. However, with a limited exploration period punctuated with practice edge-vector sequences, on-policy SR and SARSA agents learned escape routes in condition 1 but not in condition 3. Through the logic of the dual-system agent described above, this agent therefore invokes the backup homing-vector policy, mirroring mouse behavior. On-policy learning reflects a conservative learning strategy. This kind of agent only attributes value to an action if it actually led to future reward or another valuable action during the exploration period. It is therefore less likely to learn a valid escape route unless the right kind of practice has occurred.

In addition, the necessity of implementing practice runs here supports the notion that non-uniform exploratory paths are a crucial factor in modelling mice's spatial learning capabilities (McNamee et al., 2021). This makes sense given the highly

structured nature of mouse exploration, with runs to the shelter and obstacle edges being much more rapid and direct than paths in the center and perimeter. Still, our results on condition 3 are only an initial foray into the role of non-uniform exploration in navigation models. It would be interesting to run an in-depth investigation of how the exploration statistics we see in mice, such as the propensity to explore predominantly around the perimeter of the environment and along the obstacle, affect learning. Another promising approach would be to allow the agents to occasionally exploit their learned policies during exploration; this is a standard procedure in RL. Finally, it would be informative to test which RL algorithms are able to match mouse behavior under the constraint that their exploration pattern follows the exact sequence of steps that we recorded in mice.

Although these agents matched mouse escape trajectories in condition 3, one remaining difference from biological learning is the number of runs needed for learning. Mice required 1-2 runs for their so-called "gradual learning" system to learn the edge-vector route, while Q-learning and SR agents took tens of practice runs. These agents would therefore need to be at least an order of magnitude faster in order to be biologically plausible. One possibility is to construct a value function through a more data-efficient, model-based learning algorithm than the purely model-free updating mechanisms we used here (Sutton, 1991; Russek et al., 2017). Another possibility is to simply imbue certain actions (e.g. running toward salient objects) with a very high learning rate (Barto et al., 2004). A final, compatible option is to use a high-level representation of states and actions (e.g. "go from shelter area to obstacle edge" instead of "go north 10 cm") to speed up learning dramatically (Sutton et al., 1999).

Indeed, agents that decompose the arena into high-level regions and actions (e.g. a "threat-area-to-obstacle-edge" action) not only learned on a rapid timescale but they also matched mice's capacity for spatial generalization. Unlike "flat" agents, operating at the level of individual grid-world states, this agent could execute edge-vector escapes after practicing nearby but non-identical routes.

Hierarchical representations are known to allow for orders-of-magnitude increases in time and memory efficiency for planning, at the expense of overlooking routes that do not fit the agent's high-level spatial representation (Tomov et al., 2020). This nicely summarizes our intuitions about how and why mice memorize and execute subgoal escapes even after the direct homing-vector shortcut has opened up. The hierarchical state-action space also provides a straightforward explanation for our finding that subgoal escapes were selected based on by spatial rules: the initiation set could correspond to a spatial region from which mice learned a valuable "go to obstacle edge" action. How animals might cluster states within their environments into these

regions remains an interesting, open question (Solway et al., 2014; Stachenfeld et al., 2017; Tomov et al., 2020).

Integrating a hierarchical state space into Q-learning and SARSA is an example of a sophisticated spatial representation housed within a model-free learning algorithm. This illustrates a disconnect between "map-based" and "model-based" methods, which are often conflated into one joint concept (Spiers and Gilbert, 2015; Geerts et al., 2020). Here we are invoking a spatial *map* to define states and actions. However, we do not need to invoke a *model-based search* through that map to uncover routes. While we have not ruled out model-based search, we did find that model-free caching of state-action values within a hierarchical spatial map is perfectly compatible with mouse escape trajectories.

Contributions

Sebastian Lee engineered the RL environment used to run the simulations, and we collaborated on developing the algorithms to model escape behavior.

Chapter 5

General discussion

5.1 Conceptual overview

In this project, we examined at how mice deal with multistep structure in a new environment with an obstacle and a goal, on the basis of 20 minutes of self-motivated exploration. We had to look beyond individual 'action-reinforcement' or 'map-based' strategies as they are traditionally described and instead consider a hybrid approach of action-driven mapping. On the mapping side, mice dynamically selected routes to subgoal locations in an allocentric reference frame, they displayed spatial generalization and abstraction, and they learned rapidly. However, to incorporate information about subgoals into their map, they relied on a process akin to action-driven learning. In particular, there was slow updating after the obstacle was removed, there was no learning without a reward, and the subgoal memory was triggered by instinctive edge-vector runs. In brief, we found that mice use running actions to learn subgoals within a hierarchical mental map of the environment. While this work represents an initial case study in one species, we believe that this newly described strategy - using actions to learn a map of subgoals - represents an important part of the cognitive toolkit for identifying useful locations in structured environments.

5.2 Methodological innovations

The obstacle removal experiment

Prior work on multi-step spatial reasoning has focused on repeatedly placing an animal at the start of a constrained maze environment and testing how it learns to reach a food reward while minimizing erroneous turns (Tolman and Honzik, 1930; Sharma et al., 2010). These assays have several advantages: they assess long-term memory over multiple sessions/days; they induce stereotyped paths across animals;

and they rely on a particularly stable and controllable source of motivation, i.e., hunger. However, they also leave out several key aspects of spatial reasoning. For one, they disregard how animals explore and rapidly compute routes in natural environments, which include both open space (allowing a much wider range of possible actions) and obstacles (necessitating multi-step reasoning). In addition, they lack a stimulus that can trigger immediate, goal-directed behavior. Thus, it is unclear if the animal's 'errors' reflect a lack of understanding or merely a decision not to exploit a known route to the goal. Our assay - escape to shelter in the presence of an obstacle, during a mouse's first experience outside of its home cage - complements previous work on maze learning by incorporating these elements into the study of multi-step navigation.

Studying active learning with neural stimulation

In line with the enactivist framework (Ward et al., 2017), we believe that learning is best described as a dynamic interplay between the sensory stimuli we perceive, the internal neuro-cognitive mechanisms we use to update our behavior, and the actions we take in order to receive the next sensory stimulus. Thus, understanding the action policies that animals use in order to direct a learning process is just as valid of a scientific aim as investigating the relevant sensory cues or neural mechanisms. Previous studies using optogenetics, however, have always focused on the role of the stimulated or inhibited neuronal population itself (Kim et al., 2017). For example, a study might inhibit the hippocampus in order to test whether spatial memory formation is impaired when that region cannot function as normal (Siegle and Wilson, 2014). Here we have identified a novel use case for closed-loop neural stimulation: testing the causal role of particular actions in a learning process.

To test the causal role of an action, we needed to approximate an answer to the question: would the mouse have learned a subgoal if it had not decided to execute a practice edge-vector run? We could have easily just modulated mice's propensity for practice runs by modifying the environment (e.g. replacing the wall obstacle with a hole obstacle as in ch. 2) or the mouse's arousal level (at the extreme, a sleepy mouse might execute zero practice runs). However, these manipulations are too indirect, as they modulate not only practice runs but also a variety confounding factors such as the overall amount of exploration and the structure of the environment. Neural stimulation allowed us to specifically impair one class of movements and thus provided a uniquely direct way to test the causal link between practice runs and subgoal learning.

5.3 Key limitations

The behavioral assay

In our behavioral assays, we were able to examine escape routes across different amounts of experience with an obstacle (0 - 20 min), different light levels (sufficiently lit or completely dark), and different types of obstacle (a wall and a hole). However, our assay takes place in a restricted spatial scale (a 1 m environment) and time scale (learning over 0-30 min). Thus, our findings may not apply to large or tiny environment or to highly experienced mice.

The RL models

In machine learning, and reinforcement learning in particular, models can be highly sensitive to hyper-parameters. Different hyper-parameter configurations can lead to different behaviour even for the same algorithm. This makes it a challenge to meaningfully compare the algorithms. For instance, the SR model exhibits its 'classic' behavior of finding routes to the goal despite the reward not being present during training (condition 2 in ch. 4) only if its reward vector is initialized to all zeros. A non-zero reward initialization would have caused the SR model to fail to find routes to shelter in condition 2 until the agent entered and exited the shelter multiple times.

In general, due to the possibility of hyper-parameters affecting behavior and the non-exhaustive list of algorithms we tried, we are careful not to fall into the trap of labelling any particular algorithm as 'the best match' to mouse behavior. We do not even find it sensible to decide whether mice are using model-free or model-based approaches for subgoal learning. All we can do is to investigate the causes of behavior across a variety of algorithms in order to extract the overarching, high-level computational principles that we list in ch. 4.

The interpretation

The key insight from this study is that mice use practice runs to learn subgoals, which are embedded in a hierarchical spatial map of the environment. However, the exact role of practice runs is not entirely clear. The way we model this process in ch. 4 is with a preformed high-level map of states and actions, presumably generated in the mouse through the classical process of observing the environment and embedding its observations into a cognitive map. There, practice runs simply serve to add value to this pre-existing option. An alternative, appealing explanation is that the practice runs actually help to generate the hierarchical state space. In this case, these actions

would be used not only to attribute value to edge-vector runs but also to establish the subgoal action as a high-level action in the first place (Barto et al., 2004).

Similarly, it remains unclear what is going on in the trials where mice revert to the homing vector escape after having memorized a subgoal. There are several possibilities. One is that their internal map of the environment switches to one in which there is no obstacle. Second, the value of the edge-vector run could decrease or the value or the homing-vector run could increase. Third, mice could be switching to a one-step navigation strategy such as visual guidance or path integration that was otherwise blocked when the obstacle was present.

5.4 Future directions

One big remaining question is to define the scope of action-driven subgoal mapping. First, is the persistent subgoal strategy specific to short-term learning of escape behavior? Reactions to imminent threats tend to be less deliberate and flexible than less urgent behaviors such as reward seeking (Mobbs et al., 2020); this raises the possibility that the persistent usage of memorized subgoals could be specific to escape. However, studies have shown that rats (Grieves and Dudchenko, 2013) and mice (ch. 2) tend to prefer familiar, roundabout routes over new shortcut routes even during reward seeking. One way to expand on this finding would be to test whether learning with an obstacle during reward seeking translates to subgoal memorization in the context of escape, and vice versa. We predict that subgoal memorization will turn out to be a general learning strategy across task modalities.

Second, does action-driven mapping extend across species to human behavior? Clearly, an adult human in a small, well-lit room would not need to run to an obstacle edge in order to learn its location. However, humans may use analogous strategies in other scenarios. For example, De Cothi et al., 2022 showed that in a virtual environment with changing obstacles and a limited visual field, humans tend to update their spatial behavior gradually based on the paths they take, rather than immediately upon observing an obstacle. It would be interesting to adapt this assay to allow for both practice runs and slow, observational exploration and to examine the role of practice runs in human navigational learning. In addition, as observed here in mice, humans naturally decompose multi-step tasks into high-level state and action representations (Ribas-Fernandes et al., 2011; Solway et al., 2014). Indeed, Tomov et al., 2020 showed that human participants preferred paths that included sub-paths experienced during training, even if a shorter route was available. Another possibility, in both mice and humans, is that action-driven mapping may

serve as a learning stage that lays the foundation for practice-independent cognitive mapping (Piaget, 1955). This idea is similar to the role of babbling in speech learning (Petitto and Marentette, 1991). Along these lines, children prefer familiar multi-step routes to novel shortcuts until they are 4 years old (Hazen et al., 1978). In mice, we could investigate this by testing whether practice runs cease to be necessary for subgoal learning if the mouse is an experienced navigator of structured environments. Overall, we find it highly plausible that action-driven mapping forms a part of the human cognitive repertoire, both in the navigation setting and beyond.

Another direction is to delve deeper into the nuances of the subgoal-memorization algorithm. The problem of segmenting the environment into regions and waypoints, which mice appear to achieve within minutes of entering the arena, is of particular interest. Future experiments using environments with more complex structure could shed light on this. As a first step, one could present the mouse with a series of two obstacles and ask whether a sub-sub-goal is learned in the same manner as a subgoal. In addition, the behavioral-economics aspect of the work could be expanded upon. In particular, it would be interesting to better understand how mice trade off between a familiar route and a short one when deciding between the subgoal route and the homing-vector shortcut. One could make the memorized subgoal route more inefficient than it is with the linear obstacle, for instance with a semicircular obstacle that would require the mouse to run to a subgoal that is back and away from the shelter. Alternatively, one could make the threat stimulus more threatening, thereby increasing the penalty of using an inefficient subgoal route. If either of these manipulations abolish post-obstacle-removal edge-vector routes, this would suggest that mice are actively comparing the value of the subgoal and homing-vector routes during route selection.

Finally, an astute reader may have noticed that this neuroscience thesis is not really about the brain. While the neural implementation of subgoal learning remains unaddressed, our results open the door for future work elucidating the network of motor and spatial nuclei that implement subgoal memorization. One advantage we have provided for neuroscientists is a thorough characterization of this behavior; using well characterized, spontaneous behaviors for neuroscience promises to uncover aspects of brain function that could have been suppressed in traditional laboratory tasks (Krakauer et al., 2017; Mobbs et al., 2018; Datta et al., 2019). In addition, escape behavior's rapid learning profile and reliable, stimulus-locked routes make it particularly amenable to systems neuroscience techniques (although the limited number of trials present a challenge). For instance, recording and inhibiting regions putatively involved in the subgoal computation (e.g. entorhinal cortex) during edge-vector runs could be one interesting direction. Another future direction would

be to investigate interactions between the hippocampal map-based planning and the striatal action-reinforcement systems, often believed to be competing for control of behavior; the action-driven mapping strategy we have uncovered points to a tighter coordination between these regions than previously thought.

Chapter 6

Methods

Note that the experimental hardware and procedure was improved between the experiments performed in chapter 2 and chapter 3. When they differ, these methods will be presented separately in each section. If you are aiming to use this project's methodology, I would advise following the more streamlined methods of chapter 3.

6.1 Animals

All experiments were performed under the UK Animals (Scientific Procedures) Act of 1986 (PPL 70/7652) after local ethical approval by the Sainsbury Wellcome Centre Animal Welfare Ethical Review Body. We used 172 singly housed (starting from 8 weeks old), male, 8-12-week-old C57BL/6J mice (Charles River Laboratories) during the light phase of the 12-h light/dark cycle. Mice were housed at 22°C and in 55% relative humidity with ad libitum access to food and water.

6.1.0.1 Re-use over multiple sessions

For the main exploration + escape experiments in ch. 2, data come from the mice's first-ever behavioral session. However, in the following experiments, mice had experienced a session 5-7 days prior: obstacle removal with the expanded threat zone, and exploration in an environment with no shelter.

For the food-seeking experiments, each mouse experienced a session with no obstacle and then a second session using the obstacle removal procedure.

For the exploration + escape experiments in implanted mice: four of the eight mice were naive, and this was their first behavioral session of any sort. The remaining four mice had experienced a previous session 5-7 days prior. Their previous session was not allowed to be the same exact experiment as the second session but was otherwise selected randomly. For the place-preference experiment and laser-power

test, mice were randomly selected from those that had already experienced their behavioral sessions.

6.1.0.2 Exclusion criteria

Data from mice with zero escapes in the session were excluded (7% of sessions). This is due to three reasons: remaining in the shelter, not responding to the threat stimulus, or climbing down from the arena. In the experiments with implanted mice, a replacement session was performed 5-7 days later in a randomly selected mouse.

6.2 Behavioral Assays

Environment	Light	Mice	Prior trials	Pre-threat
Open field	On	10	None	10 min
Obstacle	On	24	None	10 min
Hole obstacle	On	8	None	10 min
Open field	Dark	14	None	10 min
Obstacle	Dark	14	None	10 min
Obstacle	10 min then dark	14	None	10 min
Obstacle	20 min then dark	14	3 trials	20 min
Acute obstacle removal 1	On	10	3 trials	20 min
CORE 1 (CORE-3B)	On	10	4 trials	20 min
CORE 2 (CORE-ZB)	On	10	None	20 min
Narrow corridors	On	10	1 session	20 min
CORE 3 (move the shelter)	On	10	None	20 min
Acute obstacle removal 2	On	10	None	10 min
CORE 4 (no shelter at first)	On	10	None	20 min
CORE 5 (additional barrier)	On	10	None	20 min
Open field (no shelter)	On	6	1 session	n/a
Obstacle (no shelter)	On	6	1 session	n/a
Hole obstacle (no shelter)	On	7	1 session	n/a
Open field (food seeking)	On	6	None	20 min
CORE 6 (food seeking)	On	6	1 session	20 min

Table 6.1: Experiments in chapter 2. CORE stands for chronic obstacle removal experiment. 3B stands for three baseline escape trials. ZB stands for zero baseline escape trials.

ID	Experimental setup	M2 stimulation	Mice
1	Obstacle removal	Injection/implantation, no stim	8
2	Obstacle removal	Stop edge-vector runs	8
3	Open field–no obstacle	Injection/implantation, no stim	8
4	Obstacle removal	Stop edge-vector runs after two	8
5	Obstacle removal	Stop edge-to-shelter runs	8
6	Obstacle removal	Stop threat-area-to-left-side runs	8
7	Obstacle removal–threat zone II	None	8
8	Obstacle removal–threat zone III	None	8
9	Two-chamber place preference	Paired with one chamber	8
10	Open field–no obstacle or shelter	Test effects of three laser powers	4

Table 6.2: All experiments in chapter 3

Equipment

6.2.0.1 Platforms and shelter

Experiments took place on an elevated white 5-mm-thick acrylic circular platform 92 cm in diameter. Note that the room was not totally sonically insulated and that neither the black surround nor the overhead illumination was circularly symmetric; these asymmetries could all provide spatial orientation cues. The platform and shelter were cleaned with 70% ethanol after each session.

Chapter 2: The hole obstacle consisted of a 50 cm long ×10 cm wide rectangular hole in the center of the platform. The modified platform with two narrow corridors consisted of the original platform with the obstacle, plus six additional panels. Four of these panels were 50 cm long ×12.5 cm tall ×0.5 cm thick, and two were 12.5 cm long ×12.5 cm tall ×0.5 cm thick. Together, they formed two corridors that were 50 cm long ×7.5 cm wide and were at 65°and 115°angles relative to the axis of the central obstacle. The interior panels forming the corridor were made of red acrylic so that the IR camera could see through them; all other panels were made of white acrylic. The shelter was a 10 cm cube of transparent red acrylic (opaque to the mouse). It included a mouse-hole-shaped entrance at the front and additional 2.5 cm tall square of red acrylic on top in order to prevent the mice from climbing on top.

Chapter 3: The platform had a 50 ×10 cm rectangular gap in its center. For conditions with no obstacle (all post-exploration escapes and the entirety of experiments 3 and 10), this was filled with a 50 ×10 cm white 5-mm-thick acrylic rectangular panel. For conditions with the obstacle present, this was filled with an identical panel that, attached to an obstacle: a 50 cm long ×12.5 cm tall ×5 mm thick white acrylic panel. The shelter was 20 cm wide ×10 cm deep ×15 cm tall and made of

5-mm-thick transparent red acrylic, which is opaque to the mouse but transparent to an infrared-detecting camera. The shelter had a 9cm-wide entrance at the front, which extended up to the top of the shelter and then 5 cm along its ceiling; this extension of the opening allowed the optic fiber, which was plugged into the mouse's head, to enter the shelter without twisting or giving resistive force.

6.2.0.2 Additional equipment

Chapter 2: The platform was surrounded by a black, square, plastic surrounding. A projector screen was located above the platform. The platform was illuminated with 4 infrared lights (S8100-45-A-IR, Fuloon). Experiments done "in the dark" were performed in complete darkness (0.00 cd $m^{-}2$ of visible light). At this light level, mice did not react to rapidly waving a hand in front of them, which is perceived as highly threatening when light is available. For all other experiments, light was projected onto the screen at 5.2 cd $m^{-}2$ using a projector (PF1000U, LG).

Chapter 3: The elevated platform was located in a 160 cm wide $\times$190 cm tall $\times$165 cm deep sound-proof box. A square-shaped projector screen (Xerox) was located above the platform. This screen was illuminated in uniform, gray light at 5.2 cd $m^{-}2$ using a projector (BenQ). Behavioral session were recorded with an overhead GigE camera (Basler) with a near-infrared selective filter, at 40 frames per second. Six infrared LED illuminators (TV6700, Abus) distributed above the platform illuminated it for infrared video recording.

Escape behavior

6.2.0.3 Data acquisition

All signals and stimuli, including each camera frame, were triggered and synchronized using hardware-time signals controlled with a PCIe-6351 and USB-6343 input/output board (National Instruments), operating at 10 kHz.

Chapter 2: Data acquisition was performed using custom software written in LabVIEW (2015 64-bit, National Instruments) by Kostas Bestios. To verify correct synchronization, the audio output cable was also fed in parallel to an infrared LED (850nm OLSON PowerStar IR LED), which flashed in synch with sound presentation.

Chapter 3: Data acquisition was performed using custom software in the visual reactive programming language Bonsai (Lopes et al., 2015). All signals and stimuli, including each camera frame, were triggered and synchronized using hardware-time signals controlled with a PCIe-6351 and USB-6343 input/output board (National Instruments), operating at 10 kHz.

6.2.0.4 Threat stimulus presentation

Threat stimuli were loud (84 dB), unexpected crashing sounds played from a speaker located 1 m above the center of the platform (Supplementary Audio 1 and 2). Sounds ('smashing' and 'crackling fireplace') were downloaded from soundbible.com. They were then edited using Audacity 2.3.0, such that they were 1.5 sec long and continuously loud. Stimuli alternated between the 'smashing' sound and the 'crackling' sound each trial, to prevent stimulus habituation. The volume was increased by 2 dB after time a stimulus failed to elicit an escape, up to a maximum of 88 dB. When a threat trial began, the stimuli repeated until the mouse reached the shelter or for a maximum of 9 secs. Stimuli were played from the PC, through an amplifier (TOPAZ AM10, Cambridge Audio) and speaker (L60, Pettersson). Experiments were terminated after one hour.

Chapter 2: Stimulus delivery was controlled with software custom-written in LabVIEW (2015 64-bit, National Instruments). Stimuli were triggered manually, when the mouse had been in the threat zone (demarcated on the live video) for at least one second and was facing in approximately the opposite direction from the shelter. Mice varied in how many trials they performed in each experiment. We thus limited analysis to the first three escapes in each condition (more than 50% of mice completed at least three escapes in all experiments).

Chapter 3: Stimulus delivery was controlled automatically with software custom-written in Bonsai. The criteria for activating a threat stimulus were 1) the mouse is currently in the threat zone; 2) the mouse was in the threat zone 1.5 seconds ago; 3) the mouse is moving away from the shelter at >5 cm $s^{-}1$ (this ensures that escape runs are always initiated after the stimulus onset); 4) the most recent threat stimulus occurred >45 sec ago. We limited analysis to the first six escapes in each condition (more than 50% of mice completed at least six escapes in all experiments).

6.2.0.5 Environmental manipulations

Obstacle removal, chapter 2: For experiments in which the obstacle appears or disappears, this was done by digitally triggering a custom-made pneumatic tubing system (time to raise or lower the obstacle was 100 ms). In the acute obstacle removal experiment, obstacle removal was triggered simultaneously with the stimulus onset. In chronic obstacle removal experiments, this was triggered while the mouse was in the shelter. Obstacle removal makes a whooshing sound (63 dB measured at the shelter) and usually triggers a startle response.

Obstacle removal, chapter 3: After 20 minutes of exploration were complete, as soon as the mouse entered the shelter, the experimenter quickly and quietly removed

the central panel containing the obstacle and replaced it with the flat 50x10 cm panel. Mice were then allowed to freely explore and (and trigger escapes) in this open-field platform.

Adding bedding to the platform: Bedding from the mouse's home cage was added to the platform in order to encourage exploration, rather than staying in the shelter throughout the experiment. One pinch (1 gram) of bedding was added to the center of the threat zone in all experiments when either of the following two criteria was met: 1) The mouse did not leave the shelter for five minutes; or 2) The mouse did not enter the threat zone for ten minutes.

Food-seeking behavior

6.2.0.6 Training

Mice were food restricted to 85% of their baseline weight. Training and pretraining were done in a 60cm ×15cm rectangular arena, with a shelter on one side and a reward port on the other side. The reward consisted of a 7-μL drop of condensed milk (diluted 1:1 with water) delivered through the spout. For pretraining, during which the mouse learned to associate the metal spout with reward, 100 drops of milk were manually triggered and then collected by the mouse, with a minimum interval of 1 minute between each drop. They were then trained in five, 1-hour sessions to approach and lick a metal spout in response to a 9-second, 10-kHz, 72-dB tone. Tone stimuli were triggered manually once per minute. Licking the spout while the tone was on resulted in reward. After reward delivery, there was a 5-second refractory period; thus, mice could trigger at most two rewards during the 9-second tone. On the last two day of training, the tone duration was reduced to 4.5 seconds after 30 minutes. Licks were registered with a capacitive touch sensor (Adafruit MPR121), connected to a microcontroller board (Arduino Uno). The milk was delivered through a peristaltic pump (Campden Instruments 80204E), connected to the same microcontroller.

6.2.0.7 Navigation assay

To test food-seeking paths, mice had two sessions. The first session was in the platform with no obstacle, the shelter on one side, and a lick port on the opposite end of the platform. They received practice trials of tone and milk, initially mostly when they were already near the lick port. After 20 minutes, test trials were initiated when the mouse was on the opposite side from the lick port, and these data were used for analysis. The second session followed the same protocol. However, in this session

the obstacle was initially present, and then was removed after 20 minutes while the mouse was in the shelter. Mice performed more trials than with the escape behavior, so here we examined trajectories from the first nine successful trials (greater than 50% of mice completed at least nine trials).

6.3 Neural manipulations

Surgical procedures

6.3.0.1 Viral injection and optic fiber implantation

Mice were anaesthetized with isoflurane (5%) and secured on a stereotaxic frame (Kopf Instruments). Meloxicam was administered subcutaneously for analgesia. Isoflurane (1.5-2.5% in oxygen, 1 l min^{-1}) was used to maintain anesthesia. Craniotomies were made using a 0.7 mm burr (Meisinger) on a micromotor drill (L12M, Osada), and coordinates were measured from bregma. Viral vectors were delivered using pulled glass pipettes (10 μl Wiretrol II pulled with a Sutter-97) and an injection system coupled to a hydraulic micromanipulator (Narishige), at approximately 100 nl $min^{-}1$. Implants were affixed using light-cured dental cement (3M) and the surgical wound was closed using surgical glue (Vetbond).

Mice were injected with 120 nL of AAV9/CamKIIa-ChR2-EGFP in the right, anterior premotor cortex (AP: 2.4 mm, ML: 1.0 mm, DV: -0.75 mm relative to brain surface) and implanted with a magnetic fiber-optic cannula directly above the viral injection (DV: -0.5 mm) (MFC_200/245-0.37_1.5mm_SMR_FLT, Doric). All behavioral sessions took place 2-4 weeks after the injection/implantation.

6.3.0.2 Histology

To confirm injection and implantation sites, mice were terminally anaesthetized by pentobarbital injection and decapitated for brain extraction. The brains were left in 4% PFA overnight at 4°C. 100um-thick coronal slices were acquired using a standard vibratome (Leica). The sections were then counter-stained with 4',6-diamidino-2-phenylindole (DAPI; 3 μM in PBS), and mounted on slides in SlowFade Gold antifade mountant (Thermo Fisher, S36936) before imaging (Zeiss Axio Imager 2). Histological slice images were registered to the Allen Mouse Brain Atlas (Allen Institute for Brain Science, 2015) using SHARP-Track (Shamash et al., 2018), to find the fiber tip coordinates.

Closed-loop optogenetic stimulation

Laser stimuli consisted of 2-sec, 20-HZ square-wave pulses at 30 mW (duty cycle 50%, so 15 mW average power over the two seconds) supplied by a 473-nm laser (Stradus 472, Vortran). For the experiment blocking edge-to-shelter runs, we instead used 5-sec pulses. The laser was controlled by an analog signal from our input/output board into the laser control box. At the beginning of each session, the mouse was placed in an open 10x10 cm box and the magnetic fiber-optic cannula was manually attached to a fiber-optic cable (MFP_200/230/900_0.37_1.3m_FC-SMC, Doric). A rotary joint (Doric) was used to prevent the cable from twisting. Finally, the rotary joint was connected to the laser via a 200-μm core patch cable (ThorLabs).

At the beginning of each mouse's first session, the mouse was placed in a 10x10 cm box, and two 2-sec stimuli were applied. If these did not evoke stopping and leftward turning (2/24 mice), then the mouse was assigned to one of the laser-off conditions. During laser-on sessions, the criteria for triggering laser stimuli were: 1) the mouse crosses the 'trip wire'; and 2) the mouse is moving in the 'correct' direction. For blocking edge-vector and edge-to-shelter runs, the direction was determined by a directional speed threshold: moving toward the shelter area (i.e., south) at $>$ 5 cm sec^{-}1. For blocking threat-zone-to-left-side runs, mice had to be moving toward the left side (i.e., west) at $>$ 5 cm sec^{-}1. These speed thresholds are low enough to be effective at catching all cases in which the mouse crosses the trip wire in a particular direction. These criteria were computed online using the Bonsai software described in the previous section. The laser pulses were emitted with a delay of 300-400 ms after being triggered. Up to three subsequent 2-sec pulses were triggered manually if the mouse continued moving forward.

Mice usually took 1-3 minutes to enter the shelter for the first time, and these first minute(s) of exploration typically contains relatively vigorous running. Since subgoal learning does not occur in this setting without a shelter in the environment (Shamash et al., 2021), the laser-on condition was initiated only after the mouse entered the shelter for the first time.

Place preference assay

Mice were hooked up to the optic fiber as described above and placed into a two-chamber place-preference arena. The arena was made of 5-mm-thick transparent red acrylic (opaque to the mouse) and consisted of two 18 cm long exttimes 18 cm wide exttimes 18 cm tall chambers connected by a 8cm-long opening. To make the chambers visually distinguishable, one chamber had a 10x10 cm x-shaped white acrylic piece affixed to its back wall and the other had a filled-in, 10cm-diameter

circular white acrylic piece affixed to its back wall. The stimulation chamber (left or right) was pseudoramdomly determined before each session, such that both sides ended up with four mice. After a 1-min habituation period, a series of four 2-sec laser stimuli were manually triggered whenever the mouse fully entered the stimulation chamber. A minimum of one minute was given in between each trial, and a total of six stimulation series were delivered. After the last stimulation, one minute was given so that the occupancy data would not be biased by always starting in the stimulation chamber. Then, the next 20 minutes were examined to test for place aversion in the stimulation chamber. This assay is adapted from the conditioned place preference assay (Stamatakis and Stuber, 2012) and the passive place avoidance assay (Schlesinger et al., 1983), such that it matches the conditions of our exploration/escape assay (i.e., to be relevant, place aversion must be elicited during the same session as the laser stimulation, and it must be expressed through biases in occupancy patterns)

6.4 Analysis

All analysis was done using custom software written in Python 3.8 as well as open-source libraries, notably NumPy, OpenCV, Matplotlib and DeepLabCut.

Video tracking

Videos were acquired at 30 frames per second using an overhead camera (acA1300-60gmNIR, Basler) with a near-infrared-selective filter. Video recording was performed with software custom-written in LabVIEW (2015 64-bit, National Instruments). Videos were then fisheye-distortion corrected, aligned onto a common coordinate framework, and visualized with custom Python code using the OpenCV library. We used DeepLabCut (Mathis et al., 2018) to track the mouse from the video, after labelling 1500 frames with 13 body parts: snout, left eye, right eye, left ear, neck, right ear, left upper limb, upper back, right upper limb, left hind limb, lower back, right hind limb, tail base. Post-processing includes removing low-confidence tracking, using a median filter with a width of 7 frames, and applying an affine transformation to the tracked coordinates to match the common coordinate framework.

Trajectory analysis

Calculating position, speed and heading direction: For analysis of escape trajectories and exploration, we used the average of all 13 tracked points, which we found to

be more stable and consistent than any individual point. To calculate speed, we smoothed the raw frame-by-frame speed with a Gaussian filter (σ = 100 ms). To calculate the mouse's body direction, we computed the vector between the lower body (averaging the lower left limb, lower right limb, lower back, and tail base) and the front of the body (averaging the upper left limb, upper right limb, and upper back).

The escape initiation point, chapter 2: is defined as the beginning of a homing run (see below) that goes from inside the threat zone to outside of the threat zone following a threat stimulus. This is computed in the same way for spontaneous homings. We use this this criterion in this section because it allows us to fairly compare spontaneous and stimulus-evoked homings - an important part of the analysis in chapter 2 but not chapter 3.

The escape initiation point, chapter 3: occurs when mice surpass a speed of 20 cm s^{-1}, relative to (i.e., getting closer to) the shelter location. This threshold is high enough to correctly reject non-escape locomotion bouts along the perimeter of the platform but also low enough to identify the beginning of the escape trajectory.

The escape target score: was computed by taking the vector from the mouse's position at escape initiation to its position when it was 10 cm in front of the obstacle. Vectors aimed directly at the shelter received a value of 0; those aimed at the obstacle edge received a value of 1.0; a vector halfway between these would score 0.5; and a vector that points beyond the edge would receive a value greater than 1.0. The formula is:

$$score = \frac{|offset_{HV} - offset_{EV} + offset_{HV-EV}|}{2 * offset_{HV-EV}}$$

Offset$_{HV}$ is the distance from the mouse to where the mouse would be if it took the homing vector; offset$_{EV}$ is the distance from the mouse to where the mouse would be if it took the obstacle edge vector; and offset$_{HV-EV}$ is the distance from the homing vector path to the obstacle edge vector path. The threshold for classifying a trajectory as an edge vector (scores above 0.65) was taken to be the 95th percentile of escapes in the open-field condition. Escapes with scores under 0.65 were designated as homing vectors. When escape trajectories are limited to escapes on the left side, this refers to escapes that are on the left half of the arena when they cross the center of the platform along the vertical (threat-shelter) axis.

Exploratory behavior

Extraction of homing runs and edge-vector runs: Homing runs are continuous turn-and-run movements from the threat area toward the shelter and/or obstacle edges. As in ch. 2, they are extracted by (1) computing the mouse's 'homing speed' (that is,

speed with respect to the shelter or obstacle edges with Gaussian smoothing (/sigma = 0.5 s)) and the mouse's 'angular homing speed' (the rate of change of heading direction with respect to the shelter or obstacle edges); (2) identifying all frames in which the mouse has a homing speed of >15 cm s^{-1} or is turning toward the shelter at an angular speed of >90°per sec; (3) selecting all frames within 1 s of these frames, to include individual frames that might be part of the same homing movement but do not meet the speed criteria; (4) rejecting all frames in which the mouse is not approaching or turning toward an edge or the shelter; and (5) rejecting sequences that take less than one sec or do not decrease the distance to the shelter by at least 20%. Each series of frames that meet these criteria represents one homing run. We limited analysis to he homing runs that started within the threat area. *Edge-vector runs* are homing runs that enter anywhere within the 10-cm-long (along the axis parallel to the obstacle) exttimes 5-cm-wide (along the axis perpendicular to the obstacle) rectangle centered 2.5 cm to the left of the obstacle edge.

Quantification of turning angles: Turning angles that initiated homing runs and escapes were taken as the difference between the mouse's heading direction at the start of the movement (the homing-run or escape initiation point) and the mouse's heading direction after it had traveled 15 cm away from this start location. The start location is when the mouse starts turning toward and/or moving toward the shelter or obstacle edge (see previous subsection). Left turns were defined as negative, and right turns were defined as positive. For predicting escape targets from turn movements, we first extracted all homing runs from the mouse's previous exploration experience. We then identified the homing run(s) most similar to the escape, using three different similarity metrics: the most similar turn angle, the closest starting position, and closest initial heading direction. For each homing run-escape pair, we computed what the escape target would have been if the mouse had turned the same angle that it had turned during the homing run, i.e. if it had repeated the previous egocentric action. Finally, we performed a linear regression between the predicted targets (x) and the actual escape targets (y) to find the proportion of variance (R^2) in escape targets predicted using this assumption that mice repeat previous egocentric turns. For the negative control, we disregarded the homing experience and instead predicted a random turn angle, and then extrapolated that angle to predict an escape target. We repeated this procedure 1000 times to get 1000 R^2 values and took the mean R^2.

Spontaneous exploratory traversals are paths during exploration that start at either end of the platform (within 20 cm the end) and then reach within 10 cm of the central x-axis. Traversals that go along the boundary of the platform (i.e. within 10 cm of the outer perimeter) or take longer than 2 seconds (10 cm/sec) were excluded

from analysis, as these paths contained pausing and looping behavior, hindering the analysis of trajectories.

Amount of exploration: The time spent exploring was computed as the time spent at least 5 cm away from the shelter. The amount of exploration, or distance explored, was the time exploring multiplied by the mouse's speed at each time point. Mice spent 1/3 of the session in the shelter (IQR: 20-52% of the time).

6.4.0.1 Initiation set analysis

Logistic regression: Our logistic regression analysis tests the strength of the linear relationship between each spatial metric and the log odds of performing an edge-vector escape. No regularization penalty was used. The strength of the fit was measured using McFadden's pseudo-R^2: $R^2 = 1 - \frac{LL_{full}}{LL_{null}}$, where LL_{full} is the log likelihood of the logistic regression model fitted with the predictor data and LL_{null} is the log likelihood of the logistic regression fitted with only an intercept and no predictor data. Pseudo-R^2 values of 0.2-0.4 represent "excellent fit"(McFadden, 1977). To test statistical significance of these values, we performed a permutation test, based on the distribution of pseudo-R^2 for the same predictor value, across 10,000 random shuffles of the escape responses (edge vector or homing vector).

Normalizing a metric: To normalize a spatial metric (y, e.g. distance from the center of the arena along the left-right axis) by another metric (x, e.g. distance from the shelter), we computed a linear regression on these variables. We then took the residuals of this prediction ($residual = y - \hat{y}$, where $\hat{y} = slope \times exttimes + offset$) and correlated them with proportion of edge-vector escapes in each bin. This tells us whether, at a given distance from the shelter, there is still a correlation with distance from the center.

Statistical tests

For comparisons between groups, we used a permutation test with the test statistic being the pooled group mean difference. The condition of each mouse (e.g., laser-on vs. laser-off) is randomly shuffled 10,000 times to generate a null distribution and a p-value. We used this test because it combines two advantages: 1) Having the test statistic as the pooled group mean gives weight to each trial rather than collapsing each animal's data into its mean (as in the t-test or the Mann-Whitney test); 2) It is non-parametric and does not assume Gaussian noise (unlike the repeated-measures ANOVA), in line with much of our data. Tests for increases or decreases (e.g., whether exploration decreased due to laser stimulation) were one tailed. The Wilcoxon signed-rank test was used for the place-preference assay to test whether

occupancy in the stimulation chamber was less than 50%. The sample size of the experiments in chapter 3 (n=8 mice) was selected based on a power analysis based on the data from Shamash et al., 2021 and a minimum power of 0.8. Ranges in box plots are limited from the first quartile minus 1.5 exttimes IQR to the third quartile plus 1.5 exttimes IQR. Statistically significant results are indicated in the figures using the convention *n.s.: p>0.05, *: p<0.05, **: p<0.01 and ***: p<0.001.*

6.5 Reinforcement learning simulations

General reinforcement learning setup

Reinforcement learning simulations use the formalism of a Markov Decision Process (MDP) (Sutton and Barto, 2018). An MDP consists of a tuple (S, A, T, R) where S is the set of states; A is the set of possible actions; $T : S \times A- > S'$ is the transition function defining what happens when an action a is taken in state s; $R : S \times A \times S'- > R$ is the reward function, which determines the scalar reward returned by the environment after a given state-action-next-state sequence.

We construct our environment as a 13x13 gridworld. S consists of the set of accessible positions in this map, shown in white in the figures. A, unless stated otherwise, consists of 8 actions (north, northwest, west, southwest, south, southeast, east, northeast). T is a deterministic function that moves the agent one unit in the direction of the action taken. R is a deterministic function in which a reward of 100 is given for entry to the shelter state, and a negative reward of $d(s, s')$ is given for each transition. $d(s, s')$ is the distance between a pair of states s and s' - 1.0 for side-by-side states and $\sqrt{2}$ for diagonally separated states; using this negative reward is the mechanism by which the agents take sideways actions (north, west, etc.) to be shorter than diagonal actions (northwest, etc.). This negative reward was not present when the shelter was not in the environment, i.e. the training phase of condition 2, to avoid accumulating unmitigated negative value in each state-action pair.

In general, the reinforcement learning problem is to find a policy, π, which maps states to actions, such that the expected sum of discounted future rewards is maximized (Sutton and Barto, 2018).

$$E[\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})|s_0 = s]$$

where $a_t = \pi(s_t)$, i.e. actions given by the policy and γ is the temporal discount factor, a hyperparameter specifying how much long-term reward should be weighted against

short-term reward. Each of the RL agents described below operates by searching for a policy that can optimize expected future reward. The algorithms have different limitations and compute their policies differently; thus, different algorithms often generate different policies. We compared the behavior of these various algorithms to mouse behavior, in order to end up with a concrete, computational description of mouse behavior.

Simulation details

Simulation experiments consisted of three phases: a training phase, a pre-test phase, and a test phase. Each algorithm was repeated 100 times with 100 different random seeds. Each agent started by being dropped in at a (uniform) random location in the arena. In the training phase, unless otherwise stated, the RL agent then moved around the environment with a random policy (probability of $\frac{1}{8}$ for each action) and learned based on this experience. Moving into a barrier (black) resulted in the agent remaining in the same state from which it initiated an action in the previous timestep. Trip wires acted like barriers but only when the agent was attempting to pass the trip wire in the threat-area-to-obstacle-edge direction. Each algorithm received enough training steps that all 100 seeds was able to learn an escape to shelter in condition 1, after being dropped into the threat zone, rounded up to the nearest 500 steps (for models that took <10k steps) or 5k steps (for models that took >10k steps) (Table 6.4). Thus, we are modelling only the mice that actually learn edge-vector escapes during the training phase. This number of training steps was used across all four conditions. In the pre-test phase, the agent started in the shelter and then moved randomly through the environment until reaching the threat zone square (learning was allowed to continue during this period). At this point, the test phase was initiated. The agent then stopped moving randomly and adopted its learned policy in order to navigate to the reward. After this a second and third trial (pre-test + test phase for each one) were performed. The test phase proceeded until the agent reached the shelter or for a maximum of 100 steps.

Q-learning

At test time, the Q-learning agent generates a policy by selecting the action a in the current state s that has the maximum state-action value. State-action values are incrementally learned during the training and pre-test phases using the Q-learning algorithm (Watkins, 1992) combined with an eligibility trace (Sutton and Barto, 2018). The eligibility trace is a decaying trace of recent state-action pairs. After

taking action a_t in state s_t and moving to state s_{t+1}, the agent takes three steps to update its state-action values. First, it decays its eligibility trace e, by $e \leftarrow \lambda\gamma e$, where λ is the eligibility trace decay parameter and γ is the temporal discount factor introduced above. Second, it updates its eligibility trace to add the current state-action pair: $e(s_t, a_t) \leftarrow e(s_t, a_t) + 1$. Finally, it updates its state-action-value table:

$$Q(s_t, a_t) \leftarrow Q(s, a) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]e$$

where r_t is the reward gained from this step, α is the learning rate and γ is the temporal discount factor. State-action values are initialized randomly with mean 0 and variance 0.1.

Tile coding

One limitation of tabular methods is that they are unable to generalize. Learning information (e.g. about value) in one state does not provide information about any other states. A common way to overcome this is to use function approximation to represent quantities rather than storing them explicitly in look-up tables. Among the simplest forms of function approximation is a linear map. For example, the approximate state-action value function can be defined as

$$\hat{Q}(s, a, \mathbf{w}) \equiv \mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{d} w_i x_i(s, a)$$

where $\mathbf{x}$ is the featured state with dimension d, and $\mathbf{w}$ are learnable weights. The update rule for these weights under stochastic gradient descent is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left[r_{t+1} + \gamma \max_{a'} Q(s', a') - Q(s, a) \right] Q(s, a)$$

where α is the learning rate and γ is the discount factor. One popular way to featurize a state space for linear methods is tile coding. The feature map consists of a set of overlapping receptive fields; for each field a state is said to be present—and given a feature value of 1—if it is within the receptive field, and absent—and given a feature value of 0—if it is not. We use rectangular receptive fields (tiles) of both 2x2 and 3x3, shifted by 1 in both x and y coordinates as well as iterated over the available actions. For a more detailed treatment of linear function approximation and coarse coding methods, see chapter 9 of Sutton and Barto, 2018.

Hierarchical state space

The hierarchical state space experiments took place in exactly the same gridworld environment and conditions as with the non-hierarchical (flat) learners. The difference

was that the Q-learning policy that the agent learned was in relation to a different state space. Instead of the 118 grid states an 944 state-action pairs, this regional state space contained 10 states (regional groupings of grid states, e.g. the obstacle edge areas) and 40 state-action pairs (e.g. go to the shelter area from the left obstacle edge area). During the training phase, the agent's policy was updated with respect to its transitions between these regions. For example, it would only update the value of its "go to the shelter area from the left obstacle edge area" immediately after crossing the border between those regions. Here, the distance function $d(s, s')$ that determines negative reward per timestep was equal to the distance between the centroids of the regions that the agent moved between. When the agent executes its policy at test time, it produces high-level actions. To carry out these actions, its low-level controller simply carries out an innate ability to move directly in a straight line from its current position (e.g. the threat zone) to the target location (e.g. obstacle edge area), similar to Edvardsen et al., 2020. We set up this hierarchical state space to use with Q-learning out of convenience, but it could have been used with the other gradual learners as well (Supplementary Note).

Successor Representation

The SR agent uses a model-free update rule to learn a representation of how state-action pairs predict (temporally discounted) future occupancy in each state in the environment. This successor representation, M, is thus a $SxAxS'$ tensor, where the index of the first two dimensions identify a state-action pair and the third dimension corresponds to the successor state. M can be combined with a separately learned reward vector R in order to compute value:

$$Q(S, A) = \sum_{s'} M(S, A, s')R(s')$$

This equation shows that the value of a state-action pair is the product of how much that state-action pair predicts future occupancy in the rewarded states and how much reward is those states. In our experiments, there is at most one rewarded state, so this reduces to:

$$Q(S, A) = M(S, A, shelter)R(shelter)$$

In order to learn the successor representation M, the agent applies a model-free updating rule with an eligibility trace (Gershman et al., 2012) to an entire row after each step:

$$M(s_t, a_t, :) \leftarrow M(s_t, a_t, :) + \alpha[1_{s_{t+1}} + \gamma E_a[M(s_{t+1}, a, :)] - M(s_t, a_t, :)]e$$

where α is the learning rate, $1_{s_{t+1}}$ is a one-hot vector with a 1 in the position of the successor state s', γ is the temporal discount factor, e is the eligibility trace updated similarly to Q-learning as described above, and $E_a[...]$ is the expected row in the SR for the successor state s', averaged across the possible actions taken from that state. SR values are initialized randomly with mean 0 and variance 1. Simultaneously, a reward vector must be learned. It is updated after each step:

$$R(s_t) \leftarrow R(s_t) + \alpha(r_t - R(s_t))$$

The reward vector is initialized to all zeros.

Model-based agent

The model-based agent builds up a model of the environment in the form of an undirected graph. Each time the agent encounters a new state, it stores that state as a node in the graph. Each time the agent receives a reward, it labels the node from which the reward emanated with the amount of reward. Each time the agent takes a new transition between nodes, it stores that transition as an edge in the graph. Each time the agent attempts to make a transition and is blocked by an obstacle or trip wire, it deletes that edge from the graph. The immediate learner plans using the most recent set of edges. The gradual learner stores a buffer of up to N observations per edge. During planning, edges are only used if the majority of observations in the buffer indicate that the edge is not blocked. In addition, the reward in each state is taken to be the average reward observed over the past N observations. At decision time, the model-based agent uses its model to plan the shortest possible route to the reward location, where horizontal and vertical edges have a path length of 1.0 and diagonal edges have a path length of $\sqrt{2}$. This is a heuristic that maximizes the expected future reward in this navigation-task setting. Shortest routes were calculated using an A-star tree search algorithm (Hart et al., 1968). Equally effective actions (according to the A-star algorithm, which finds the shortest route to the goal) were sampled with equal probability.

Practice runs

We augmented the random exploration policy during the training phase with practice edge-vector and shelter-vector runs. Edge-vector runs were hard-coded action trajectories taking the agent from the threat area directly to an obstacle edge. The initiation and termination states are shown in Fig. 4.2. Each time the agent entered one of these states, the hard-coded trajectory was triggered with a probability of 0.2.

Classifying escape runs

We used four classifications for simulated escape runs: homing-vector routes, edge-vector routes, tortuous routes and non-escapes. Homing-vector routes went from the threat zone to one of the three middle states above the obstacle location, and then continued toward the shelter (south, southwest or southeast) from there. Edge-vector routes went from the threat zone to the obstacle edge, without deviating from its path by more than one step to go around the trip wire. Tortuous routes are homing-vector or edge-vector routes that deviate from that path (to go around a trip wire location) by at least two steps. Non-escapes did not reach the shelter within the 50-step time limit.

Algorithm	Hyperparameter	Value
Q-learning	temporal discount factor γ	0.9
Q-learning	TD(λ) decay factor γ	0.5
Q-learning	learning rate, α	0.1
Q-learning	neg. reward per step	0.01
SR	temporal discount factor γ	0.9
SR	TD(λ) decay factor γ	0.5
SR	learning rate, α	0.1
SARSA	temporal discount factor γ	0.99
SARSA	TD(λ) decay factor γ	0.5
SARSA	learning rate, α	0.1
SARSA	neg. reward per step	0.001
Tile coding	tile size	[2x2, 3x3]
MB-G	model buffer window, N	15

Table 6.3: Hyper-parameters used in the RL models. While we did not conduct extensive comparison over hyper-parameters, we endeavored to use comparable settings across models and chose from typical ranges for grid-world environments in the RL literature (e.g. https://github.com/karpathy/reinforcejs).

Algorithm	Exploration	# steps to learn
Tabular Q-learning	Random	45k
Tabular Q-learning	Random + practice runs	30k
Hierarchical Q-learning	Random	2.5k
Hierarchical Q-learning	Random + practice runs	1.5k
Tile-coding Q-learning	Random	285k
Successor Representation	Random	125k
Successor Representation	Random + practice runs	20k
Model-based (Immediate)	Random + practice runs	3k
Model-based (Gradual)	Random + practice runs	3k
Tabular SARSA	Random + practice runs	35k
Hierarchical SARSA	Random + practice runs	2k

Table 6.4: Training steps needed for the RL models to learn escape routes

Bibliography

Able, Kenneth P. “Mechanisms of orientation, navigation and homing”. In: *Animal migration, orientation and navigation* (1980), pp. 283–373 (cit. on p. 26).

Allen Institute for Brain Science. *Allen Mouse Brain Atlas*. 2015. URL: https://mouse.brain-map.org/static/atlas (cit. on p. 99).

Alyan, Sofyan and Rudolf Jander. “Short-range homing in the house mouse, Mus musculus: stages in the learning of directions”. In: *Animal Behaviour* 48.2 (1994), pp. 285–298 (cit. on pp. 32, 53).

Aoun, Peter, Timothy Jones, Gordon L Shaw, and Mark Bodner. “Long-term enhancement of maze learning in mice via a generalized Mozart effect”. In: *Neurological research* 27.8 (2005), pp. 791–796 (cit. on p. 27).

Ballard, D H, M M Hayhoe, P K Pook, and R P Rao. “Deictic codes for the embodiment of cognition”. en. In: *Behav. Brain Sci.* 20.4 (Dec. 1997), 723–42, discussion 743–67 (cit. on p. 54).

Banino, Andrea, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. “Vector-based navigation using grid-like representations in artificial agents”. en. In: *Nature* 557.7705 (May 2018), pp. 429–433 (cit. on p. 82).

Barnes, Carol A. “Memory deficits associated with senescence: a neurophysiological and behavioral study in the rat.” In: *Journal of comparative and physiological psychology* 93.1 (1979), p. 74 (cit. on p. 32).

Baron, S P and L T Meltzer. “Mouse strains differ under a simple schedule of operant learning”. en. In: *Behav. Brain Res.* 118.2 (Jan. 2001), pp. 143–152 (cit. on pp. 24, 69).

Barto, Andrew G, Satinder Singh, Nuttapong Chentanez, et al. “Intrinsically motivated learning of hierarchical collections of skills”. In: *Proceedings of the 3rd International Conference on Development and Learning*. Piscataway, NJ. 2004, pp. 112–19 (cit. on pp. 84, 90).

Behrens, Timothy EJ, Timothy H Muller, James CR Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-Nelson. "What is a cognitive map? Organizing knowledge for flexible behavior". In: *Neuron* 100.2 (2018), pp. 490–509 (cit. on p. 27).

Benhamou, Simon. "An analysis of movements of the wood mouse Apodemus sylvaticus in its home range". In: *Behavioural Processes* 22.3 (1991), pp. 235–250 (cit. on pp. 27, 31).

Biro, Dora, Jessica Meade, and Tim Guilford. "Familiar route loyalty implies visual pilotage in the homing pigeon". In: *Proceedings of the National Academy of Sciences* 101.50 (2004), pp. 17440–17443 (cit. on p. 26).

Bittner, Katie C, Aaron D Milstein, Christine Grienberger, Sandro Romani, and Jeffrey C Magee. "Behavioral time scale synaptic plasticity underlies CA1 place fields". In: *Science* 357.6355 (2017), pp. 1033–1036 (cit. on p. 24).

Blanchard, D Caroline and Robert J Blanchard. "Ethoexperimental approaches to the biology of emotion". In: *Annual review of psychology* 39.1 (1988), pp. 43–68 (cit. on p. 31).

Bovet, Pierre and Simon Benhamou. "Spatial analysis of animals' movements using a correlated random walk model". In: *Journal of theoretical biology* 131.4 (1988), pp. 419–433 (cit. on p. 25).

Burgess, Neil, Michael Recce, and John O'Keefe. "A model of hippocampal function". In: *Neural Netw.* 7.6 (Jan. 1994), pp. 1065–1081 (cit. on pp. 29, 53).

Burnham, Kenneth P and David R Anderson. "Multimodel inference". In: *Sociol. Methods Res.* 33.2 (Nov. 2004), pp. 261–304 (cit. on p. 68).

Chapuis, N, Catherine Thinus-Blanc, and Bruno Poucet. "Dissociation of mechanisms involved in dogs' oriented displacements". In: *The Quarterly Journal of Experimental Psychology* 35.3 (1983), pp. 213–219 (cit. on p. 28).

Chapuis, Nicole. "Detour and shortcut abilities in several species of mammals". In: *Cognitive processes and spatial orientation in animal and man.* Springer, 1987, pp. 97–106 (cit. on p. 23).

Chase, William G and Herbert A Simon. "Perception in chess". In: *Cogn. Psychol.* 4.1 (Jan. 1973), pp. 55–81 (cit. on p. 54).

Cheng, Ken, Janellen Huttenlocher, and Nora S Newcombe. "25 years of research on the use of geometry in spatial reorientation: a current theoretical perspective". en. In: *Psychon. Bull. Rev.* 20.6 (Dec. 2013), pp. 1033–1054 (cit. on p. 30).

Cheng, Ken, Ajay Narendra, Stefan Sommer, and Rüdiger Wehner. "Traveling in clutter: navigation in the Central Australian desert ant Melophorus bagoti". In: *Behavioural Processes* 80.3 (2009), pp. 261–268 (cit. on p. 26).

Chernetsov, Nikita, Dmitry Kishkinev, and Henrik Mouritsen. “A long-distance avian migrant compensates for longitudinal displacement during spring migration”. In: *Current Biology* 18.3 (2008), pp. 188–190 (cit. on p. 26).

Clark, A. “An embodied cognitive science?” en. In: *Trends Cogn. Sci.* 3.9 (Sept. 1999), pp. 345–351 (cit. on p. 54).

Collett, Thomas S, Matthew Collett, and Rüdiger Wehner. “The guidance of desert ants by extended landmarks”. In: *Journal of Experimental Biology* 204.9 (2001), pp. 1635–1639 (cit. on p. 28).

Collett, TS. “Do toads plan routes? A study of the detour behaviour of Bufo viridis”. In: *Journal of Comparative Physiology* 146.2 (1982), pp. 261–271 (cit. on pp. 28, 53).

— “Making learning easy: the acquisition of visual information during the orientation flights of social wasps”. In: *Journal of Comparative Physiology A* 177.6 (1995), pp. 737–747 (cit. on p. 54).

Cooper, William E and Daniel T Blumstein. *Escaping from predators: an integrative view of escape decisions*. Cambridge University Press, 2015 (cit. on p. 31).

Crowcroft, Peter. *Mice all over*. Foulis, 1966 (cit. on pp. 31, 68).

Dashiell, John Frederick. “Some Transfer Factors in Maze Learning by the White Rat.” In: *Psychobiology* 2.4 (1920), p. 329 (cit. on p. 28).

Datta, Sandeep Robert, David J Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. “Computational neuroethology: a call to action”. In: *Neuron* 104.1 (2019), pp. 11–24 (cit. on p. 91).

Daw, Nathaniel D, Yael Niv, and Peter Dayan. “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nature neuroscience* 8.12 (2005), pp. 1704–1711 (cit. on pp. 24, 78, 83).

Dayan, Peter. “Improving Generalization for Temporal Difference Learning: The Successor Representation”. In: *Neural Comput.* 5.4 (July 1993), pp. 613–624 (cit. on p. 76).

De Cothi, William, Nils Nyberg, Eva-Maria Griesbauer, Carole Ghanamé, Fiona Zisch, Julie M Lefort, Lydia Fletcher, Coco Newton, Sophie Renaudineau, Daniel Bendor, et al. “Predictive maps in rats and humans for spatial navigation”. In: *Current Biology* (2022) (cit. on pp. 74–76, 90).

Dill, LM and RC Ydenberg. “The group size–flight distance relationship in water striders (Gerris remigis)”. In: *Canadian Journal of Zoology* 65.2 (1987), pp. 223–226 (cit. on p. 31).

Doeller, Christian F, John A King, and Neil Burgess. “Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.15 (Apr. 2008), pp. 5915–5920 (cit. on p. 69).

Drickamer, Lee. C. and James Stuart. “Peromyscus: snow tracking and possible cues used for navigation”. In: *American Midland Naturalist* (1984), pp. 202–204 (cit. on p. 31).

Edvardsen, Vegard, Andrej Bicanski, and Neil Burgess. “Navigating with grid and place cells in cluttered environments”. en. In: *Hippocampus* 30.3 (Mar. 2020), pp. 220–232 (cit. on pp. 30, 53, 68, 108).

Eichenbaum, Howard, Caroline Stewart, and RG Morris. “Hippocampal representation in place learning”. In: *Journal of Neuroscience* 10.11 (1990), pp. 3531–3542 (cit. on p. 22).

Ellard, Colin G and Meghan C Eller. “Spatial cognition in the gerbil: computing optimal escape routes from visual threats”. In: *Animal cognition* 12.2 (2009), pp. 333–345 (cit. on pp. 28, 32, 53).

Etienne, Ariane S and Kathryn J Jeffery. “Path integration in mammals”. In: *Hippocampus* 14.2 (2004), pp. 180–192 (cit. on p. 25).

Etienne, Ariane S, Evelyne Teroni, Roland Maurer, Véronique Portenier, and Francis Saucy. “Short-distance homing in a small mammal: the role of exteroceptive cues and path integration”. In: *Experientia* 41.1 (1985), pp. 122–125 (cit. on p. 53).

Evans, Dominic A, A Vanessa Stempel, Ruben Vale, and Tiago Branco. “Cognitive Control of Escape Behaviour”. en. In: *Trends Cogn. Sci.* 23.4 (Apr. 2019), pp. 334–348 (cit. on pp. 31, 32).

Geerts, Jesse P, Fabian Chersi, Kimberly L Stachenfeld, and Neil Burgess. “A general model of hippocampal and dorsal striatal learning and decision making”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.49 (Dec. 2020), pp. 31427–31437 (cit. on pp. 24, 69, 78, 85).

Gershman, Samuel J, Christopher D Moore, Michael T Todd, Kenneth A Norman, and Per B Sederberg. “The successor representation and temporal context”. In: *Neural Computation* 24.6 (2012), pp. 1553–1568 (cit. on p. 108).

Gollub, L. “Conditioned reinforcement: Schedule effects”. In: *Handbook of operant behavior* (1977), pp. 288–312 (cit. on p. 70).

Gradinaru, Viviana, Kimberly R Thompson, Feng Zhang, Murtaza Mogri, Kenneth Kay, M Bret Schneider, and Karl Deisseroth. “Targeting and readout strategies for fast optical neural control in vitro and in vivo”. en. In: *J. Neurosci.* 27.52 (Dec. 2007), pp. 14231–14238 (cit. on p. 57).

Grieves, Roderick M and Paul A Dudchenko. “Cognitive maps and spatial inference in animals: Rats fail to take a novel shortcut, but can take a previously experienced one”. In: *Learn. Motiv.* 44.2 (May 2013), pp. 81–92 (cit. on p. 90).

Hamilton, Derek A, Cory S Rosenfelt, and Ian Q Whishaw. "Sequential control of navigation by locale and taxon cues in the Morris water task". In: *Behavioural brain research* 154.2 (2004), pp. 385–397 (cit. on p. 20).

Harrison, Fiona E, Randall S Reiserer, Andrew J Tomarken, and Michael P McDonald. "Spatial and nonspatial escape strategies in the Barnes maze". In: *Learning & memory* 13.6 (2006), pp. 809–819 (cit. on p. 32).

Hart, Peter E, Nils J Nilsson, and Bertram Raphael. "A formal basis for the heuristic determination of minimum cost paths". In: *IEEE transactions on Systems Science and Cybernetics* 4.2 (1968), pp. 100–107 (cit. on p. 109).

Hazen, Nancy L, Jeffrey J Lockman, and Herbert L Pick. "The development of children's representations of large-scale environments". In: *Child Dev.* 49.3 (Sept. 1978), pp. 623–636 (cit. on p. 91).

Hobhouse, LT. *Mind in Evolution/Eds Wozniak rH Series: Thoemmes Press, Classics in Psychology*. 1901 (cit. on p. 28).

Horton, Travis W, Richard N Holdaway, Alexandre N Zerbini, Nan Hauser, Claire Garrigue, Artur Andriolo, and Phillip J Clapham. "Straight as an arrow: humpback whales swim constant course tracks during long-distance migration". In: *Biology letters* 7.5 (2011), pp. 674–679 (cit. on p. 26).

Hsiao, Hsiao Hung. *An experimental study of the rat's "insight" within a spatial complex*. Vol. 4. 4. Univesity of California Press, 1929 (cit. on p. 24).

Huber, Roman and Markus Knaden. "Egocentric and geocentric navigation during extremely long foraging paths of desert ants". In: *Journal of Comparative Physiology A* 201.6 (2015), pp. 609–616 (cit. on p. 26).

Hull, C L. "The Concept of the Habit-Family Hierarchy, and Maze Learning. Part I". In: *Psychol. Rev.* 41.1 (1934), pp. 33–54 (cit. on pp. 19, 70).

Ingle, David J. "Visually elicited evasive behavior in frogs". In: *Bioscience* 40.4 (1990), pp. 284–291 (cit. on p. 28).

Janus, Christopher. "Search strategies used by APP transgenic mice during navigation in the Morris water maze". In: *Learning & memory* 11.3 (2004), pp. 337–346 (cit. on p. 25).

Kabadayi, Can, Katarzyna Bobrowicz, and Mathias Osvath. "The detour paradigm in animal cognition". In: *Animal Cognition* 21.1 (2018), pp. 21–35 (cit. on p. 28).

Kim, Christina K, Avishek Adhikari, and Karl Deisseroth. "Integration of optogenetics with complementary methodologies in systems neuroscience". In: *Nature Reviews Neuroscience* 18.4 (2017), pp. 222–235 (cit. on p. 88).

Kohler, Martin and Rüdiger Wehner. "Idiosyncratic route-based memories in desert ants, Melophorus bagoti: how do they interact with path-integration vectors?" In: *Neurobiology of learning and memory* 83.1 (2005), pp. 1–12 (cit. on p. 26).

Kohler, W. *The Mentality of Apes.* 1925 (cit. on p. 28).

Krakauer, John W, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. "Neuroscience needs behavior: correcting a reductionist bias". In: *Neuron* 93.3 (2017), pp. 480–490 (cit. on p. 91).

Krejcova, Gabriela, Jiri Patocka, and Jirina Slaninova. "Effect of humanin analogues on experimentally induced impairment of spatial memory in rats". In: *Journal of Peptide Science: An Official Publication of the European Peptide Society* 10.10 (2004), pp. 636–639 (cit. on p. 27).

Lagos, Patricio A, Andrea Meier, Liliana Ortiz Tolhuysen, Rodrigo A Castro, Francisco Bozinovic, and Luis A Ebensperger. "Flight initiation distance is differentially sensitive to the costs of staying and leaving food patches in a small-mammal prey". In: *Canadian Journal of Zoology* 87.11 (2009), pp. 1016–1023 (cit. on p. 31).

Layne, John E, W Jon P Barnes, and Lindsey MJ Duncan. "Mechanisms of homing in the fiddler crab Uca rapax 1. Spatial and temporal characteristics of a system of small-scale navigation". In: *Journal of Experimental Biology* 206.24 (2003), pp. 4413–4423 (cit. on pp. 28, 53).

Lima, Steven L and Lawrence M Dill. "Behavioral decisions made under the risk of predation: a review and prospectus". In: *Canadian journal of zoology* 68.4 (1990), pp. 619–640 (cit. on p. 31).

Liu, Annie, Andrew E Papale, James Hengenius, Khusbu Patel, Bard Ermentrout, and Nathan N Urban. "Mouse navigation strategies for odor source localization". In: *Frontiers in neuroscience* 14 (2020), p. 218 (cit. on p. 53).

Lockman, Jeffrey J and Christina D Adams. "Going around transparent and grid-like barriers: detour ability as a perception–action skill". In: *Developmental Science* 4.4 (2001), pp. 463–471 (cit. on p. 28).

Lopes, Gonçalo, Niccolò Bonacchi, João Frazão, Joana P Neto, Bassam V Atallah, Sofia Soares, Luis Moreira, Sara Matias, Pavel M Itskov, Patricia A Correia, Roberto E Medina, Lorenza Calcaterra, Elena Dreosti, Joseph J Paton, and Adam R Kampff. "Bonsai: an event-based framework for processing and controlling data streams". en. In: *Front. Neuroinform.* 9 (Apr. 2015), p. 7 (cit. on p. 96).

Lorenz, Konrad. *King Solomon's ring.* Routledge, 1949 (cit. on p. 26).

Maaswinkel, H and I Q Whishaw. "Homing with locale, taxon, and dead reckoning strategies by foraging rats: sensory hierarchy in spatial navigation". en. In: *Behav. Brain Res.* 99.2 (Mar. 1999), pp. 143–152 (cit. on pp. 32, 83).

Magno, Luiz Alexandre Viana, Helia Tenza-Ferrer, Mélcar Collodetti, Matheus Felipe Guimarães Aguiar, Ana Paula Carneiro Rodrigues, Rodrigo Souza da Silva, Joice do Prado Silva, Nycolle Ferreira Nicolau, Daniela Valadão Freitas Rosa,

Alexander Birbrair, Débora Marques Miranda, and Marco Aurélio Romano-Silva. "Optogenetic Stimulation of the M2 Cortex Reverts Motor Dysfunction in a Mouse Model of Parkinson's Disease". en. In: *J. Neurosci.* 39.17 (Apr. 2019), pp. 3234–3248 (cit. on p. 57).

Mataric, M J. "Integration of representation into goal-driven behavior-based robots". In: *IEEE Trans. Rob. Autom.* 8.3 (June 1992), pp. 304–312 (cit. on p. 54).

Mateo, Jill M. "The development of alarm-call response behaviour in free-living juvenile Belding's ground squirrels". In: *Animal Behaviour* 52.3 (1996), pp. 489–505 (cit. on p. 31).

Mathis, Alexander, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature neuroscience* 21.9 (2018), pp. 1281–1289 (cit. on p. 101).

McCreery, Helen F, Zachary A Dix, Michael D Breed, and Radhika Nagpal. "Collective strategy for obstacle navigation during cooperative transport by ants". In: *Journal of Experimental Biology* 219.21 (2016), pp. 3366–3375 (cit. on p. 53).

McFadden, Daniel. *Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments*. en. Ed. by David A Hensher and Peter R Stopher. Institute of Transportation Studies, University of California, 1977 (cit. on pp. 67, 104).

McMillan, Brock R and Donald W Kaufman. "Travel path characteristics for free-living white-footed mice (Peromyscus leucopus)". In: *Canadian Journal of Zoology* 73.8 (1995), pp. 1474–1478 (cit. on p. 31).

McNamee, Daniel C, Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. "Flexible modulation of sequence generation in the entorhinal-hippocampal system". en. In: *Nat. Neurosci.* 24.6 (June 2021), pp. 851–862 (cit. on p. 83).

Mobbs, Dean, Drew B Headley, Weilun Ding, and Peter Dayan. "Space, time, and fear: survival computations along defensive circuits". In: *Trends in cognitive sciences* 24.3 (2020), pp. 228–241 (cit. on pp. 55, 90).

Mobbs, Dean, Pete C Trimmer, Daniel T Blumstein, and Peter Dayan. "Foraging for foundations in decision neuroscience: insights from ethology". In: *Nature Reviews Neuroscience* 19.7 (2018), pp. 419–427 (cit. on p. 91).

Morris, Richard G M. "Spatial localization does not require the presence of local cues". In: *Learn. Motiv.* 12.2 (May 1981), pp. 239–260 (cit. on pp. 20, 30).

Müller, Martin and Rüdiger Wehner. "Path integration in desert ants, Cataglyphis fortis". In: *Proceedings of the National Academy of Sciences* 85.14 (1988), pp. 5287–5290 (cit. on p. 25).

O'Keefe, John and Jonathan Dostrovsky. "The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat." In: *Brain research* (1971) (cit. on p. 21).

O'Keefe, John and Lynn Nadel. *The hippocampus as a cognitive map*. en. Clarendon Press, 1978 (cit. on pp. 19, 23, 24, 30).

Packard, M G, R Hirsh, and N M White. "Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems". en. In: *J. Neurosci.* 9.5 (May 1989), pp. 1465–1472 (cit. on pp. 21, 69).

Packard, Mark G and James L McGaugh. "Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: further evidence for multiple memory systems." In: *Behavioral neuroscience* 106.3 (1992), p. 439 (cit. on pp. 21, 22).

— "Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning". In: *Neurobiology of learning and memory* 65.1 (1996), pp. 65–72 (cit. on p. 21).

Patil, Sudarshan S, Berta Sunyer, Harald Höger, and Gert Lubec. "Evaluation of spatial memory of C57BL/6J and CD1 mice in the Barnes maze, the Multiple T-maze and in the Morris water maze". In: *Behavioural brain research* 198.1 (2009), pp. 58–68 (cit. on p. 25).

Paxinos, George and Keith B J Franklin. *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates*. en. Academic Press, Apr. 2019 (cit. on p. 60).

Petitto, L A and P F Marentette. "Babbling in the manual mode: evidence for the ontogeny of language". en. In: *Science* 251.5000 (Mar. 1991), pp. 1493–1496 (cit. on pp. 54, 91).

Piaget, Jean. *The child's construction of reality*. Routledge & Kegan Paul, 1955 (cit. on p. 91).

Poucet, Bruno, Catherine Thinus-Blanc, and Nicole Chapuis. "Route planning in cats, in relation to the visibility of the goal". In: *Animal Behaviour* 31.2 (1983), pp. 594–599 (cit. on p. 28).

Rauscher, Frances, Desix Robinson, and Jason Jens. "Improved maze learning through early music exposure in rats". In: *Neurological research* 20.5 (1998), pp. 427–432 (cit. on p. 27).

Redish, A David et al. *Beyond the cognitive map: from place cells to episodic memory*. MIT press, 1999 (cit. on p. 21).

Regolin, Lucia, Giorgio Vallortigara, and Mario Zanforlin. "Object and spatial representations in detour problems by chicks". In: *Animal Behaviour* 49.1 (1995), pp. 195–199 (cit. on p. 28).

Reimers, Eigil and Sindre Eftestøl. “Response behaviors of Svalbard reindeer towards humans and humans disguised as polar bears on Edgeøya”. In: *Arctic, antarctic, and alpine research* 44.4 (2012), pp. 483–489 (cit. on p. 31).

Reppert, Steven M and Jacobus C de Roode. “Demystifying monarch butterfly migration”. In: *Current Biology* 28.17 (2018), R1009–R1022 (cit. on p. 26).

Rescorla, Robert A. “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement”. In: *Current research and theory* (1972), pp. 64–99 (cit. on p. 20).

Restle, F. “Discrimination of cues in mazes: a resolution of the place-vs.-response question”. en. In: *Psychol. Rev.* 64.4 (July 1957), pp. 217–228 (cit. on pp. 20, 30, 41, 69).

Ribas-Fernandes, José J F, Alec Solway, Carlos Diuk, Joseph T McGuire, Andrew G Barto, Yael Niv, and Matthew M Botvinick. “A neural signature of hierarchical reinforcement learning”. en. In: *Neuron* 71.2 (July 2011), pp. 370–379 (cit. on pp. 75, 90).

Russek, Evan M, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw. “Predictive representations can link model-based reinforcement learning to model-free mechanisms”. en. In: *PLoS Comput. Biol.* 13.9 (Sept. 2017), e1005768 (cit. on pp. 75, 76, 84).

Schlesinger, K, D U Lipsitz, P L Peck, M A Pelleymounter, J M Stewart, and T N Chase. “Substance P enhancement of passive and active avoidance conditioning in mice”. en. In: *Pharmacol. Biochem. Behav.* 19.4 (Oct. 1983), pp. 655–661 (cit. on p. 101).

Schmitzer-Torbert, Neil and A David Redish. “Development of path stereotypy in a single day in rats on a multiple-T maze.” In: *Archives italiennes de biologie* 140.4 (2002), pp. 295–301 (cit. on pp. 27, 28).

Schölkopf, Bernhard and Hanspeter A Mallot. “View-based cognitive mapping and path planning”. In: *Adaptive Behavior* 3.3 (1995), pp. 311–348 (cit. on p. 54).

Schulz, Eric and Samuel J Gershman. “The algorithmic architecture of exploration in the human brain”. en. In: *Curr. Opin. Neurobiol.* 55 (Apr. 2019), pp. 7–14 (cit. on p. 68).

Schulz, Eric, Edgar D Klenske, Neil R Bramley, and Maarten Speekenbrink. “Strategic exploration in human adaptive control”. en. May 2017 (cit. on p. 68).

Shamash, Philip, Matteo Carandini, Kenneth Harris, and Nick Steinmetz. “A tool for analyzing electrode tracks from slice histology”. In: *bioRxiv* (2018) (cit. on p. 99).

Shamash, Philip, Sarah F. Olesen, Panagiota Iordanidou, Dario Campagner, Nabhojit Banerjee, and Tiago Branco. “Mice learn multi-step routes by memorizing subgoal locations”. In: *Nature Neuroscience* (July 2021) (cit. on pp. 100, 105).

Sharma, Sunita, Sharlene Rakoczy, and Holly Brown-Borg. “Assessment of spatial memory in mice”. In: *Life sciences* 87.17-18 (2010), pp. 521–536 (cit. on pp. 27, 87).

Siegel, Shepard and Lorraine G Allan. “The widespread influence of the Rescorla-Wagner model”. In: *Psychonomic Bulletin & Review* 3.3 (1996), pp. 314–321 (cit. on p. 20).

Siegle, Joshua H and Matthew A Wilson. “Enhancement of encoding and retrieval functions through theta phase-specific manipulation of hippocampus”. In: *elife* 3 (2014) (cit. on p. 88).

Solway, Alec, Carlos Diuk, Natalia Córdova, Debbie Yee, Andrew G Barto, Yael Niv, and Matthew M Botvinick. “Optimal behavioral hierarchy”. In: *PLoS computational biology* 10.8 (2014), e1003779 (cit. on pp. 30, 85, 90).

Spiers, Hugo J and Sam J Gilbert. “Solving the detour problem in navigation: a model of prefrontal and hippocampal interactions”. en. In: *Front. Hum. Neurosci.* 9 (Mar. 2015), p. 125 (cit. on pp. 29, 53, 68, 74, 85).

Stachenfeld, Kimberly L, Matthew M Botvinick, and Samuel J Gershman. “The hippocampus as a predictive map”. In: *Nat. Neurosci.* 20.11 (Nov. 2017), pp. 1643–1653 (cit. on pp. 53, 74, 75, 85).

Stamatakis, Alice M and Garret D Stuber. “Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance”. en. In: *Nat. Neurosci.* 15.8 (June 2012), pp. 1105–1107 (cit. on p. 101).

Stopka, Pavel and David W Macdonald. “Way-marking behaviour: an aid to spatial navigation in the wood mouse (Apodemus sylvaticus)”. In: *BMC ecology* 3.1 (2003), pp. 1–9 (cit. on p. 27).

Sutton, Richard S. “Dyna, an integrated architecture for learning, planning, and reacting”. In: *ACM Sigart Bulletin* 2.4 (1991), pp. 160–163 (cit. on p. 84).

— “Generalization in reinforcement learning: Successful examples using sparse coarse coding”. In: *Advances in neural information processing systems* 8 (1995) (cit. on p. 80).

Sutton, Richard S and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on pp. 73, 74, 79, 105–107).

Sutton, Richard S, Doina Precup, and Satinder Singh. “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning”. In: *Artif. Intell.* 112.1 (Aug. 1999), pp. 181–211 (cit. on pp. 75, 80, 84).

Teichroeb, Julie Annette and Eve Ann Smeltzer. “Vervet monkey (Chlorocebus pygerythrus) behavior in a multi-destination route: Evidence for planning ahead when heuristics fail”. In: *Plos one* 13.5 (2018), e0198076 (cit. on p. 30).

Thompson, Steven D. “Spatial utilization and foraging behavior of the desert woodrat, Neotoma lepida lepida”. In: *Journal of Mammalogy* 63.4 (1982), pp. 570–581 (cit. on pp. 27, 31).

Thorndike, EL. *Animal intelligence; experimental studies. On cover: the animal behavior series*. 1911 (cit. on p. 28).

Tolman, E C. “Cognitive maps in rats and men”. en. In: *Psychol. Rev.* 55.4 (July 1948), pp. 189–208 (cit. on p. 19).

Tolman, E C and C H Honzik. “Introduction and removal of reward, and maze performance in rats”. In: *Publ. Psychol.* 4 (1930), pp. 257–275 (cit. on pp. 22, 23, 27, 30, 87).

Tomov, Momchil S, Samyukta Yagati, Agni Kumar, Wanqian Yang, and Samuel J Gershman. “Discovery of hierarchical representations for efficient planning”. In: *PLoS computational biology* 16.4 (2020), e1007594 (cit. on pp. 75, 84, 85, 90).

Trullier, Olivier, Sidney I Wiener, Alain Berthoz, and Jean-Arcady Meyer. “Biologically based artificial navigation systems: Review and prospects”. In: *Progress in neurobiology* 51.5 (1997), pp. 483–544 (cit. on pp. 21, 22).

Tsoar, Asaf, Ran Nathan, Yoav Bartan, Alexei Vyssotski, Giacomo Dell’Omo, and Nachum Ulanovsky. “Large-scale navigational map in a mammal”. In: *Proceedings of the National Academy of Sciences* 108.37 (2011), E718–E724 (cit. on p. 26).

Tulving, Endel and Stephen A Madigan. “Memory and verbal learning”. In: *Annual review of psychology* 21.1 (1970), pp. 437–484 (cit. on p. 19).

Vale, Ruben, Dominic Evans, and Tiago Branco. “A Behavioral Assay for Investigating the Role of Spatial Memory During Instinctive Defense in Mice”. In: *JoVE (Journal of Visualized Experiments)* 137 (2018), e56988 (cit. on p. 32).

Vale, Ruben, Dominic A Evans, and Tiago Branco. “Rapid Spatial Learning Controls Instinctive Defensive Behavior in Mice”. In: *Curr. Biol.* 27.9 (May 2017), pp. 1342–1349 (cit. on pp. 31, 32, 35, 53, 59, 83).

Viswanathan, GM, V Afanasyev, Sergey V Buldyrev, Shlomo Havlin, MGE Da Luz, EP Raposo, and H Eugene Stanley. “Lévy flights in random searches”. In: *Physica A: Statistical Mechanics and its Applications* 282.1-2 (2000), pp. 1–12 (cit. on p. 25).

Wallace, Douglas G, Bogdan Gorny, and Ian Q Whishaw. “Rats can track odors, other rats, and themselves: implications for the study of spatial behavior”. In: *Behavioural brain research* 131.1-2 (2002), pp. 185–192 (cit. on p. 53).

Ward, Dave, David Silverman, and Mario Villalobos. “Introduction: The Varieties of Enactivism”. In: *Topoi* 36.3 (Sept. 2017), pp. 365–375 (cit. on pp. 54, 68, 88).

Watkins, Christopher J. “Daya. P: Technical Note: Q-Learning”. In: *Machine learning* 8.3 (1992), pp. 279–292 (cit. on p. 106).

Wehner, Rüdiger, Martin Boyer, Florian Loertscher, Stefan Sommer, and Ursula Menzi. “Ant navigation: one-way routes rather than maps”. In: *Current biology* 16.1 (2006), pp. 75–79 (cit. on p. 26).

Wells, MJ. “Short-term learning and interocular transfer in detour experiments with octopuses”. In: *Journal of Experimental Biology* 47.3 (1967), pp. 393–408 (cit. on p. 28).

Yilmaz, Melis and Markus Meister. “Rapid innate defensive responses of mice to looming visual stimuli”. en. In: *Curr. Biol.* 23.20 (Oct. 2013), pp. 2011–2015 (cit. on pp. 31, 32).

Zani, PA, TD Jones, RA Neuhaus, and JE Milgrom. “Effect of refuge distance on escape behavior of side-blotched lizards (Uta stansburiana)”. In: *Canadian Journal of Zoology* 87.5 (2009), pp. 407–414 (cit. on p. 31).