Biological Intelligence: From Behavior to Learning Theory

Thesis by
Tony Zhang

In Partial Fulfillment of the Requirements for the
Degree of
Computation and Neural Systems

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended November 23, 2021

Tony Zhang
ORCID: 0000-0002-5198-499X

ACKNOWLEDGEMENTS

Many made this thesis possible. First, I would like to thank my advisor, **Pietro**, for letting me join this surprisingly wonderful and sociable group of people that I get to call my academic family. You've inspired my confidence in posing bigger research questions and tackling more challenging problems, and have shown me the importance of detail and clarity in science. I also want to thank my secondary "advisor," **Markus**: you've demonstrated what it takes to do rigorous science. Thank you for always welcoming me to be a part of your group meetings, and for your frequent advice on our projects. Finally, I'm grateful for the many other faculty who have served on my various committees throughout the PhD: **Richard**, **Thanos**, **Katie**, and **Yisong**. Thank you for your support and advice.

To the Vision Lab Slackers I've had the privilege of befriending all these years: thank you for always being there and for all the good times. From our daily lab lunch and lab walks to our frequent outings, you've made my time here a blast: **Oisin**, **Mason**, **Grant**, **Alvita**, **Cristina**, **Matteo**, **Joe**, **Sara**, **Serim**, **Jennifer**, **Eli**, **Neehar**, **Laure**, and **Kevin**.

I also want to thank those I've befriended in other places. First, members of the Meister Lab, who often helped me reassert my identify as a neuroscientist-by-association in times of identity crisis: especially **Matt** and **Mu**, my closest collaborators, who have each contributed enormously to this thesis, because without them, the mouse data collected would have not been possible. Additionally, I am grateful to many passionate individuals I've befriended across Caltech who have made late-night conversations enjoyable. Special shout out to my close friends: **Sharan**, my tennis partner and two-time roommate, and **Aryeh**, whom I've befriended since I first arrived in Los Angeles. Thanks to all of my CNS friends for reinforcing the passion that first brought me here, and reminding me that the fundamental questions in our field are worth asking. Special shout out to the rest of the five-member gang in my cohort: **Jon**, **Matt**, **Jeremy**, and **Anish**. It warms my heart whenever I'm in your company, and I'll miss our lively dinner conversations.

On the personal front, I'm grateful to **my parents** for supporting all of my decisions in pursuing my studies. They've been by my side, each step along the way, despite being thousands of miles away. And of course, to **Anna**—thank you for supporting me during the pivotal parts of my PhD, and for your love.

ABSTRACT

Knowing how to learn, think, and act is not just a hallmark of intelligence, but a necessity of survival for many organisms. Behavior, the complete set of actions of species, allows us to glimpse into the minds of humans and animals, and by extension, intelligence itself. Biological intelligence is characterized by fast adaptation to changes and challenges, which is what allows species to survive in natural environments from starvation and predation. To study learning in a controlled setting, we can observe the behavior evoked through decision-making tasks that make it possible to quantify and analyze learning. By modeling the extracted behavioral features, we could start to understand the possible underlying mechanisms by proposing neural theory models, and look for those signals in the brain. Understanding the neural mechanisms of learning also strengthens the basis for building intelligent machines that are flexible and adaptive to the nonstationary world we live in. In this thesis, I present works in **(1)** automating behavioral setups and modeling suboptimal behavior in a traditional decision-making task [5], **(2)** using an ethological navigation task to characterize fast-sequence learning [6], and **(3)** how neural theory can explain some core behavioral phenomena in **(2)**, and be used to solve a central problem in graph search [8].

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Matthew Rosenberg*, Tony Zhang*, Pietro Perona, and Markus Meister. "Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration". In: *Elife* 10 (2021), e66175. URL: https://elifesciences.org/articles/66175.
Conceptualization, data collection setup, tracking, software, analysis.

[2] Tony Zhang, Matthew Rosenberg, Pietro Perona, and Markus Meister. "Endotaxis: A Universal Algorithm for Mapping, Goal-Learning, and Navigation". In: *bioRxiv* (2021). URL: https://www.biorxiv.org/content/10.1101/2021.09.24.461751v1.
Conceptualization, simulations, analysis.

[3] Matthew Rosenberg*, Tony Zhang*, Pietro Perona, and Markus Meister. "Rapid learning and efficient exploration by mice navigating a complex maze". In: *NeurIPS Biological and Artificial Reinforcement Learning Workshop* (2019). URL: https://sites.google.com/view/biologicalandartificialrl/home?authuser=0.
Conceptualization, data collection setup, tracking, software, analysis.

[4] Mu Qiao, Tony Zhang, Cristina Segalin, Sarah Sam, Pietro Perona, and Markus Meister. "Mouse Academy: high-throughput automated training and trial-by-trial behavioral analysis during learning". In: *bioRxiv* (2018), p. 467878. URL: https://www.biorxiv.org/content/10.1101/467878v2.
Tracking, behavioral trajectory analysis.

CONTENTS

LIST OF FIGURES

Number *Page*

Part I

Introduction

Chapter 1

BACKGROUND

Biological intelligence can be broadly defined as the ability of a biological organism to respond or adapt to its surroundings in an efficient way [23]. While learning complex concepts quickly can be a challenge for AI, biology seems to have converged on good ways to learn or adapt to changes in the environment both quickly and robustly. For humans, that might be learning to ride a bike, navigating in a new city, or adaptive procedural tasks like cleaning up a messy kitchen. The ability of biological organisms to learn quickly could be explained by evolutionary pressure for survival, which necessitates fast adaptation, whether it is one-shot association of an unseen organism as threat, or remembering where food is located and reliably returning to that location [46]. In constrained navigation, the learning can require the integration of multiple decisions into a sequence in order to achieve one goal, or even learning a map for reliably navigating to multiple locations [65].

From the perspective of neuroscience, behavior can be perceived as the low-dimensional projection of an organism's neural activities. An organism can act directly in response to events driven by salient sensory inputs, such as those of a predator or prey, or they could behave according to some more abstract goal, such as navigating to a particular location that requires previously learning a sequence of actions [46]. To study this in behavioral and systems neuroscience, researchers use a variety of experimental setups that characterize simple decision-making. These experiments often involve learning a few-bit decision-making task with 1 or 2 actions [43]. Fewer tasks are used that involve the learning of complex sequences of actions, which might be naturalistic but more difficult to analyze or control for head fixation, an important setup for certain neural recording tools [49]. As the behavior exhibited under each task can differ substantially, it is possible that the underlying neural mechanisms are thereby also different.

With increased advancements in hardware and software tools for automating the capture of all aspects of behavior, we can begin to tease apart the underpinnings of complex learning. Experimenters now have full control over the environment of the organism and are able to record all of its experiences when learning to perform novel tasks with increasing proficiency. This allows us to return to the behavior data

later and characterize the presence and absence of key behavioral phenomena across time, and build predictive models that capture those behaviors. Key behavioral phenomena also allow us to build more plausible models of the neural mechanisms of behavior, which help to inform neural recording experiments, such as what brain regions to target or what types of cells to look for. In the following sections we briefly cover related background from the perspective of hardware sensing, the feature representations of behavior, and prediction and theory, which is important for understanding the mechanisms of learning, and for use in many applied disciplines in the real world.

Chapter 2

EXPERIMENTAL SETUP

Learning and behavior are tightly coupled as they operate in a closed-loop system, since behavior affects the organism's sensory experience, which affects learning. In the early stages of learning a new task, the behavior itself will also differ substantially from the final learned behavior. Because the entirety of an animal's experience can potentially drive learning, it is useful to examine learning changes on a fine-grained level by fully characterizing the animal's behavior from naive to learned. Task design and sensing is a big component of this process, since it is important to work with tasks that allow for the possibility of fully characterizing the evoked behavior and track any relevant stimuli without occlusion or interruption. By designing tasks and sensing setups in a practical way, we could capture the entirety of the animal's experience, and subsequently use the collected behavioral data for analysis or developing theory.

2.1 Task Design

In many disciplines, in order to study a particular behavior under controlled lab settings, the experimenter first designs a task that could allow the behavior to conform to some trial-based structure, which makes it easier to model or analyze. Setting aside the technology required for sensing, the task itself is critical, as the behavior exhibited may differ depending on the perceived difficulty of the task to the subject, which depends on many factors.

In behavioral neuroscience, the model organism can range from insects, rodents, macaques, to humans [15, 63]. Each organism has specific evolutionary niches that make them more adept at certain tasks than others. In the earlier days of psychology, complex mazes were often used to study rodent learning, despite the lack of computational tools for tracking and quantifying behavior [57]. However, decades of reductionist experimental designs have led much of the neuroscience community to converge on tasks that are simple variations of the 2-AFC (two-alternative forced choice) task, which is a single-decision problem whereby there are two choices available for each trial, and typically the animal learns to associate the correct decision with a stimulus or several stimuli [43]. Such tasks have many uses and may be especially suited for understanding sensory thresholds or slow learning of simple but abstract concepts, but it may be difficult to answer questions related to the

hallmarks of behavior exhibited by biological intelligence: fast, complex learning.

The task design should also be guided strongly by the research question. For instance, to study certain innate behaviors, it may not be necessary to use a specialized task at all, as simply the presence of a salient stimulus, such as another mate, could bring about the behavior of interest [25]. Similarly, for perception related tasks which can involve quantifying saccades, it may be sufficient to head-fix the individual and record reactions to a large quantity of stimuli without considering the need for deliberation [5]. However, for decision-making tasks in animals, a key challenge is that the animal may not understand the purpose of the task, and thus perform arbitrarily poorly, especially when it comes to maximizing reward, a human-defined proficiency, typically characterized by percent correctness over many trials [11, 43]. This proficiency metric may not be the objective an animal optimizes, as animals may derive satisfaction from non-water-based sources, such as self-perceived novelty from exploring suboptimal policies [43]. Therefore, for tasks that require learned behavior, it is important for a task to have some purpose in order to gain insight into goal-oriented learning or decision-making.

2.2 Sensing Hardware

Data can be collected using hardware sensors that capture one or multiple modalities simultaneously, depending on the task, environment, and subject. For many organisms, the primary form of behavior is movement, and thus, most of the data could be collected with some form of optical imaging and saved into images or videos, from which structured data can be extracted using computer vision models, such as keypoints ([35]).

While vision is an important sensing modality, not all organisms behave similarly, as there may be other actions that organisms can use to interact with itself, others, or the environment. Humans tend to place strong priors on visual sensing as it is our primary sense, but for many animals, other modalities such as tactility, acoustics, pheromones, and odors are also strongly used for guiding behavior, and / or are part of the behavior itself. For instance, many insects emit pheromones as a way to communicate, and rats can use ultrasonic vocalization, both of which would be imperceptible to humans [10, 14, 51]. Some of these can be sensed by existing, mature hardware sensors, especially for imaging or acoustics, and some may require the development of novel sensing technologies and models [66]. In the case of detecting airborne molecules such as semiochemicals, a sensing device could be

metal oxide sensors such as the metal–oxide–semiconductor field-effect transistor (MOSFET), which has been explored as one potential method for sensing odor and pheromone compounds in the air [58]. Similar to how machine learning has enabled key computer vision models for tracking behavior through optical imaging, we should also think about modalities that are additionally ethologically relevant for certain organisms and develop tasks around those modalities. For species where optical imaging is not possible, other sensing tools can be developed to supplement existing optical tools for imaging. This could open up explorations into new areas of brain regions that are more relevant for behavior in the wild. Additionally, since behavior is driven largely by the sensory stimuli received by the nervous system, capturing the full set of environmental conditions can help explain some aspects of the behavior in naturalistic environments.

In this thesis, we focus entirely on optical imaging, where the primary behavior is movement and can be captured by pose estimation. In the future, with the increasing availability of sensing tools for other modalities, we could begin to image the entirety of an animal's behavioral set as well as the sensory inputs received by the animal, such as the possible generation of self-odor.

Chapter 3

HIERARCHY OF BEHAVIORAL FEATURE REPRESENTATIONS

To quantify behavioral states across time, we need to extract features from the raw data that are relevant for analysis. The representation of the extracted features should depend on the research question. The features themselves may also be useful downstream when building models for predicting behavior, since those models may be fitted directly to the behavioral features. While there is no single correct way to do this, all behavior can be characterized at multiple levels of abstractions, where with each increased level of abstraction the representation is further simplified to place emphasis on the most relevant parts of the behavior for the research question.

For example, in the case of video or image data collected from optical cameras, at the lowest level of the feature abstraction hierarchy is the raw video. This data can be very high dimensional, and depending on the task and camera view, the animal may not even occupy a large fraction of the captured view. Thus, it is useful to extract features from this data. One way to do that would be to extract the pose, composed of a series of keypoints tied to specific body-parts of the animal that are of particular interest to us. In a mice navigation task, those keypoints could include the nose, centroid, and tail-base. These keypoints can be extracted with off-the-shelf pose estimation models used in computer vision trained on just several hundred frames of supervised annotations [35]. With these keypoints, we can stack them and create trajectories that inform how the positions of these body landmarks evolve across time.

Pose keypoints are able to capture a great deal of information, which sometimes may not be directly related to the task. Many species are not simple deterministic machines, and thus often behave with alternate intents and goals from those directly related to solving a task, such as resting or grooming, especially when the task is unintuitive to the animal. By modeling the variability in behavior that exists in gory detail, it may be hard to derive a global understanding of the more task-relevant behavior. Therefore, it can be useful to introduce a more curated set of extracted features from the keypoints.

For example, in a complex maze task where animals would be subjected to many junctions where a decision could be made, a continuous trajectory of x-y coordinates

of a particularly important keypoint could be discretized into a uniform grid, from which the actual decision at each choice point could by computed [46]. The decision sequences can then be encoded in a vector of actions, $\mathbf{a}$, where a_t is the decision at decision step t. By analyzing behavior at this level of resolution, we can aggregate across more important task-relevant features and focus on the critical parts of behavior.

To go up an additional level, the sequence of decisions could be aggregated across the temporal dimension so that we could derive summary statistics of behavior with action probabilities that are specific to each decision junction in space. A caveat of this approach is the requirement for large quantities of decisions in the data, which may be implausible for certain tasks that only capture one behavior, such as homing [50]. While feature abstraction is important, it is still useful to examine data at all levels, as some may yield insight into previously unknown behaviors. It is wise to start with the least abstracted form of behavior and work one's way up in the analysis.

Chapter 4

PREDICTION AND THEORY

Building predictive models is important from a theoretical perspective, but can also be useful in the applied domain in many industries. Many forms of models could be used, broadly categorized into agent-based and observer-based models [16]. In agent-based modeling approaches, the model captures some underlying mechanistic process related to the agent and simulates its behavior based on the model of the internal dynamics. That simulated behavior can then be compared with real data to decide which model behaves more similarly. In observer-based models, the goals are to understand behavior from a less mechanistic perspective, by aggregating statistics or fitting statistical models that may have no biological realism to the data, but may still be useful from a purely statistical perspective. These models can capture the distribution of the underlying data directly and may be useful for examining changes in the distributions over the course of the experiment.

4.1 Neuroscience

In behavioral neuroscience, behavior is often first shaped through pretraining on a specified task, which can then guide the search for neural changes in the brain when recording directly from the brain using imaging or electrodes [29]. Finding these neural correlates of behavior is a fundamental goal of behavioral neuroscience, and understanding the neural mechanisms of behavior sometimes additionally requires the development of computational models of neural dynamics. From a neural theory perspective, prediction can help us understand the core tenets of task-relevant behavior. With the multiple levels in the abstraction of behavioral features, models and prediction can also be built and assessed at each level. On a fine-grained level of feature representation, behavior prediction can gather low-level insights about an organism's motor movements, such as stereotyped sequences of actions like mounting or attack [48]. In decision-making tasks with a denser form of features like sequences of actions, we can understand how animals deliberate and choose actions to achieve goals. And finally, by observing key phenomena, such as the ability of species to learn complex structures, we can propose theory about the neural circuits that make this behavior possible. Several candidate neural mechanistic models may then be used to guide the search for neural signals in the brain, as these models can

predict the types of neurons we may expect to find. By iterating on both the theory and the experiments, we may get closer to a mechanistic understanding of how the brain computes decisions and learns to efficiently solve tasks [65].

4.2 Other Disciplines

Behavioral models for prediction are also commonly used in many other diverse disciplines, such as psychiatry, ecology, and economics [18, 33, 39, 61]. In each discipline, prediction tends to operate at a different level of implementation, but is no less critical for both theoretical understanding and uses in applied settings. In pharmacology, behavior can be used as a proxy for inferring the effects of a particular pharmaceutical, task, or stimulus' effects of behavioral state changes or learning. In those settings, fine grained behavioral changes can be used as an indication of toxicity [26].

In other species-specific fields such as entomology, while the mechanism being studied may focus on other behaviors such as mating or dynamics such as population growth across generations, the models that have been developed retain the same goal of behavioral understanding. Additionally, the use of these dynamics models can go far beyond just theoretical understanding, as they can be applied in industries that require understanding insect behavior in areas such as insect control, to better target harmful insects, or protect beneficial ones [28, 34].

Chapter 5

REVIEW OF RELATED WORKS

Biological intelligence is characterized by many desirable properties shaped through evolutionary pressure for survival. Properties relevant in scope to this thesis are the ability to learn quickly and adapt robustly, which are critical functions for handling the nonstationary world around us. Questions around how this is accomplished in science are asked at various levels of implementation across multiple disciplines, particularly in psychology, cognitive science, and systems neuroscience. There are commonalities across disciplines on how human and animal experiments are designed, because most studies in this space rely on observing and quantifying some form of behavior [19, 41, 42]. One common method for studying learning is to use a decision-making task where some learned action(s) can be quantified, sometimes in addition to other sets of simultaneously observed variables in order to postulate factors that directly or indirectly relate to the behavior [13, 25, 27, 29, 42, 52].

The way psychology and cognitive science approach the question of learning is to develop theories around abstract cognitive processes underlying structural or reward learning, with behavior as the core inspiration [27, 31, 41, 55, 57]. While there may be some neural inspiration for these models, theorists in these disciplines are not confined by the immediate plausibility of biological implementations. This contrasts with systems and theoretical neuroscience, where the emphasis is on teasing apart the neural mechanisms and dynamics that give rise to or are the results of some behavior [2, 3, 13, 31]. Across disciplines, the goals can be similarly summarized as understanding the general mechanisms of learning, how ever "mechanism" may be defined in spatiotemporal scale. To start, the behavioral task enables comparisons of decision statistics across subjects or within one subject over the course of learning [13, 46]. The objective is to find a class of tasks where the subject could exhibit some nontrivial behavioral phenomenon which could then allow specialized cognitive models to be formulated based on [59]. This is followed by testing the proposed models of behavior on additional experimental tasks and building extensions of existing models to capture any unexplained behavior. Because of the reduced focus on neural mechanisms, neural recordings are not required as part of the data collection, allowing most studies in this space use human subjects. Using human subjects also enable the use of a broader space of tasks that rely on reasoning in abstract space

(e.g. language learning) as opposed to physical space, since the experimenter could pre-define some abstract goals for the subject [19, 41]. In cognitive neuroscience, the gap between cognition and neural substrates is narrowed through non-invasive neuroimaging setups that can indirectly record some form of coarse brain activity, usually at much lower resolutions than the imaging setups available in animals [17, 39, 40].

In systems neuroscience, the resolution of neural implementation is often addressed at the population or circuit level [5, 13, 31]. The finer-scale of operations presents unique challenges, as the commonly available tools for single-cell recording are invasive and less accessible in humans [2]. Additionally, the vast amount of available genetic tools for rodents make them a more suitable model organism for neural circuit-level manipulations [13, 32]. As an interdisciplinary field, neuroscience tends to adopt tasks that were originally used in psychology or cognitive science for studying learning [20] . These tasks from human psychology and cognition are adapted to mice through substantial simplification to make the learning problem easier, while retaining the independent trial-based structure and related core methods of analysis [4, 24, 42]. For example: a multi-armed bandit task designed for humans based on picking payouts in a contextual bandit task might be simplified to a just a binary choice task with the trial-start stimuli being an audible tone or a salient cue [4, 9, 64]. Another important consideration in systems neuroscience is the preference of experimenters to isolate the simplest learned behavior in order to make the subsequent data analysis when searching for interpretable neural correlates of that behavior in the high dimensional neural data more tractable [2, 13, 25].

In the last several decades, the combination of anthropomorphizing mice with human-inspired tasks and the fixation on behavioral reductionism have led much of the neuroscience community to converge on simple-learning tasks that counterintuitively also produced poor learning proficiencies [12, 43]. Unlike the types of complex learning often seen in nature, the types of learning in neuroscience are overwhelmingly variations of associative learning, where arbitrary stimuli are reinforced through tiny water rewards to associate with a simple behavior, such as nose-poking a port. Part of the cause for poor performance in such a simple task potentially lies in the tendency of experimenters to treat behavioral events across trials independently, which is highly implausible ethologically. In the real world, events often are correlated across time and space. It would not be surprising if the concept of “independent trials” is a difficult concept to grasp for a mouse. After all, it is more common for rewards to be

associated with space than arbitrary stimuli in nature. This problem is addressed through experiments and policy inference in Chapter 6.

To cite a specific example: the quintessential paradigm used in studying learning is the two-alternative forced choice (2-AFC) task, for which the subject learns to associate some stimuli with a decision in a binary choice task. Despite its simplicity, 2-AFC tasks can require thousands of trials over weeks for just one mouse to achieve 80% proficiency [12, 43]. With many mice, this can take up enormous experimental space and labor before neural recordings could even take place. There is an additional complication on the theory front due to the slowness and variability in learning. The inability of animals to comprehend cognition-inspired tasks makes it implausible to compare their behavior with predictions from more deliberate and sophisticated learning models proposed in cognitive science. These models are often adapted based on results from human experiments, where the subjects excel at understanding the task structure and can even describe their intentions post hoc.

Thankfully, not all hope is lost. An example of a task where animals excel at is navigation, which is embodied in many common scenarios. In mice, this could be learning to reliably navigate away from and back to the home burrow when finding resources. To rear their young, rodents might first determine a safe location to dig a burrow, then reliably forage for resources located at faraway locations, and return to the den to feed the pups—all while avoiding predation and adapting to any changes in the environment [60]. Feats like this likely recruit alternate mechanisms from those used in associative learning, and require the integration of various sources of sensory information to learn a map or to make complex decisions both for localization and trajectory planning. Questions in this space have been asked for decades in psychology, with an emphasis on defining and proposing what a "cognitive map" might look like [56].

While there have been many works that highlighted learned behavior in complex environments for mice, they represent a small minority due to the technical difficulties presented in freely-roaming neural recording and analysis, and there is often less clarity about how neural data could actually relate to the behavior over the course of learning [22, 49]. In studies that do use tasks that involve humans or animals navigating in a complex environment, they often feature environments with simple topologies, such as several connected compartments, or environments with a high degree of regularity, such as a maze with spatially-repeating motifs [1, 6, 62].

Perhaps due to the history of simple learning in experiments, of the theory developed

around learning and decision-making across disciplines, there remains a strong prior on leveraging traditional modeling frameworks that were originally proposed for modeling the cognitive processes of simple conditioning tasks [45, 54]. Such a framework is reinforcement learning (RL). RL was originally developed to model learning in classical and operant conditioning, and thus does not prescribe general methods for learning to reason about spatial or abstract concepts, nor does it prescribe clear paths for concrete neural implementations. In spite of this, over the last several decades, many theories based on extensions of classic RL algorithms have been developed, leveraging existing RL methods for value estimation or policy updates and strapping on complex cognition models to enable learning that depend less on environmental rewards [7, 8, 21, 30, 36, 38, 47]. A complication with these increasingly sophisticated models is that their overparameterization / complexity can be a burden for model falsification when comparing to behavioral data, which are often obtained from simple environments. The complexity of some decision models also makes it more challenging to imagine implementing with known biological networks [36], and thus, difficult to relate back to systems neuroscience.

On the other hand, circuit-level models in theoretical neuroscience for decision-making tend to focus on simple input-output mappings or involve complex temporal dynamics encoded via fixed-point or non-convergent recurrent dynamical networks [37, 44, 53]. These models often exhibit such complex dynamics that they are difficult to reproduce patterns robustly even in simulation. Additionally, they are frequently trained with a complex gradient-based optimizer to reproduce the some given ground-truth dynamics [53].

What is clearly needed are methods that can tie together the types of complex but fast learning behavior seen in nature with a new theory built from the ground-up for explaining non-reward based, map-like sequence learning. For tasks—there is a need to develop specialized complex tasks for animals that could evoke the type of learning possible in natural settings, which allows the few-shot learning of long sequences of actions. This might also mean moving away from the traditional type of rigid trial-based learning, where the start and end of each trial is predefined by the experimenter. Moving to a more continuous task would also require new analysis tools that are designed to handle non-trial-based data types. This could include the ability to track the occurrence of key learned behavioral phenomena throughout the entirety of learning, that can also later be analyzed with simultaneously recorded neural data. Additionally, there would need to be ways of assessing the frequency of

the learned behavior against some control behavior. A ethological task that achieves this with the relevant methods of analysis is explored in Chapter 7, while a general theory for learning a map for use in sequential decisions is proposed in Chapter 8.

Finally, methods for model validation should be assessed over the entire course of learning and across multiple tasks to characterize any sudden changes in behavior, since they could be directly related to synaptic learning. Especially in the early stages of learning, there may be rapid behavioral changes, and unsupervised analysis that could spot this would be more useful than tracking a single experimenter-defined measurement, such as choice probability computed over a large time window. While prior works have studied neural recordings in complex spaces, there is a disconnect between the level of behavioral understanding sought after in cognitive neuroscience and that in neuroscience [2, 13, 59].

To conclude, if one wanted to understand some general mechanisms related to learning, it could be more useful for the behavior to resemble the types of learning that biology is already adept at solving in nature, which gives us the confidence that it may recruit some general circuitry for learning across species. Perhaps then we could start moving away from reward-based learning theory and truly understand how goals are self-defined are then dynamically and flexibly solved in the real world. This thesis represents an effort to step in that direction.

References

[1] Alejandra Alonso, Levan Bokeria, Jacqueline van der Meij, Anumita Samanta, Ronny Eichler, Ali Lotfi, Patrick Spooner, Irene Navarro Lobato, and Lisa Genzel. "The HexMaze: A previous knowledge task on map learning for mice". In: *Eneuro* 8.4 (2021).

[2] Richard A Andersen and He Cui. "Intention, action planning, and decision making in parietal-frontal circuits". In: *Neuron* 63.5 (2009), pp. 568–583.

[3] David J Anderson and Pietro Perona. "Toward a science of computational ethology". In: *Neuron* 84.1 (2014), pp. 18–31.

[4] Ryo Aoki, Tadashi Tsubota, Yuki Goya, and Andrea Benucci. "An automated platform for high-throughput mouse behavior and physiology with voluntary head-fixation". In: *Nature communications* 8.1 (2017), pp. 1–9.

[5] Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. "A map of object space in primate inferotemporal cortex". In: *Nature* 583.7814 (2020), pp. 103–108.

[6] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Vıctor Valdés, Amir Sadik, et al. "Deepmind lab". In: *arXiv preprint arXiv:1612.03801* (2016).

[7] Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. "Reinforcement learning, fast and slow". In: *Trends in cognitive sciences* 23.5 (2019), pp. 408–422.

[8] Matthew Botvinick, Jane X Wang, Will Dabney, Kevin J Miller, and Zeb Kurth-Nelson. "Deep reinforcement learning and its neuroscientific implications". In: *Neuron* (2020).

[9] K. Branson, A. A. Robie, J. Bender, P. Perona, and M. H. Dickinson. "High–Throughput Ethomics in Large Groups of Drosophila." In: *Nat Methods* 6 (June 2009). The following values have no corresponding Zotero field:
auth-address: California Institute of Technology, Pasadena, California, USA.
number: 6
user16: PMC2734963
accession-num: 19412169, pp. 451–7. DOI: 10.1038/nmeth.1328.

[10] Stefan M Brudzynski. "Ethotransmission: communication of emotional states through ultrasonic vocalization in rats". In: *Current opinion in neurobiology* 23.3 (2013), pp. 310–317.

[11] Christopher P Burgess, Armin Lak, Nicholas A Steinmetz, Peter Zatka-Haas, Charu Bai Reddy, Elina AK Jacobs, Jennifer F Linden, Joseph J Paton, Adam Ranson, Sylvia Schröder, et al. "High-yield methods for accurate two-alternative visual psychophysics in head-fixed mice". In: *Cell reports* 20.10 (2017), pp. 2513–2524.

[12] Christopher P. Burgess, Armin Lak, Nicholas A. Steinmetz, Peter Zatka-Haas, Charu Bai Reddy, Elina A. K. Jacobs, Jennifer F. Linden, Joseph J. Paton, Adam Ranson, Sylvia Schröder, Sofia Soares, Miles J. Wells, Lauren E. Wool, Kenneth D. Harris, and Matteo Carandini. "High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice". In: *Cell Reports* 20.10 (Sept. 2017), pp. 2513–2524. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.08.047.

[13] M. Carandini and Anne K Churchland. "Probing Perceptual Decisions in Rodents." In: *Nature Neuroscience* 16 (July 2013), pp. 824–31. DOI: 10.1038/nn.3410.

[14] Ring T Carde and Albert K Minks. "Control of moth pests by mating disruption: successes and constraints". In: *Annual review of entomology* 40.1 (1995), pp. 559–585.

[15] Le Chang and Doris Y Tsao. "The code for facial identity in the primate brain". In: *Cell* 169.6 (2017), pp. 1013–1028.

[16] Donald L DeAngelis and Wolf M Mooij. "Individual-based modeling of ecological and evolutionary processes". In: *Annu. Rev. Ecol. Evol. Syst.* 36 (2005), pp. 147–168.

[17] Lesley K Fellows. "The cognitive neuroscience of human decision making: a review and conceptual framework". In: *Behavioral and cognitive neuroscience reviews* 3.3 (2004), pp. 159–172.

[18] Lucas A Garibaldi, Ingolf Steffan-Dewenter, Rachael Winfree, Marcelo A Aizen, Riccardo Bommarco, Saul A Cunningham, Claire Kremen, Luısa G Carvalheiro, Lawrence D Harder, Ohad Afik, et al. "Wild pollinators enhance fruit set of crops regardless of honey bee abundance". In: *science* 339.6127 (2013), pp. 1608–1611.

[19] James J Gibson. "A critical review of the concept of set in contemporary experimental psychology." In: *Psychological bulletin* 38.9 (1941), p. 781.

[20] Alex Gomez-Marin, Joseph J Paton, Adam R Kampff, Rui M Costa, and Zachary F Mainen. "Big behavioral data: psychology, ethology and the foundations of neuroscience". In: *Nature neuroscience* 17.11 (2014), pp. 1455–1462.

[21] David Ha and Jürgen Schmidhuber. "Recurrent World Models Facilitate Policy Evolution". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

[22] Christopher D Harvey, Forrest Collman, Daniel A Dombeck, and David W Tank. "Intracellular dynamics of hippocampal place cells during virtual navigation". In: *Nature* 461.7266 (2009), pp. 941–946.

[23] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. "Neuroscience-inspired artificial intelligence". In: *Neuron* 95.2 (2017), pp. 245–258.

[24] Ashley L Juavinett, Jeffrey C Erlich, and Anne K Churchland. "Decision-making behaviors: weighing ethology, complexity, and sensorimotor compatibility". In: *Current opinion in neurobiology* 49 (2018), pp. 42–50.

[25] Tomomi Karigo, Ann Kennedy, Bin Yang, Mengyu Liu, Derek Tai, Iman A Wahle, and David J Anderson. "Distinct hypothalamic control of same-and opposite-sex mounting behaviour in mice". In: *Nature* 589.7841 (2021), pp. 258–263.

[26] Ann Kennedy. "Computational behavior analysis takes on drug development". In: *Nature neuroscience* 23.11 (2020), pp. 1314–1316.

[27] Jeansok J Kim and Min Whan Jung. "Neural circuits and mechanisms involved in Pavlovian fear conditioning: a critical review". In: *Neuroscience & Biobehavioral Reviews* 30.2 (2006), pp. 188–202.

[28] AL Knight. "Adjusting the phenology model of codling moth (Lepidoptera: Tortricidae) in Washington state apple orchards". In: *Environmental entomology* 36.6 (2014), pp. 1485–1493.

[29] Chi-Tat Law and Joshua I Gold. "Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area". In: *Nature neuroscience* 11.4 (2008), pp. 505–513.

[30] Mark R Lepper and David Greene. *The hidden costs of reward: New perspectives on the psychology of human motivation*. Psychology Press, 2015.

[31] Dayu Lin, Maureen P Boyle, Piotr Dollar, Hyosang Lee, ES Lein, Pietro Perona, and David J Anderson. "Functional identification of an aggression locus in the mouse hypothalamus". In: *Nature* 470.7333 (2011), pp. 221–226.

[32] Liqun Luo, Edward M Callaway, and Karel Svoboda. "Genetic dissection of neural circuits: a decade of progress". In: *Neuron* 98.2 (2018), pp. 256–281.

[33] Tiago V Maia and Michael J Frank. "From reinforcement learning models to psychiatric and neurological disorders". In: *Nature neuroscience* 14.2 (2011), pp. 154–162.

[34] Thomas J Manetsch. "Time-varying distributed delays and their use in aggregative models of large systems". In: *IEEE Transactions on systems, man, and cybernetics* 8 (1976), pp. 547–553.

[35] Alexander Mathis, 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". en. In: *Nature Neuroscience* (Aug. 2018), p. 1. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0209-y.

[36] Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. "Deep neuroethology of a virtual rodent". In: *arXiv preprint arXiv:1911.09451* (2019).

[37] Thomas Miconi. "Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks". In: *Elife* 6 (2017), e20899.

[38] Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. "The successor representation in human reinforcement learning". In: *Nature human behaviour* 1.9 (2017), pp. 680–692.

[39] Thomas Naselaris, Danielle S Bassett, Alyson K Fletcher, Konrad Kording, Nikolaus Kriegeskorte, Hendrikje Nienborg, Russell A Poldrack, Daphna Shohamy, and Kendrick Kay. "Cognitive Computational Neuroscience: a new conference for an emerging discipline". In: *Trends in cognitive sciences* 22.5 (2018), pp. 365–367.

[40] John O'Doherty, Morten L Kringelbach, Edmund T Rolls, Julia Hornak, and Caroline Andrews. "Abstract reward and punishment representations in the human orbitofrontal cortex". In: *Nature neuroscience* 4.1 (2001), pp. 95–102.

[41] John P O'Doherty, Jeffrey Cockburn, and Wolfgang M Pauli. "Learning, reward, and decision making". In: *Annual review of psychology* 68 (2017), pp. 73–100.

[42] Talmo D Pereira, Joshua W Shaevitz, and Mala Murthy. "Quantifying behavior to understand the brain". In: *Nature neuroscience* 23.12 (2020), pp. 1537–1549.

[43] Mu Qiao, Tony Zhang, Cristina Segalin, Sarah Sam, Pietro Perona, and Markus Meister. "Mouse Academy: high-throughput automated training and trial-by-trial behavioral analysis during learning". In: *bioRxiv* (2018), p. 467878. URL: https://www.biorxiv.org/content/10.1101/467878v2.

[44] Kanaka Rajan, Christopher D Harvey, and David W Tank. "Recurrent network models of sequence generation and memory". In: *Neuron* 90.1 (2016), pp. 128–142.

[45] Robert Rescorla and Allan R. Wagner. "A Theory of Pavlovian Conditioning : Variations in the Effectiveness of Reinforcement and Nonreinforcement". In: *Classical Conditioning II: Current Research and Theory*. Ed. by A. H. Black and Prokasy W. F. New York: Appleton-Century-Crofts, 1972, pp. 64–69.

[46] Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. "Mice in a Labyrinth Exhibit Rapid Learning, Sudden Insight, and Efficient Exploration". In: *eLife* 10 (July 2021). Ed. by Mackenzie W Mathis, e66175. ISSN: 2050-084X. DOI: 10.7554/eLife.66175.

[47] Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw. "Predictive representations can link model-based reinforcement learning to model-free mechanisms". In: *PLoS computational biology* 13.9 (2017), e1005768.

[48] Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J Sun, Pietro Perona, David J Anderson, and Ann Kennedy. "The Mouse Action Recognition System (MARS): a software pipeline for automated analysis of social behaviors in mice". In: *BioRxiv* (2020).

[49] Sunita Sharma, Sharlene Rakoczy, and Holly Brown-Borg. "Assessment of spatial memory in mice". In: *Life sciences* 87.17-18 (2010), pp. 521–536.

[50] Keith T Sillar, Laurence D Picton, and William J Heitler. *The neuroethology of predation and escape*. John Wiley & Sons, 2016.

[51] Keith N Slessor, Mark L Winston, and Yves Le Conte. "Pheromone communication in the honeybee (Apis mellifera L.)" In: *Journal of chemical ecology* 31.11 (2005), pp. 2731–2745.

[52] Chess Stetson and Richard A Andersen. "The parietal reach region selectively anti-synchronizes with dorsal premotor cortex during planning". In: *Journal of Neuroscience* 34.36 (2014), pp. 11948–11958.

[53] David Sussillo and Larry F Abbott. "Generating coherent patterns of activity from chaotic neural networks". In: *Neuron* 63.4 (2009), pp. 544–557.

[54] Richard S Sutton. "Learning to predict by the methods of temporal differences". In: *Machine learning* 3.1 (1988), pp. 9–44.

[55] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. "How to grow a mind: Statistics, structure, and abstraction". In: *science* 331.6022 (2011), pp. 1279–1285.

[56] E. C. Tolman. "Cognitive Maps in Rats and Men". In: *Psychological Review* 55.4 (1948), pp. 189–208. ISSN: 0033-295X. DOI: 10.1037/h0061626.

[57] Edward Chace Tolman and C. H. Honzik. ""Insight" in Rats". In: *Univ. Calif. Publ. Psychol.* 4 (1930), pp. 215–232.

[58] Anthony PF Turner and Naresh Magan. "Electronic noses and disease diagnostics". In: *Nature Reviews Microbiology* 2.2 (2004), pp. 161–166.

[59] Felix Warneken and Michael Tomasello. "Extrinsic rewards undermine altruistic tendencies in 20-month-olds." In: *Developmental psychology* 44.6 (2008), p. 1785.

[60] Jesse N. Weber, Brant K. Peterson, and Hopi E. Hoekstra. "Discrete Genetic Modules Are Responsible for Complex Burrow Evolution in Peromyscus Mice". en. In: *Nature* 493.7432 (Jan. 2013), pp. 402–405. ISSN: 1476-4687. DOI: 10.1038/nature11816.

[61] Nick Wilkinson and Matthias Klaes. *An introduction to behavioral economics*. Macmillan International Higher Education, 2017.

[62] Jonathan J Wilson, Elizabeth Harding, Mathilde Fortier, Benjamin James, Megan Donnett, Alasdair Kerslake, Alice O'Leary, Ningyu Zhang, and Kate Jeffery. "Spatial learning by mice in three dimensions". In: *Behavioural brain research* 289 (2015), pp. 125–132.

[63] Michael M Yartsev. "The emperor's new wardrobe: rebalancing diversity of animal models in neuroscience research". In: *Science* 358.6362 (2017), pp. 466–469.

[64] Shunan Zhang and Angela J Yu. "Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting". In: *Advances in neural information processing systems* 26 (2013).

[65] Tony Zhang, Matthew Rosenberg, Pietro Perona, and Markus Meister. "Endotaxis: A Universal Algorithm for Mapping, Goal-Learning, and Navigation". In: *bioRxiv* (2021). URL: https://www.biorxiv.org/content/10.1101/2021.09.24.461751v1.

[66] Tony Zhang, Szymon Zmyslony, Sergei Nozdrenkov, Matthew Smith, and Brandon Hopkins. "Semi-Supervised Audio Representation Learning for Modeling Beehive Strengths". In: *arXiv preprint arXiv:2105.10536* (2021).

Part II

Methods and Tools

Chapter 6

MOUSE ACADEMY: HIGH-THROUGHPUT AUTOMATED TRAINING AND TRIAL-BY-TRIAL BEHAVIORAL ANALYSIS DURING LEARNING

[1] Mu Qiao, Tony Zhang, Cristina Segalin, Sarah Sam, Pietro Perona, and Markus Meister. "Mouse Academy: high-throughput automated training and trial-by-trial behavioral analysis during learning". In: *bioRxiv* (2018), p. 467878. URL: https://www.biorxiv.org/content/10.1101/467878v2.

Progress in understanding how individual animals learn will require high-throughput standardized methods for behavioral training but also advances in the analysis of the resulting behavioral data. In the course of training with multiple trials, an animal may change its behavior abruptly, and capturing such events calls for a trial-by-trial analysis of the animal's strategy. To address this challenge, we developed an integrated platform for automated animal training and analysis of behavioral data. A low-cost and space-efficient apparatus serves to train entire cohorts of mice on a decision-making task under identical conditions. A generalized linear model (GLM) analyzes each animal's performance at single-trial resolution. This model infers the momentary decision-making strategy and can predict the animal's choice on each trial with an accuracy of ~80%. We also assess the animal's detailed trajectories and body poses within the apparatus. Unsupervised analysis of these features revealed unusual trajectories that represent hesitation in the response. This integrated hardware/software platform promises to accelerate the understanding of animal learning.

6.1 Introduction

Learning—the change of neural representation and behavior that results from past experience and the consequences of actions—is important for animals to survive and forms a central topic in neuroscience [35]. Different individuals may apply different strategies to the learning process, reflecting their individual personalities. Indeed, substantial differences in sensory biases, locomotion, motivation, and cognitive competence have been observed in populations of fruit flies [7, 18], rodents and primates [3, 25, 36]. Thus, it is critical to investigate learning at the individual level.

Rodents, especially the mouse, have become popular experimental animals in studying associative learning and decision-making, because of the wide availability of transgenic resources [10, 16, 22, 23]. They can learn to perform complex decision-making tasks that probe cognitive components such as working memory and selective attention [1, 29, 42]. However, differences in learning strategies across individuals have rarely been addressed, partly owing to the limitations of data gathering and analysis.

Studying differences among individuals requires training and collecting data from multiple animals in a standardized and high-throughput fashion. The training procedures are often time-consuming, requiring several days to many weeks [11, 16], depending on the task. Although there have been advances in training automation, existing systems either require an experimenter to move animals from the home cage to the training apparatus [9, 30, 39], or training animals within their own cages [15, 28, 33]. The former introduces additional sources of variability [12, 20], and the latter precludes tasks that require a large training arena. Following data acquisition, the analysis of behavior aims at understanding the learning process. Present approaches tend to focus on the averaged performance over many trials [20]. However, changes in behavior may happen at a single trial, and thus the modeling of behavior should similarly offer a time resolution of single trials to assess each animal's individual approach to learning.

To address these challenges, we present Mouse Academy, an integrated platform for automated training of group-housed mice and analysis of behavioral changes in learning a decision-making task. We designed hardware that makes use of implanted radio frequency identification (RFID) chips to identify each mouse, and guides the animal into a behavior training box. Synchronized video recordings and decision-making sequences are acquired during animal learning. To analyze the decision-making sequences, we developed an iterative generalized linear model

(GLM). This model makes a prediction of the animal's choice in each trial and gets updated based on the animal's actual choice. This iterative GLM model achieves a prediction accuracy of ~80%, and also reveals the decision-making strategy of the animal and how it changes over time. To analyze the animal's behavior during the task in greater detail, we computed the movement trajectories of each mouse. These trajectories allowed us to perform an unsupervised analysis of each animal's behavior, and discover individual traits of behavioral learning that were not apparent from the simple choice sequences.

6.2 Results

The Mouse Academy platform consists of three components (Figure 6.1): an automated RFID sorting and animal training system, an iterative GLM to analyze decision-making sequences, and behavior-assessment software that analyzes animal trajectories computed from video data.

Automated RFID sorting supports individual training programs

We designed the equipment in the following manner (Figure 6.1): RFID-tagged mice are grouped in a common home cage where food and bedding is supplied. The home cage connects to a behavior training box through a gated tunnel. The gates are controlled by a home-made RFID animal sorting system [43]: three RFID antennas are placed along the tunnel, with one near the home cage, one near the training box and one between the two; the motorized gates are placed between the RFID sensors, separating the tunnel into three compartments. An Arduino microcontroller integrates information from the RFID readers to open and shut the gates, allowing only one animal at a time to pass through the tunnel (Figure 6.9 a, c, d). The behavior box is outfitted with three ports, each of which contains a photo-transistor to detect snout entry, a solenoid valve to deliver water reward, and a light emitting diode (LED) to present visual cues. To maintain a controlled environment, the training box is isolated from the outside by a light- and sound-proof chamber (Figure 6.9b).

Once a mouse enters the training box, a protocol is set up to train the mouse to perform a certain task. In the experiments reported here, the animal must nose-poke the center port to initialize a trial and then hold the position for a short period. Visual or auditory stimuli are delivered, and based on these stimuli, the animal must choose to poke one of the side ports. If the correct response is chosen, the animal gets a water reward from a lick tube in the response port, otherwise a timeout punishment is applied. This training process is controlled by Bpod, an Arduino microcontroller

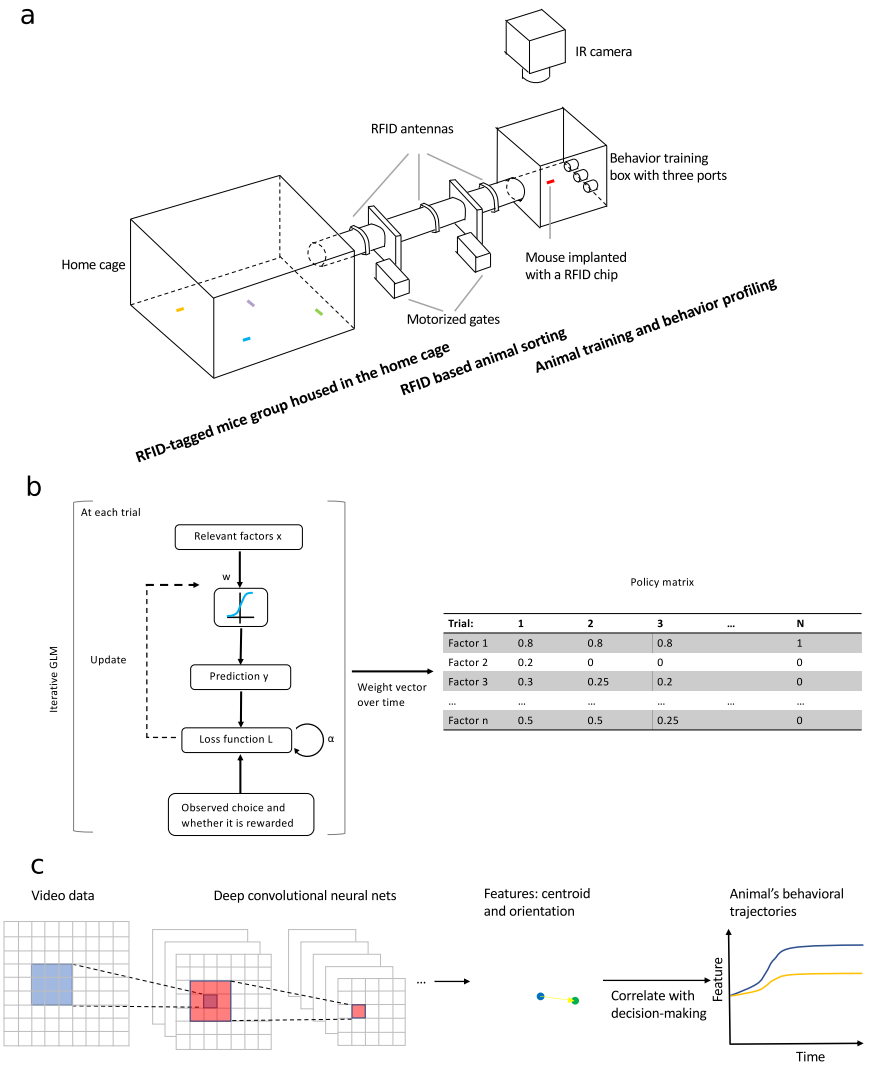

Figure 6.1: **Components of Mouse Academy**. (a) An automated RFID sorting and animal training system. Mice implanted with RFID chips are group-housed in the home cage. The RFID sorting system identifies each mouse by its implanted chip. One animal at a time gains access to a behavioral training box. As the animal learns a task, its decision sequences and video recordings are acquired. (b) An iterative generalized linear model. For each trial, the model predicts the animal's choice based on the relevant factors and then evaluates the difference from the actual choice. This difference, after temporal weighting, is fed back to the loss function, which gets minimized by updating the weights of the input factors. The model produces a policy matrix in which the rows indicate the weights of the relevant factors and the columns are the trials. (c) An automated behavior assessment program using deep convolutional neural networks to extract the location and pose information of an animal.

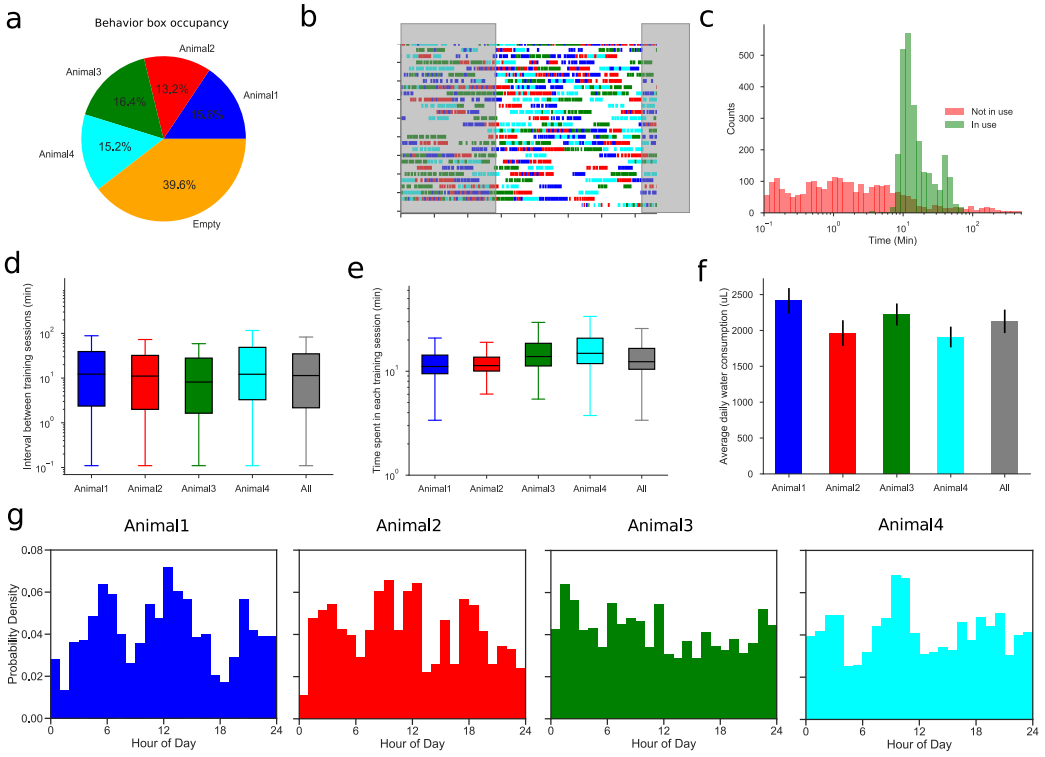

Figure 6.2: **Performance of the automated training system on a sample cohort**. (a) Fraction of time the behavior box was occupied by each of the four animals. (b) Activity trace of each animal in the behavior box for the entire training period of 28 days. Shadow indicates the dark cycle from 8pm to 8am. (c) Distribution of time intervals during which the behavior box is occupied or empty. (d) Box plot of intervals between each animal's sessions (median, quartiles, and range). (e) Box plot of the time spent in a session for each animal. (f) Averaged daily water consumption of each animal. Error bars indicate standard errors. (g) Circadian histograms of each animal's activity in the behavior box.

that interfaces with the three ports. Data from the response ports as well as video recordings from an overhead camera are acquired simultaneously as the animal is trained.

The entire apparatus is orchestrated by a master program that coordinates the RFID sorting device, the Bpod system, synchronized video recording, data management and logging (Figure 6.9e). The program monitors the amount of water each animal consumes per day and regulates the time each animal can spend in the training box per session. In addition, the software updates the training protocol for each animal based on its performance, for example switching to a harder task once a simpler one has been mastered (Figure 6.10). This lets each animal learn at its own pace.

The apparatus can be assembled at a materials cost of $1500–2500, with the cheaper option using a Raspberry Pi computer as the controller (Figure 6.9). Compared with designs in which each animal is automatically trained in its own home cage [28, 30], the system saves considerable space. Because housing and training are independent modules, the same system can be used for diverse training environments.

We tested the automated RFID sorting and animal training system by training group-house mice to learn a variety of decision-making tasks, following similar procedures as reported previously [22, 29] (Figure 6.10 and **Online Methods**). The training period lasted 28 days, with up to five mice in the common home cage. Each animal occupied the training box for 3-4 hours per day (13-15% of the 24 hours) throughout the entire training period (Figure 6.2a, b, Figure 6.11). For a sample cohort of four animals trained in sessions of 90 trials each, we found that the behavior box was occupied most of the time, with brief empty intervals of <10 min (Figure 6.2c, d, e). Each animal was trained for over 900 trials (10 sessions), and consumed more than 1.9 mL of water per day (Figure 6.2f). Interestingly there was no circadian pattern to the animals' training activity, even though the setup was illuminated on a daily light cycle (12 h on / 12 h off) (Figure 6.2g). As observed previously, it appears that animals working for a goal can avoid circadian modulation of the locomotor pattern [13, 21].

A generalized linear model accurately predicts decision-making during training
In a decision-making task, an animal is asked to associate distinct stimuli with distinct responses. Although this is the ultimate goal, during learning, it is often observed that the animal begins by basing its decisions on unrelated input variables and gradually switches to using the stimulus variables that actually predict reward. We define a policy as a mapping of these variables to the animal's decisions. A fundamental goal in the study of learning is to infer what policy the animal follows at any given time and to determine how the policy evolves with experience.

We applied a generalized linear model (GLM) to map factors relevant to the animal's decision-making to its choices through logistic regression. A common way to build such a GLM is by fitting data of an entire session [9, 31]. However, this loses resolution in single trials within the session. During learning, a change of policy can happen at each trial. Thus, we developed the model to make trial-by-trial choice predictions based on various factors the animal might plausibly use. The model works in an iterative two-step process (Figure 6.1b). In the prediction step, the model

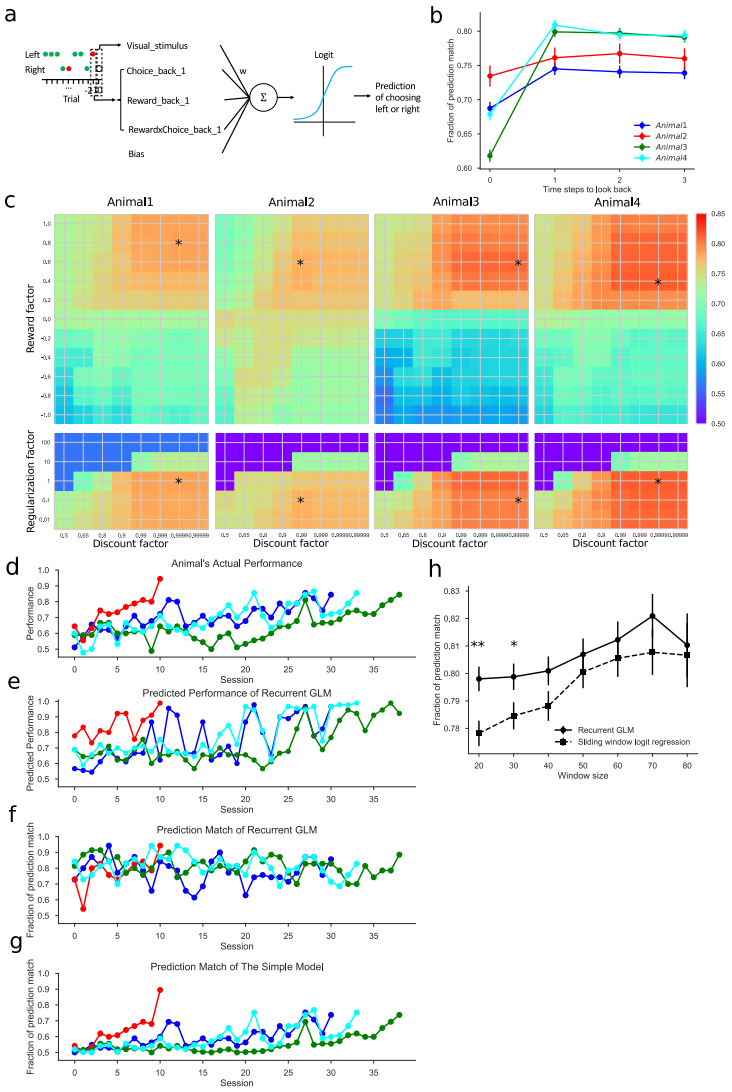

Figure 6.3: **Iterative generalized linear model and its prediction accuracy**. (a) Illustration of the GLM as applied to a visual discrimination task. The model's prediction is based on the output of a logistic function whose input is the weighted sum of a visual stimulus term, a bias term, and three history dependent terms. The stimulus can be on the left or right and the choice can be rewarded (consistent with the stimulus, indicated by a green dot) or unrewarded (opposite to the stimulus, indicated by a red dot). (b) Selection of the history dependent terms based on the model prediction accuracy. Error bars indicate standard errors. (c) Hyperparameters for each of the animals: reward factor, discount factor, and regularization factor. The optimal values are marked with a star. (d) The actual performance of each animal over time in the visual task. (e) Performance as predicted by the GLM. (f) Fraction of choices predicted correctly by the GLM. (g) Fraction of choices predicted correctly by a simple model based on the animal's average performance in the task. (h) Fraction of predictions matched by the iterative GLM and the sliding window logistic regression model. Error bars indicate standard errors. **, * indicate P < 0.01, 0.05. Random prediction would give 50% match.

makes a prediction for the next decision based on the input factors. Once the outcome of the animal's decision is observed, an error term between the model's prediction and the observation is computed. This error, after weighting by a reward factor and a temporal discount factor, is fed back to the loss function. In the update step, the model is updated by minimizing the regularized loss function. This iteration happens after every trial. The temporal discount factor accounts for the possibility that the most recent trials impact the current decision more than remote trials. The reward factor accounts for the fact that water rewards and timeout punishments may have effects of different magnitude on the updates of the animal's policy.

We illustrate the utility of this model by fitting results from an easy visual task, in which one of the two choice ports lights up to indicate the location of the reward, and the optimal policy is to simply poke the port with the light (Figure 6.10a, a', a"). All the mice eventually reached a >83% performance level, comparable to what mice achieve in similar tasks [8, 15]. The GLM makes a prediction for the outcome of each trial based on a weighted combination of several input variables: the current visual stimulus, a constant bias term, and three terms representing the history of previous trials (Figure 6.3a). These inputs from a previous trial include the port choice, whether that choice was rewarded, and a term indicating the multiplicative interaction between the choice and reward (Choice × Reward). This term supports a strategy called win-stay-lose-switch (WSLS), which chooses the same port if it was rewarded previously and the opposite one if not. Since a GLM cannot multiply two inputs, we provided this interaction term explicitly. Each of the above terms has a weight coefficient that can be positive or negative. For instance, a positive weight for the visual stimulus supports turning toward the light, and a negative weight away from the light.

To determine the extent of trial history that affects the animal's behavior, we fitted the model to the response data including history dependent terms up to three previous trials. We found that only the immediately preceding trial had an appreciable effect on the prediction accuracy, and thus restricted further analysis to those inputs (Figure 6.3b). The model also has three hyperparameters (the temporal discount factor α, the reward factor r, and the regularization factor λ), and we optimized them for each animal by grid search. We found that each animal had a different set of hyperparameters, reflecting differences in the learning process across individuals (Figure 6.3c). Among the four sample mice, Animal 2 had the lowest temporal discount factor, suggesting that it weighed recent trials more heavily and updated the

policy more quickly. Indeed, this is the animal that learned the fastest among the four (Figure 6.3d).

Predictions from the iterative GLM matched ~80% of the animals' actual choices (Figure 6.3f), and the predicted accuracy of each animal captured the actual fluctuations of its learning curve (Figure 6.3e and Figure 6.12). We compared the performance of the GLM with two other modeling approaches (**Online Methods**). The first model was fit to the animal's average performance in the task; its trial-by-trial match of the animal's actual choices was only ~59% (Figure 6.3g). The second model was a logistic regression fitted to data in a sliding window of N trials. This sliding window model performed worse than the iterative GLM when the window size was small (N = 20 and 30 trials, Figure 6.3h); for larger windows the performance was comparable. Overall, the iterative model is advantageous because it makes predictions online as every trial occurs and adapts dynamically to the growing data set.

Individual learning policies can be inferred from iterative GLM fitting

The iterative GLM serves to infer what policy the animal follows in making decisions. The linear weight of each input term reflects its relative importance for the decision. By following this weight vector across trials, one obtains a policy matrix that documents how the animal's policy changes during learning (Figure 6.1b, Figure 6.4c). To test that the model can correctly capture a time-varying policy, we simulated decision-making data from a ground truth policy that changed at a certain frequency, including a certain level of noise in the behavioral output (Figure 6.4a). Over a wide range of policy change frequencies and noise levels, the GLM was able to capture the ground truth policy (Figure 6.4a, b). In addition, different values of policy change frequency and noise levels led to different sets of hyperparameters fitted from the model, showing that the GLM can adapt to individuals with diverse learning characteristics (Figure 6.13a–e).

We then recovered the policy matrix of each animal from the GLM fits. All four animals started with the non-optimal policy of WSLS. Subsequently each animal followed its own learning process (Figure 6.4c): Animal 2 had a clear bias towards the right port at the beginning but it rapidly found the optimal policy of following the light. The other three animals were slower learners. Animal 3 and Animal 4 followed similar processes to converge to the optimal policy. Animal 1 was distinct from the others. At the early stages, it had a strong bias toward the left port, and it made decisions based on whether the previous choice was rewarded.

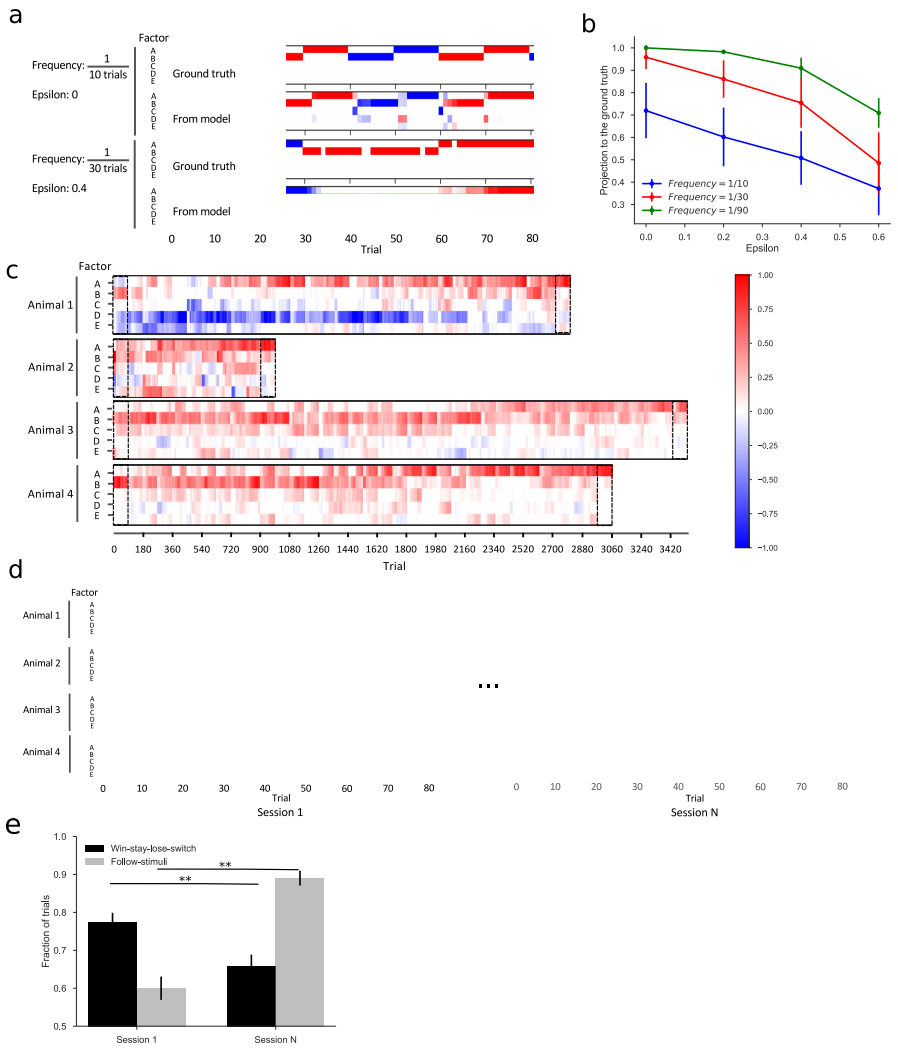

Figure 6.4: **Interpretation of policies during learning**. (a) Policy vectors recovered by the iterative GLM capture the ground truth policies. The policy matrix plots in each trial (horizontal) the weights associated with each of five factors (vertical), encoded with a color scale (see Panel c). The factors are: A = Visual_stimulus, B = Choice × Reward_back_1, C = Choice_back_1, D = Reward_back_1, E = Bias. Two examples are shown of ground truth policies used to simulate data and the corresponding trial-by-trial estimates from the GLM. Blanks in the ground truth matrix indicate instances where the simulated choice is opposite to the policy. (b) Similarity between the recovered policy and the ground truth, measured by the cosine between the two policy vectors. Error bars indicate standard deviation. (c) Policy matrices recovered for the four animals show distinct individual learning processes. Dashed rectangles highlight the first and last sessions of each animal, as enlarged in d. (d) Recovered policy matrices for the first and last sessions of each animal. (e) Fraction of trials explained by two candidate policies (win-stay-lose-switch and following the stimuli) in the first and last sessions. Error bars indicate standard errors. ** indicates $P < 0.01$.

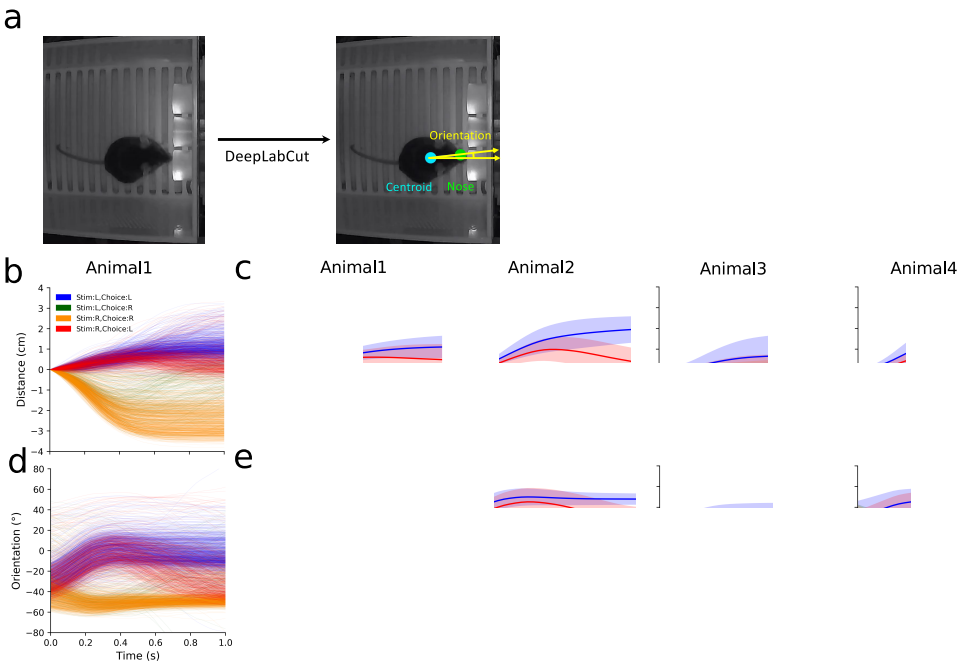

Figure 6.5: **Supervised analysis using features extracted by automated behavior assessment**. (a) DeepLabCut extracts the centroid and the orientation as the angle between the horizontal axis and the line connecting the centroid and the nose. (b) Centroid distance along the left-right axis vs time during the movement, for animal 1. The starting position is set to zero, positive values indicate movement to the left, negative to the right. The four trial types are indicated by different colors. (c) Average centroid trajectory for each animal. Shaded region indicates standard error. (d-e) Orientation vs time, displayed as in panels b-c. Positive angle points to the left, negative to the right.

We further validated the transition between policies during learning by analyzing the first and last sessions of each animal and counting how many choices could be explained by each policy (Figure 6.4d). Indeed, we found a clear switch from the (non-optimal) WSLS policy to the (optimal) stimulus-based policy (Figure 6.4e and Figure 6.13f). The animals might have been biased towards the WSLS strategy by a shaping method we used during training, which offered the animal a repeat of the same stimulus every time it made a mistake (**Online Methods**). To test whether these correlations in the trial sequence influenced the final policy we performed two additional analyses. First, we only included trials following a correct trial, and performed logistic regression on these trials for each session. This analysis showed that at least on these trials, all the animals based their decisions on the light stimulus by the end of learning (Figure 6.14a). Second, we compared the error rate on trials following an incorrect choice with that following a correct one. We found

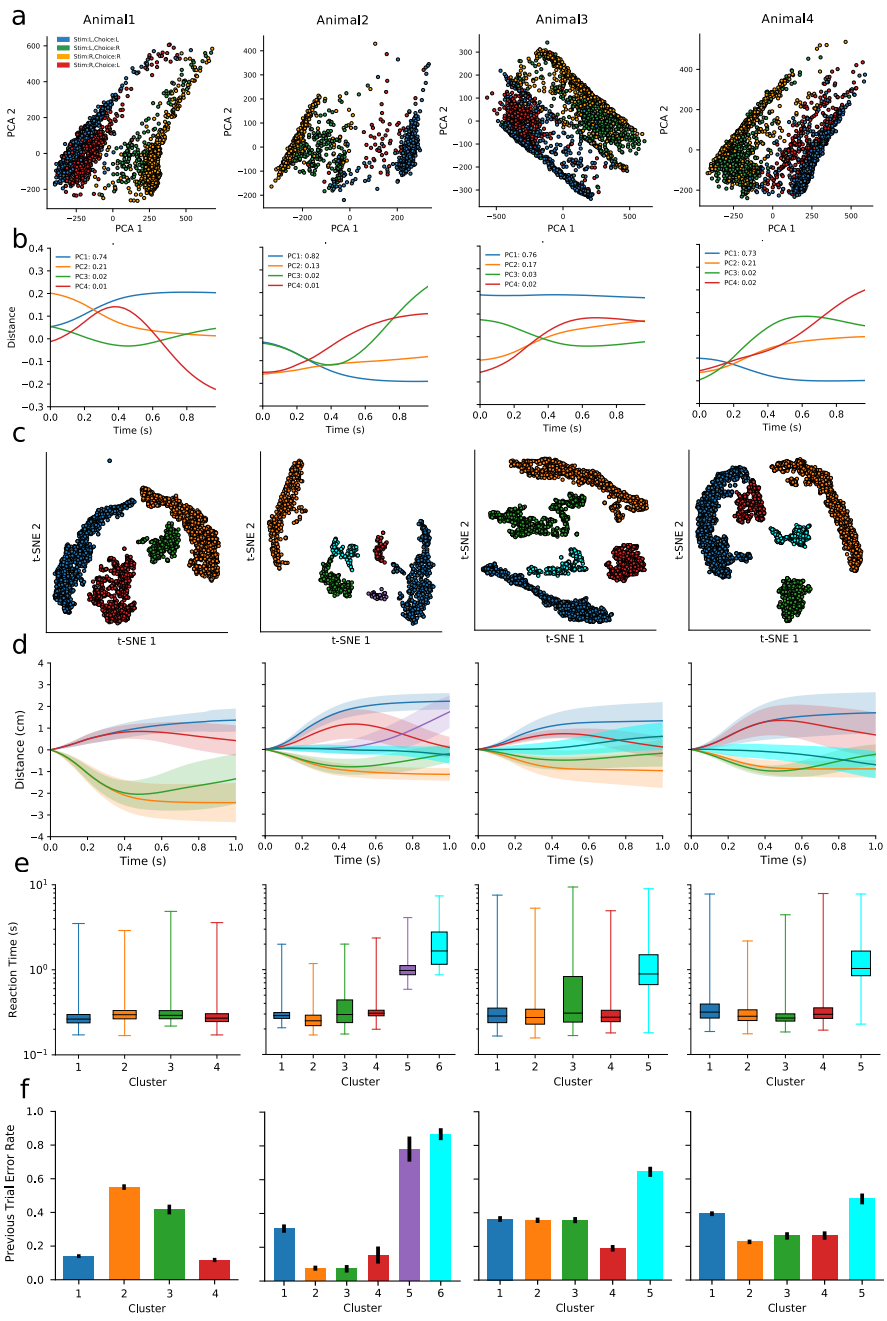

Figure 6.6: **Unsupervised analysis of the behavior trajectories**. (a) Principal component projections onto PC1 and PC2 of the centroid-vs-time trajectories from Figure 5. The four trial types are indicated by different colors. (b) The centroid trajectories corresponding to the first four principal components (PCs). The variance explained by each PC is shown in the plot legend. (c) Clustering trials by their trajectories using t-SNE analysis. Distinct clusters are marked with different colors for use in subsequent panels. (d) Averaged centroid distance vs time for each cluster, plotted as in Figure 5b. (e) Box plot of the reaction time for each cluster. (f) The error rate on the preceding trial for each cluster. Error bars indicate standard errors.

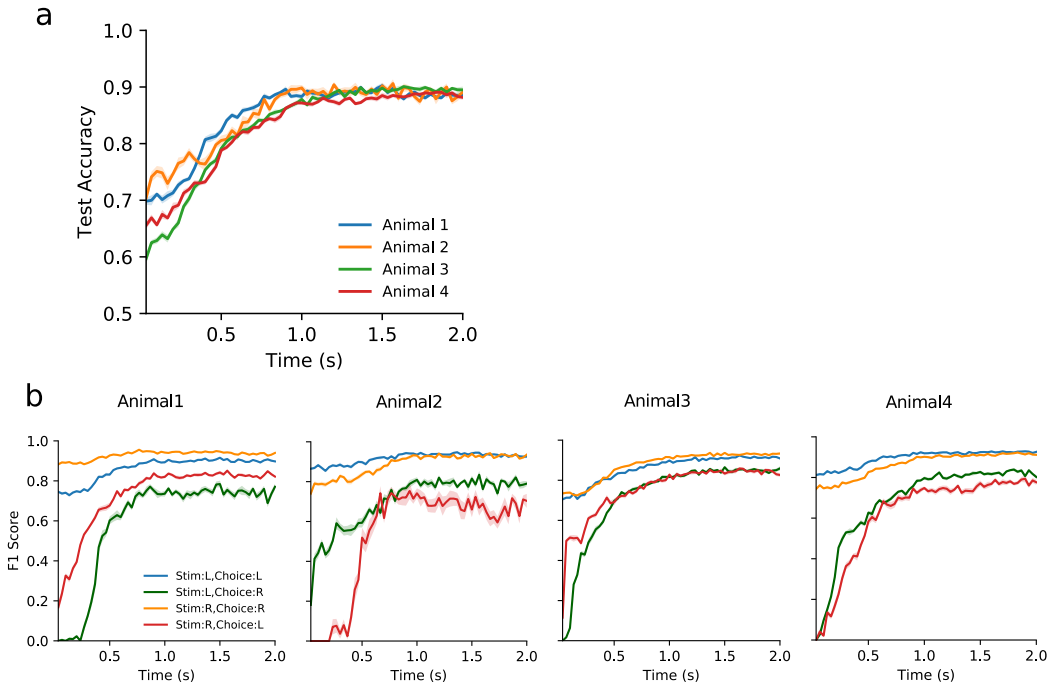

Figure 6.7: **Performance of the support vector machine to infer trial category from mouse trajectories**. (a) Prediction accuracy of the SVMs for individual animals. (b) F1 score of the SVM fitted for the decision categories of each animal. Shaded region denotes standard error. x-axis indicates the time starting from when the animal leaves the center port to make a choice. SVMs were trained using features up to a certain time point.

no significant difference between the two error rates during the last session (Figure 6.14b, c), suggesting that the animals treated these two types of trials identically.

Automated movement tracking reveals fine structure of behavioral responses

Thus far the report has focused on the animal's responses only as sensed by the nose pokes into response ports. The GLM fits of those responses already revealed differences in policy across individuals. To gain further insight into these individual preferences, it is essential to track each animal's behavior along the way from stimuli to responses [22]. We thus employed computer vision software to automatically, quantitatively, and accurately assess each animal's behavior during decision-making. Having compared several tracking algorithms, we eventually used DeepLabCut [24], a deep learning-based program, because it is easy to use and accurate in identifying body landmarks (Figure 6.1c, Figure 6.5a). We identified two body landmarks: the nose and the centroid of the animal, and further calculated the orientation as the angle of the line connecting the centroid and the nose (Figure 6.5a).

To illustrate use of these behavioral trajectories, we focus on the period of the visual choice task where the animal reports its decision: from the time it leaves the center port to when it pokes one of the side ports. The trials fall into four groups based on location of the stimulus and the response. As expected, the trajectories of position and orientation clearly distinguish left from right choices (Figure 6.5b, d). Interestingly, the trajectories also reveal whether the decision was correct: On incorrect decision, the trajectories reversed direction after ~0.5 s, because the animal quickly turned back to the center after finding no reward in the chosen port (Figure 6.5c, e). A linear kernel support vector machine (SVM), trained to predict the category of each trial from a 1 s trajectory, was able to correctly distinguish correct and incorrect choices with an accuracy of over 90% (Figure 6.15). In addition, many of the trajectories were highly asymmetric and again revealed differences across individuals. For instance, Animal 2 and Animal 4 started from a location close to the right port, Animal 1 closer to the left port (Figure 6.5c). This asymmetry correlates with the bias revealed by the iterative GLM: each animal prefers to select the port closer to its body location.

Unsupervised behavioral analysis reveals moments of hesitation

Whereas the supervised learning discussed above relies on prior classification of stimuli and responses, an unsupervised analysis has the potential to discover unexpected structures in the animal's behavior [17]. We thus performed an unsupervised classification of the behavioral trajectories.

After subjecting all the trajectories of a given animal to principal component analysis (PCA) we projected the data onto the top three components, which explained over 95% of the variance (Figure 6.6a, b). Importantly, without any labels from trial types, these three PCs captured meaningful features that differentiated the animal's responses. The first PC separated movements to the left from those to the right. The third PC captured the turning-back behavior after an incorrect choice. The second PC captured different baseline positions. Each animal has its own preference for a baseline position somewhere off the midline of the chamber (Figure 6.16a).

We also projected the trajectories into 2 dimensions using a non-linear embedding method, t-distributed stochastic neighbor embedding [5, 38]. Unlike PCA, this graph prioritizes the preservation of local structures within the data instead of the global structure [38]. In the t-SNE space the trajectories formed clear clusters (Figure 6.6c). Most of the clusters are dominated by one of the decision categories (Figure 6.6c and Figure 6.16b). Interestingly, we found clusters in Animals 2, 3, and 4, in which the

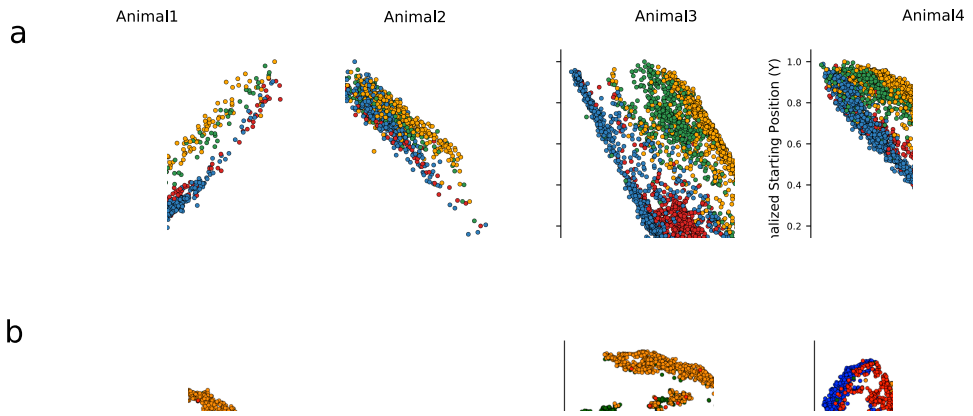

Figure 6.8: **Additional information on the unsupervised analysis of behavior trajectories**. (a) Scatter plot of starting positions along the left-right axis against PC2 shows correlation between the two. Starting positions are normalized to range from 0 (the leftmost position) and 1 (the rightmost position). (b) t-SNE plots with colors indicating different decision categories.

centroid trajectories were flat, unlike the trajectories of the four decision categories (Figure 6.6d), suggesting that animals hesitated in these trials and made decisions only after a delay. Indeed, in trials flagged by these clusters, the animals had longer reaction times (Figure 6.6e). Furthermore, such hesitating responses were more common following an incorrect trial (Figure 6.6f); they may reflect a behavioral adjustment to prevent further mistakes [6].

6.3 Discussion

Despite the fact that rodents can be trained to perform interesting decision-making tasks [10, 16, 22, 23], the learning progress of individual animals has rarely been addressed. Doing so requires training and observing many animals in parallel under identical conditions, and the ability to analyze the decision policy of each animal on a trial-by-trial basis. To meet these demands, we developed Mouse Academy, an integrated platform for automated training and behavior analysis of individual animals.

We demonstrate here that Mouse Academy can train group-housed mice in an automated and highly efficient manner while simultaneously acquiring decision-making sequences and video recordings. Automated animal training has been of

great interest in recent years and efforts have focused on two directions. In one design, multiple animals are trained in parallel within stacks of training boxes. This requires a technician to transfer animals from their home cages to the behavior boxes [9, 30, 40]. Such animal handling has been reported to introduce additional variability [12, 20], and even the mere presence of an experimenter can influence behavioral outcomes [34]. Thus, eliminating the requirement for human intervention, as in Mouse Academy, likely reduces experimental variation. In another design, a training setup is incorporated within the animals' home cage [15, 28, 33]. By contrast, Mouse Academy separates the functions of housing and training, and that modular design allows easy adaptation to a different purpose. For instance, one can replace the 3-port discrimination box with a maze to study spatial navigation learning [4, 26], or with an apparatus for training under voluntary head-fixation [2]. In each case, a single training apparatus can serve many mice, potentially from multiple home cages.

To understand how an animal's decision-making policies change in the course of learning, we developed a trial-by-trial iterative GLM. The evolution of the model is similar to online machine learning [32] in which the data are streamed in sequentially, rather than in batch mode. The linear nature of the model supports a straightforward definition of the animal's decision policy, namely as the vector of weights associated with different input variables. In addition, the simple linear structure allows rapid execution of the algorithm, which favors its use in real-time closed-loop behavior experiments. The model also allows several parametric adjustments. One specifies how much the recent trials are weighted over more distant ones in shaping the animal's policy. Another rates the relative influence of reward versus punishments. Fitting these parameters to each animal already revealed differences in learning style. This model can have a broader use beyond mouse decision-making, for instance to track the progress of human learners from their answers to a series of quizzes [27].

Finally we analyzed behavioral trajectories of individual animals. The behavioral trajectories can reveal intricate aspects of the animal's decision process that are hidden from a mere record of the binary choices. The large data volume again calls for automated analysis, and both supervised machine learning methods [14, 17, 19] and unsupervised classification [5, 17, 38, 41] have been employed for this purpose. Unsupervised analysis is not constrained by class labels, and can identify hidden structure in the data in an unbiased manner. In the present case, we discovered a motif wherein the animal hesitates on certain trials before taking action.

Mouse Academy can be combined with chronic wireless recording [37, 44], to allow

synchronized data acquisition of neural responses. Researchers can seek correlations between neural activity and the policy matrix or even the behavioral trajectories. This will open the door to a mechanistic understanding of how neural representations and dynamics change in the course of animal learning.

Online Methods

Animals

Subjects were C57BL/6J male mice aged 8–12 weeks. All experiments were conducted in accordance with protocols approved by the Institutional Animal Care and Use Committee of the California Institute of Technology.

Hardware setup

The hardware setup comprises a behavioral training box, an engineered home cage, and a radio frequency identification (RFID) sorting system, which allows animals to move between the home cage and the training box. These components are coordinated by customized software.

The design file for the behavior box was modified from that of Sanworks LLC (`https://github.com/sanworks/Bpod-CAD`) using Solid-Works computer-aided design software and the customized behavioral training box was manufactured in the lab. The behavior box is controlled by a Bpod state machine (r0.8, Sanworks LLC). To monitor the animal's behavior, an IR webcam (Ailipu Technology or OpenMV Camera M7) is installed above the behavior box. The behavior box and the webcam are placed within a light- and sound-proof chamber. The chamber is made of particle board with walls covered by acoustic foam. A tunnel made of red plastic tubes connects the behavior box to a home cage (Figure 6.9b).

For the RFID access control system, an Arduino Mega 2560 microcontroller is connected with three RFID readers (ID-12LA, Sparkfun) with custom antenna coils spaced along the access tunnel. The microcontroller controls two generic servo motors fitted with plastic gates to grant individual access to the training box (Figure 6.9a).

The microcontroller identifies each animal by its implanted RFID chip and permits only one animal to go through the tunnel connecting the home cage and the behavioral training box (Figure 6.9c). It also communicates the animal's identity to a master program running on a PC or Raspberry Pi (Matlab or Python). The master program

coordinates the following programs: Bpod (https://github.com/sanworks/Bpod), synchronized video recording, data management, and logging. A repository containing the design files, the firmware code for the microcontroller, and the software can be found in https://github.com/muqiao0626/Mouse_Academy.

Behavior training

The training procedures of mice to perform a selective attention task are similar to those previously reported [29, 42]. Mice were water restricted for seven days before training, and habituated in the automated training system to collect reward freely for several sessions. Then the mice were trained in sessions, each of which was made of 90 trials, to collect water rewards by performing two alternative forced choice tasks. Briefly, the animal had to nose-poke one of two choice ports based on the presented stimuli. If the decision was correct, 10% sucrose-sweetened water (3 µL) was delivered to the animal. For incorrect responses, the animal was punished with a five-second timeout. Following an incorrect response, the animal was presented with the identical trial again; this simple shaping procedure helps counter-act biases in the behavior. Over 28 days of training the animals learned increasingly complex tasks, from visual discrimination to a two-modality cued attention switching task [29, 42]. The training progressed through six stages (Figure 6.10):

1. A simple visual task: In this task, the animal initiates a trial by poking the center port and holding the position for 100 ms. Then either the left- or right-side port light up briefly until the animal moves away from the center port. The animal must then poke one of the two side ports within the decision period of 10 s. Choice of the port flagged by the light leads to a water reward, and choice of the other port leads to a timeout period during which no trials can be initiated. Data presented in the main text are from this stage of training only.

2. A simple auditory task: As Stage 1, except that the stimulus was white noise sound either the left or the right side to flag the reward port.

3. A cued single-modality (visual or auditory) switching task: Blocks of 15 trials consisting of single-modality (visual or auditory) stimulus presentation. Each block was like Stages 1 or 2, except that the trial type was indicated by a 7 kHz (visual) or 18 kHz (auditory) pure tone.

4. A cued single- and double-modality switching task: Like Stage 3, but distracting trials were introduced in which both visual and auditory stimuli were present,

but only one of the modalities was relevant to the decision. The relevant modality was again indicated by the pure tone cues. In repeating blocks, four types of trials were presented: a. five visual-only trials; b. ten "attend to vision" trials with auditory distractors; c. five auditory-only trials; d. ten "attend to audition" trials with visual distractors. During the training, the time that the animal had to hold in the center port was gradually increased to 0.5 s, and the duration of the stimuli was gradually shortened to 0.2 s.

5. A cued double-modality switching task: Like Stage 4 except that the single-modality trials were removed, and the block length was gradually shortened to three trials.

6. A selective attention task: Like Stage 5, but the block structure was abandoned and all eight possible trial types were randomized: (audition vs vision) × (sound left or right) × (light left or right).

Iterative generalized linear model

We modeled the animal's choice probability by a logistic regression. At each trial number t, the choice probability is defined as

$$p(y_t = 1|\mathbf{w}_{t-1}) = \frac{1}{1 + \exp -\mathbf{w}_{t-1}^T x_t} \tag{6.1}$$

$$p(y_t = -1|\mathbf{w}_{t-1}) = 1 - p(y_t = 1|\mathbf{w}_{t-1}) \tag{6.2}$$

where y_t indicates the binary choice of the animal (1 = right, –1 = left), $\mathbf{x}_t$ is the vector of input factors on trial t, and $\mathbf{w}_{t-1}$ is the vector of weights for these factors obtained from fitting up to the preceding trial. The prediction y_t^* for the animal's choice is simply that with the higher model probability:

$$y_t^* = \underset{y \in \{-1,+1\}}{\operatorname{argmax}}\ p(y|\mathbf{x}_t, \mathbf{w}_{t-1}) \tag{6.3}$$

After observing the animal's actual choice z_t, the cross-entropy error E_t between the observation and model prediction is calculated as

$$E_t = -\log p(z_t|\mathbf{x}_t, \mathbf{w}_{t-1}) \tag{6.4}$$

We weight the error term by a reward factor , and apply exponential temporal smoothing to get the loss function L_t:

$$L_t = R_t E_t + \alpha L_{t-1} \tag{6.5}$$

where is the smoothing discount factor accounting for the effect that distant trials have less impact on decision-making than immediately preceding trials, and R_t is defined as

$$R_t = \begin{cases} 1, & \text{if the choice is rewarded} \\ r, & \text{otherwise} \end{cases} \tag{6.6}$$

The values of R_t for rewarded and unrewarded trials may be different, accounting for the fact that rewards and punishments may have different effects on learning. For each time point, the weights in the model are determined by minimizing the loss function subject to L1 (lasso) regularization, namely

$$w_t = \underset{w}{\operatorname{argmin}}(L_t + \lambda \|\mathbf{w}\|_1) \tag{6.7}$$

Then $\mathbf{w}_t$ is used for prediction of the next trial. For subsequent analysis, we only used predictions starting at the 15th trial. The three hyperparameters for the temporal discount factor α, the reward factor r, and the regularization factor λ were selected by grid search.

To fit the decision-making sequences of the simple visual task, we included the following terms in the input vector $\mathbf{x}_t$:

1. Visual_stimulus: +1 = light on right, −1 = light on left.
2. Bias: A constant value of +1. The associated weight determines whether the animal favors the left (negative) or the right (positive) port.
3. Choice_back_n: The choice the animal made n trials ago (+1 = right, −1 = left).
4. Reward_back_n: The reward the animal received n trials ago (+1 = reward, −1 = punishment).

5. Choice × Reward_back_n: The product of terms 4 and 5. This term corresponds to the win-stay-lose-switch (WSLS) strategy of repeating the last choice if it was rewarded and switching if it was punished.

To determine the extent of history-dependence of the animal's decisions, we fitted the model including terms 3–5 from up to three previous trials ($n = 1, 2, 3$), and found that only the immediately preceding trial had an appreciable effect on the model's prediction accuracy. For the subsequent analysis, we therefore included terms 3–5 for the preceding trial ($n = 1$).

We compared the iterative generalized linear model (GLM) with two other models. The first only captures the animal's average performance over all trials. If the fraction of the correct responses is z, then the model simply predicts a correct response with probability z, and an error with probability $1 - z$. Thus, the fraction of trials where the prediction matches the observation is $z^2 + (1 - z)^2$.

The second model is a sliding window logistic regression. To make a prediction for trial t, we fitted the logistic model presented above (Equations 6.1–6.2) to the preceding n trials. The loss function is:

$$L_t = -\sum_{i=t-n}^{t-1} \log p(z_t|\mathbf{x}_t, \mathbf{w}_{t-1}) \tag{6.8}$$

and the weights are again optimized as in Equation 6.7.

Recovering policy matrices from simulated data To test the model's ability in recovering policy matrices, we trained the model on data generated from pre-defined ground truth policies. The ground truth policies changed every 10 trials, 30 trials, or 90 trials. Binary choices were simulated with different noise levels using the algorithm "ϵ-greedy": with a probability of ϵ, the simulator made a random choice, and with a probability of $1 - \epsilon$, it chose the action indicated by the ground truth policy. The noise levels (ϵ values) ranged from 0 to 0.6. The similarity between the recovered policy and the ground truth policy was evaluated by the cosine between the recovered weight vector and the ground truth weight vector.

Supervised and unsupervised analysis of behavioral trajectories We annotated two body landmarks, the nose and the centroid, on ~100 frames of the video, and used them to train DeepLabCut [24]. Tested on a separate set of annotated frames, more than 85% the nose positions are inferred within an error radius of 0.25 cm, and more

than 85% of the centroid positions are inferred within an error radius of 0.4 cm. From the nose and the centroid, we calculated the orientation as the angle of the line connecting the centroid and the nose. For each trial, the centroid and orientation were extracted for n frames (n = 30 (1 s) in most cases), thus the data dimension for each trial is 3n (the two centroid coordinates and the orientation).

To determine whether the behavioral trajectories contain information about the decision categories, a support vector machine (SVM) with a linear kernel was trained for each decision category. The training set was labelled with the decision category based on information about the visual stimulus and the animal's choice (for example, "Stim: R, Choice: L" means that the light is on the right and the animal chooses the left port). Performance of the trained SVM was examined by prediction accuracy on the test set, and the F1 score, which is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{true positive}}{2 \cdot \text{true positive} + \text{false positive} + \text{false positive}} \quad (6.9)$$

The performance was computed as the average across 10 repeated analyses (Figure 6.15).

We performed a non-linear embedding method, t-distributed stochastic neighbor embedding (t-SNE) analysis as previously described [5, 38]. Briefly, the trajectory data of each trial were projected into a 2D t-SNE space. Point clouds on the t-SNE map represented candidate clusters. Density clustering identified these regions. We then plotted trajectories and reaction time distributions to confirm that the clusters were distinct from each other. A repository of the analysis scripts can be found in https://github.com/tonyzhang25/MouseAcademyBehavior.

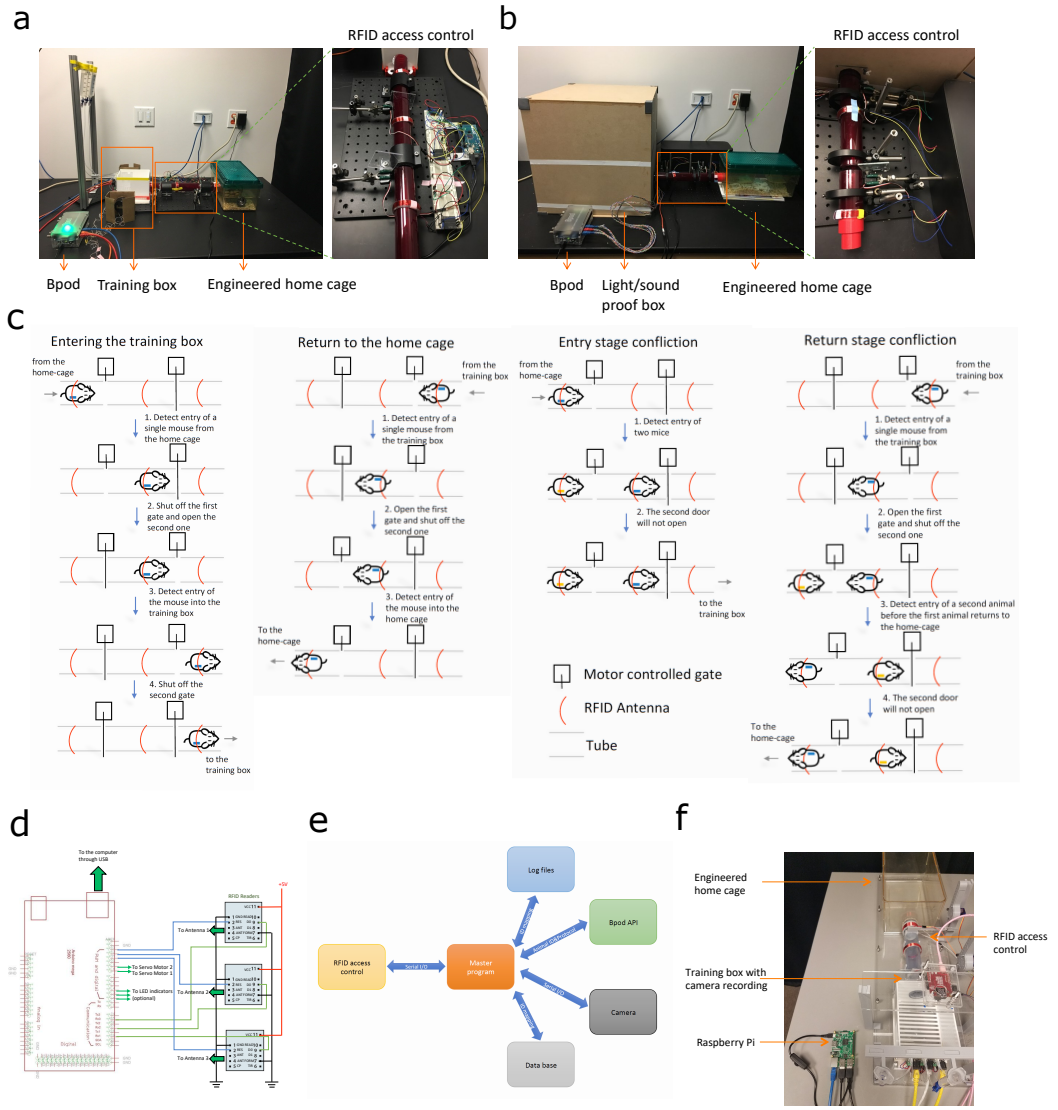

Figure 6.9: **Technical details of the hardware design**. (a-b) Side view of the setup (a) packed into a light- and sound-proof box (b). (c) RFID sorting process. For an animal to enter the behavior box, only when the left and the middle detectors detect the same RFID chip, the left gate is closed and the right gate is open so that the animal can access the behavior box. For an animal to return to the home cage, only when the right and the middle detectors detect the same RFID chip, the right gate is closed and the left gate is open so that the animal can go back to the home cage. In the entry and the return processes, if the left and the middle detectors detect different RFID chips, the animals have to leave the tube and the detectors get reset afterwards. (d) Schematic of RFID access control circuit. (e) Schematic of the software controlling the devices. A master program receives input from the RFID sorting device and controls four other modules including Bpod, synchronized video recording, data management, and logging. (f) Top view of a Raspberry Pi version of the setup.

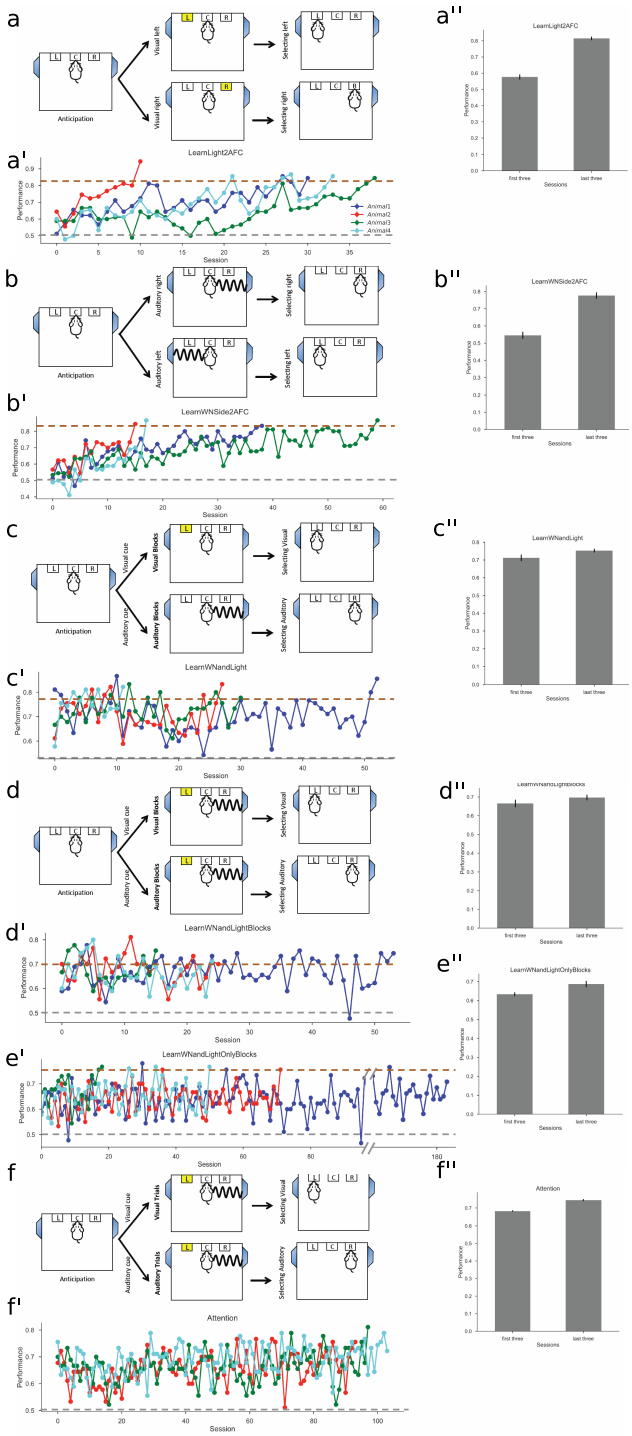

Figure 6.10: **Illustration of training procedures**. Training proceeds through six stages (**Online Methods**). The design, learning curves, and animal performance of the simple visual task (a, a', a''), the simple auditory task (b, b', b''), the cued single-modality (visual or auditory) switching task (c, c', c''), the cued single- (visual or auditory) and double-modality (attend to vision or audition) switching task (d, d', d''), the cued double-modality (attend to vision or audition) switching task (e', e''), and the final selective attention task (f, f', f'') are shown here. a' displays performance data as in Figure 3e. Brown and gray dashed lines indicate the performance thresholds for upgrading to the next stage and downgrading to the previous stage respectively.

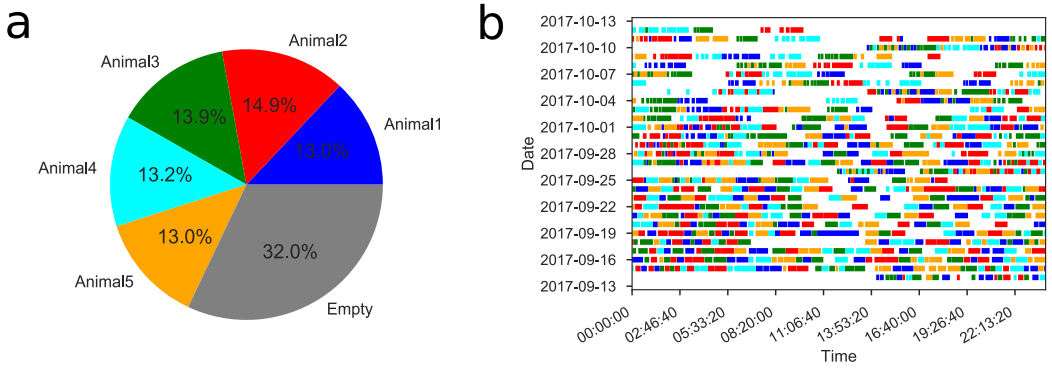

Figure 6.11: **Automated training system allows efficient use of the behavior box**. For a sample cohort of five animals, this shows the fraction of time each animal used the behavior box (a) and the activity trace of each animal throughout one month.

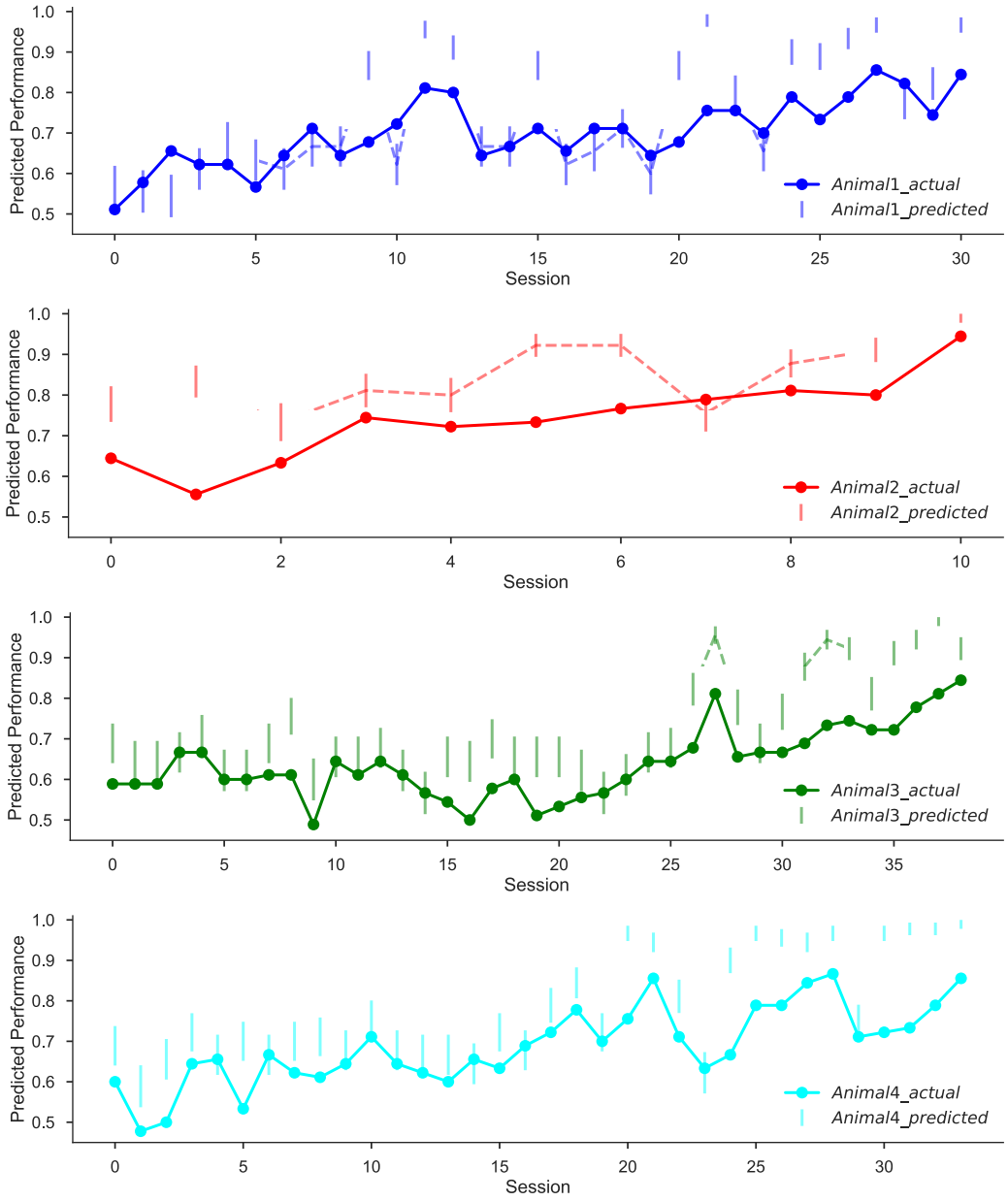

Figure 6.12: **Additional analysis on the iterative generalized linear model's prediction accuracy**. The actual performance and the performance predicted by the model, for each of the four animals. Note that the predictions recapitulate the more prominent fluctuations in the actual learning curves. Error bars indicate standard errors.

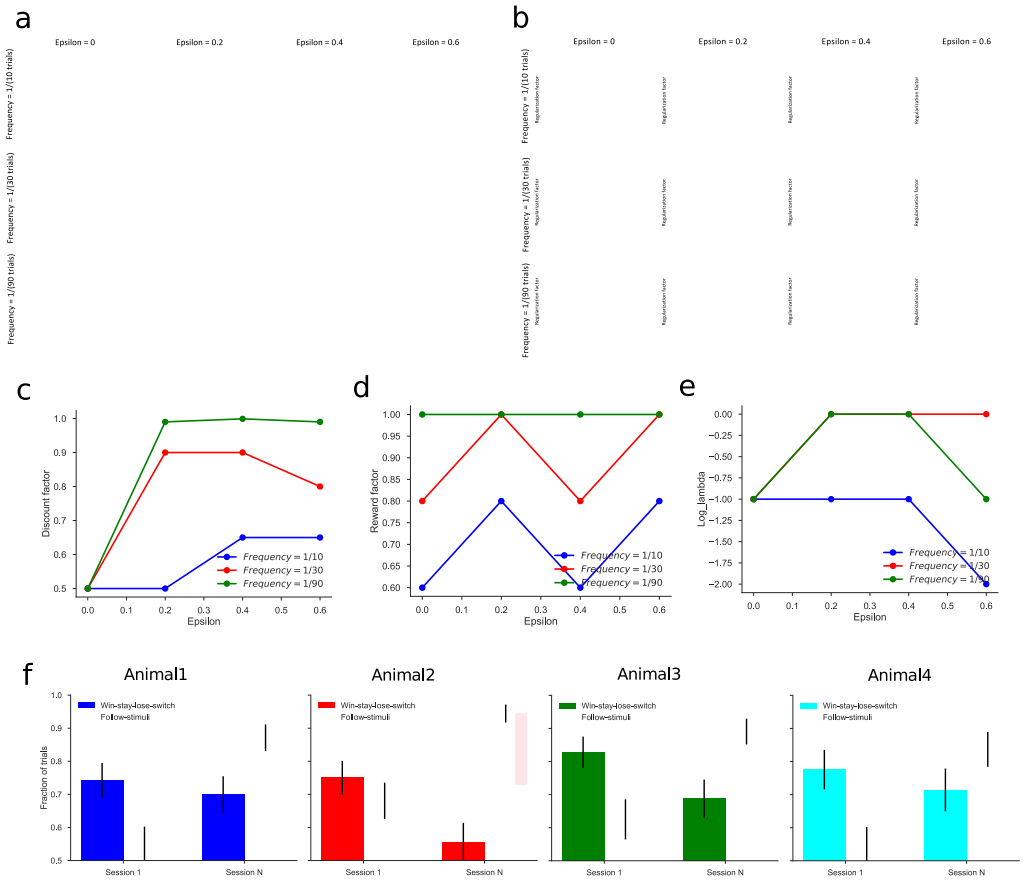

Figure 6.13: **Iterative generalized linear model captures differences between individuals and policy changes**. (a-b) Hyperparameter selection for the GLMs fitted to the simulated data generated from the ground truth policies. Different values of policy change frequency and noise level (ϵ) lead to different landscapes of the hyperparameters. (c-e) Selected temporal discount factor (c), reward factor (d) and regularization factor (e) for different values of policy change frequency and noise level (ϵ). (f) Fraction of the trials explained by the two policies (win-stay-lose-switch or WSLS, and following the stimuli) in the first and last sessions, for each of the four animals.

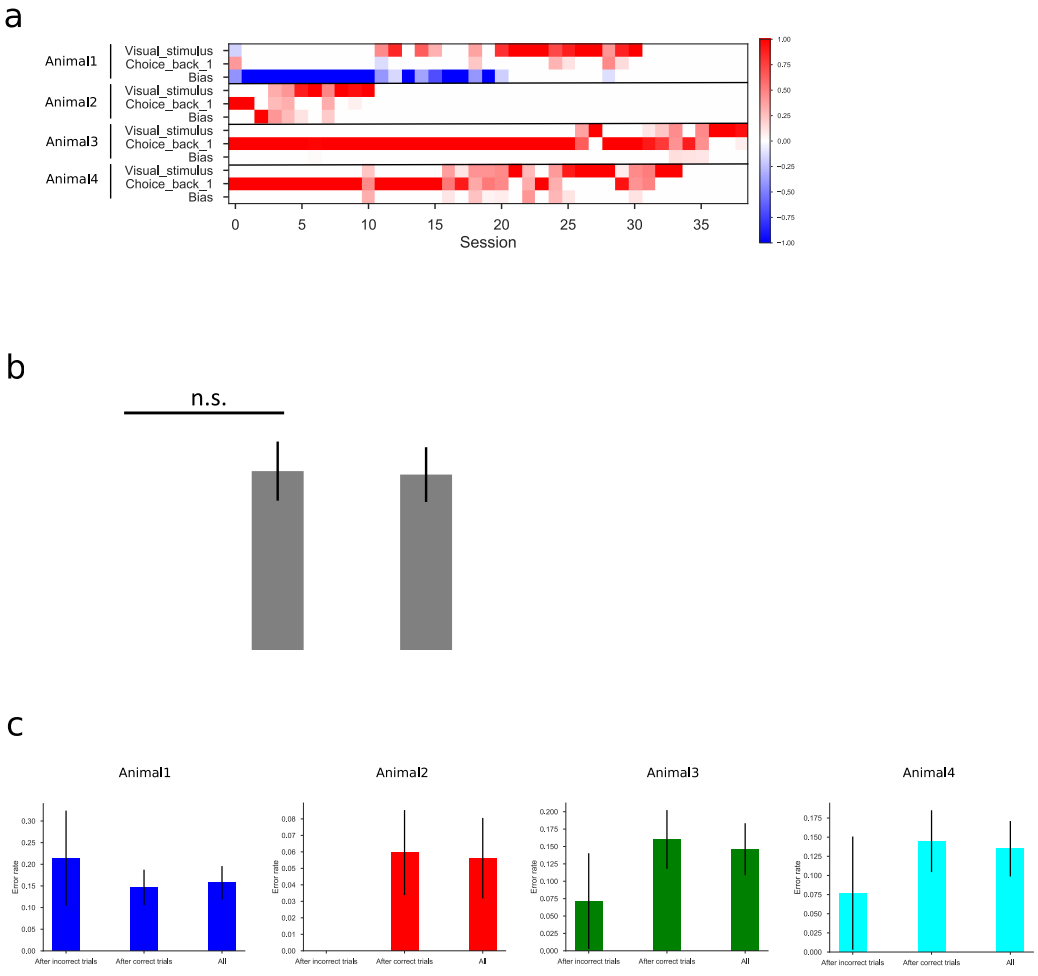

Figure 6.14: **Additional analyses of policy changes during learning**. (a) Policy matrices over sessions of the four animals. Here the policy matrices are recovered from logistic regression using only the trials following a correct response. Because the reward of the last trial is always +1, the term Reward_back_1 is the same as Bias, and the term RewardxChoice_back_1 is equal to Choice_back_1, so we drop them to avoid redundancy. (b-c) Quantification of the error rate during the last session, comparing trials following a correct response to those following a mistake. Averaged over all four animals (b) and for each of the four animals (c). n.s. indicates not significant.

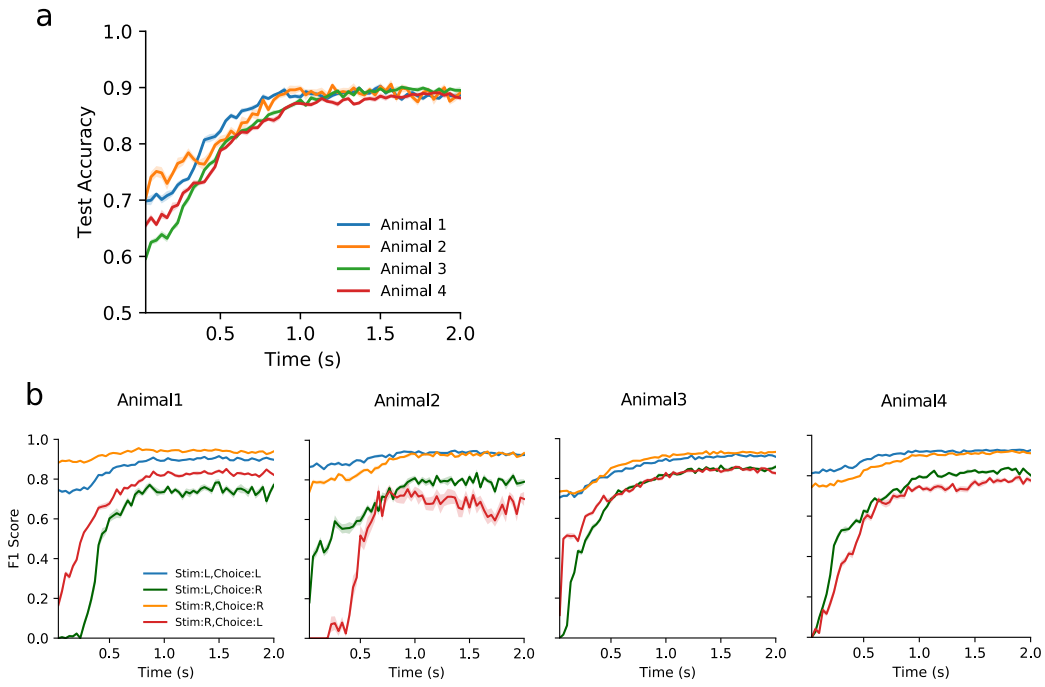

Figure 6.15: **Performance of the support vector machine to infer trial category from mouse trajectories**. (a) Prediction accuracy of the SVMs for individual animals. (b) F1 score of the SVM fitted for the decision categories of each animal. Shaded region denotes standard error. x-axis indicates the time starting from when the animal leaves the center port to make a choice. SVMs were trained using features up to a certain time point.

Figure 6.16: **Additional information on the unsupervised analysis of behavior trajectories**. (a) Scatter plot of starting positions along the left-right axis against PC2 shows correlation between the two. Starting positions are normalized to range from 0 (the leftmost position) and 1 (the rightmost position). (b) t-SNE plots with colors indicating different decision categories.

References

[1] Athena Akrami, Charles D Kopec, Mathew E Diamond, and Carlos D Brody. "Posterior parietal cortex represents sensory history and mediates its effects on behaviour". In: *Nature* 554.7692 (2018), pp. 368–372.

[2] Ryo Aoki, Tadashi Tsubota, Yuki Goya, and Andrea Benucci. "An automated platform for high-throughput mouse behavior and physiology with voluntary head-fixation". In: *Nature communications* 8.1 (2017), pp. 1–9.

[3] John W Atkinson and Alfred C Raphelson. "Individual differences in motivation and behavior in particular situations." In: *Journal of Personality* (1956).

[4] Carol A Barnes. "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.

[5] Gordon J Berman, Daniel M Choi, William Bialek, and Joshua W Shaevitz. "Mapping the structure of drosophilid behavior". In: *bioRxiv* (2014), p. 002873.

[6] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. "Conflict monitoring and cognitive control." In: *Psychological review* 108.3 (2001), p. 624.

[7] Kristin Branson, Alice A Robie, John Bender, Pietro Perona, and Michael H Dickinson. "High-throughput ethomics in large groups of Drosophila". In: *Nature methods* 6.6 (2009), pp. 451–457.

[8] Christopher P. Burgess, Armin Lak, Nicholas A. Steinmetz, Peter Zatka-Haas, Charu Bai Reddy, Elina A. K. Jacobs, Jennifer F. Linden, Joseph J. Paton, Adam Ranson, Sylvia Schröder, Sofia Soares, Miles J. Wells, Lauren E. Wool, Kenneth D. Harris, and Matteo Carandini. "High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice". In: *Cell Reports* 20.10 (Sept. 2017), pp. 2513–2524. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.08.047.

[9] Laura Busse, Asli Ayaz, Neel T. Dhruv, Steffen Katzner, Aman B. Saleem, Marieke L. Schoelvinck, Andrew D. Zaharia, and Matteo Carandini. "The Detection of Visual Contrast in the Behaving Mouse". English. In: *Journal of Neuroscience* 31.31 (Aug. 2011), pp. 11351–11361. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.6689-10.2011.

[10] M. Carandini and Anne K Churchland. "Probing Perceptual Decisions in Rodents." In: *Nature Neuroscience* 16 (July 2013), pp. 824–31. DOI: 10.1038/nn.3410.

[11] Matteo Carandini and Anne K Churchland. "Probing perceptual decisions in rodents". In: *Nature neuroscience* 16.7 (2013), pp. 824–831.

[12] John C Crabbe, Douglas Wahlsten, and Bruce C Dudek. “Genetics of mouse behavior: interactions with laboratory environment”. In: *Science* 284.5420 (1999), pp. 1670–1672.

[13] Serge Daan, Kamiel Spoelstra, Urs Albrecht, Isabelle Schmutz, Moritz Daan, Berte Daan, Froukje Rienks, Inga Poletaeva, Giacomo Dell’Omo, Alexei Vyssotski, et al. “Lab mice in the field: unorthodox daily activity and effects of a dysfunctional circadian clock allele”. In: *Journal of biological rhythms* 26.2 (2011), pp. 118–129.

[14] SE Roian Egnor and Kristin Branson. “Computational analysis of behavior”. In: *Annual review of neuroscience* 39 (2016), pp. 217–236.

[15] Andrew Erskine, Thorsten Bus, Jan T Herb, and Andreas T Schaefer. “AutonoMouse: High throughput automated operant conditioning shows progressive behavioural impairment with graded olfactory bulb lesions”. In: *bioRxiv* (2018), p. 291815.

[16] Alex Gomez-Marin, Joseph J Paton, Adam R Kampff, Rui M Costa, and Zachary F Mainen. “Big behavioral data: psychology, ethology and the foundations of neuroscience”. In: *Nature neuroscience* 17.11 (2014), pp. 1455–1462.

[17] Katsiaryna V Gris, Jean-Philippe Coutu, and Denis Gris. “Supervised and unsupervised learning technology in the study of rodent behavior”. In: *Frontiers in behavioral neuroscience* 11 (2017), p. 141.

[18] Kyle Honegger and Benjamin de Bivort. “Stochasticity, individuality and behavior”. In: *Current Biology* 28.1 (2018), R8–R12.

[19] Weizhe Hong, Ann Kennedy, Xavier P Burgos-Artizzu, Moriel Zelikowsky, Santiago G Navonne, Pietro Perona, and David J Anderson. “Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning”. In: *Proceedings of the National Academy of Sciences* 112.38 (2015), E5351–E5360.

[20] Jane L Hurst and Rebecca S West. “Taming anxiety in laboratory mice”. In: *Nature methods* 7.10 (2010), pp. 825–826.

[21] Roelof A Hut, Violetta Pilorz, Ate S Boerema, Arjen M Strijkstra, and Serge Daan. “Working for food shifts nocturnal mouse activity into the day”. In: *PloS one* 6.3 (2011), e17527.

[22] Ashley L Juavinett, Jeffrey C Erlich, and Anne K Churchland. “Decision-making behaviors: weighing ethology, complexity, and sensorimotor compatibility”. In: *Current opinion in neurobiology* 49 (2018), pp. 42–50.

[23] Liqun Luo, Edward M Callaway, and Karel Svoboda. “Genetic dissection of neural circuits: a decade of progress”. In: *Neuron* 98.2 (2018), pp. 256–281.

[24] Alexander Mathis, 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". en. In: *Nature Neuroscience* (Aug. 2018), p. 1. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0209-y.

[25] Louis D Matzel and Bruno Sauce. "Individual differences: Case studies of rodent and primate intelligence." In: *Journal of Experimental Psychology: Animal Learning and Cognition* 43.4 (2017), p. 325.

[26] David S Olton, John A Walker, and Fred H Gage. "Hippocampal connections and spatial discrimination". In: *Brain research* 139.2 (1978), pp. 295–308.

[27] Chris Piech, Jonathan Spencer, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. "Deep knowledge tracing". In: *arXiv preprint arXiv:1506.05908* (2015).

[28] Rajesh Poddar, Risa Kawai, and Bence P Ölveczky. "A fully automated high-throughput training system for rodents". In: *PloS one* 8.12 (2013), e83171.

[29] L Ian Schmitt, Ralf D Wimmer, Miho Nakajima, Michael Happ, Sima Mofakham, and Michael M Halassa. "Thalamic amplification of cortical connectivity sustains attentional control". In: *Nature* 545.7653 (2017), pp. 219–223.

[30] Benjamin B Scott, Carlos D Brody, and David W Tank. "Cellular resolution functional imaging in behaving rats using voluntary head restraint". In: *Neuron* 80.2 (2013), pp. 371–384.

[31] Eyal Jacob Seidemann. *Neuronal mechanisms mediating conversion of visual signals into perceptual decisions in a direction discrimination task*. Stanford University, 1998.

[32] Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and trends in Machine Learning* 4.2 (2011), pp. 107–194.

[33] Gergely Silasi, Jamie D Boyd, Federico Bolanos, Jeff M LeDue, Stephen H Scott, and Timothy H Murphy. "Individualized tracking of self-directed motor learning in group-housed mice performing a skilled lever positioning task in the home cage". In: *Journal of neurophysiology* 119.1 (2018), pp. 337–346.

[34] Robert E Sorge, Loren J Martin, Kelsey A Isbester, Susana G Sotocinal, Sarah Rosen, Alexander H Tuttle, Jeffrey S Wieskopf, Erinn L Acland, Anastassia Dokova, Basil Kadoura, et al. "Olfactory exposure to males, including men, causes stress and related analgesia in rodents". In: *Nature methods* 11.6 (2014), pp. 629–632.

[35] Norman E Spear, James S Miller, and Joyce A Jagielo. "Animal memory and learning". In: *Annual review of psychology* 41.1 (1990), pp. 169–211.

[36] Elsbeth Stern. "Individual differences in the learning potential of human beings". In: *npj Science of Learning* 2.1 (2017), pp. 1–7.

[37] Tobi A Szuts, Vitaliy Fadeyev, Sergei Kachiguine, Alexander Sher, Matthew V Grivich, Margarida Agrochão, Pawel Hottowy, Wladyslaw Dabrowski, Evgueniy V Lubenov, Athanassios G Siapas, et al. "A wireless multi-channel neural amplifier for freely moving animals". In: *Nature neuroscience* 14.2 (2011), pp. 263–269.

[38] Jeremy G Todd, Jamey S Kain, and Benjamin L de Bivort. "Systematic exploration of unsupervised methods for mapping behavior". In: *Physical biology* 14.1 (2017), p. 015002.

[39] Naoshige Uchida and Zachary F Mainen. "Speed and accuracy of olfactory discrimination in the rat". In: *Nature neuroscience* 6.11 (2003), pp. 1224–1229.

[40] Naoshige Uchida and Zachary F. Mainen. "Speed and Accuracy of Olfactory Discrimination in the Rat". eng. In: *Nature Neuroscience* 6.11 (Nov. 2003), pp. 1224–1229. ISSN: 1097-6256. DOI: 10.1038/nn1142.

[41] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. "Mapping sub-second structure in mouse behavior". In: *Neuron* 88.6 (2015), pp. 1121–1135.

[42] R. D. Wimmer, L. I. Schmitt, T. J. Davidson, M. Nakajima, K. Deisseroth, and M. M. Halassa. "Thalamic Control of Sensory Selection in Divided Attention." In: *Nature* 526 (Oct. 2015), pp. 705–9. DOI: 10.1038/nature15398.

[43] York Winter and Andrea TU Schaefers. "A sorting system with automated gates permits individual operant experiments with mice from a social home cage". In: *Journal of neuroscience methods* 196.2 (2011), pp. 276–280.

[44] Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. "Long-term dynamics of CA1 hippocampal place codes". In: *Nature neuroscience* 16.3 (2013), pp. 264–266.

Part III

Learning and Behavior

Chapter 7

MICE IN A LABYRINTH EXHIBIT RAPID LEARNING, SUDDEN INSIGHT, AND EFFICIENT EXPLORATION

[1] Matthew Rosenberg*, Tony Zhang*, Pietro Perona, and Markus Meister. "Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration". In: *Elife* 10 (2021), e66175. URL: https://elifesciences.org/articles/66175.

[1] Matthew Rosenberg*, Tony Zhang*, Pietro Perona, and Markus Meister. "Rapid learning and efficient exploration by mice navigating a complex maze". In: *NeurIPS Biological and Artificial Reinforcement Learning Workshop* (2019). URL: https://sites.google.com/view/biologicalandartificialrl/home?authuser=0.

Animals learn certain complex tasks remarkably fast, sometimes after a single experience. What behavioral algorithms support this efficiency? Many contemporary studies based on two-alternative-forced-choice (2-AFC) tasks observe only slow or incomplete learning. As an alternative, we study the unconstrained behavior of mice in a complex labyrinth and measure the dynamics of learning and the behaviors that enable it. A mouse in the labyrinth makes ~2000 navigation decisions per hour. The animal explores the maze, quickly discovers the location of a reward, and executes correct 10-bit choices after only 10 reward experiences—a learning rate 1000-fold higher than in 2-AFC experiments. Many mice improve discontinuously from one minute to the next, suggesting moments of sudden insight about the structure of the labyrinth. The underlying search algorithm does not require a global memory of places visited and is largely explained by purely local turning rules.

7.1 Introduction

How can animals or machines acquire the ability for complex behaviors from one or a few experiences? Canonical examples include language learning in children, where new words are learned after just a few instances of their use, or learning to balance a bicycle, where humans progress from complete incompetence to near perfection after crashing once or a few times. Clearly such rapid acquisition of new associations or of new motor skills can confer enormous survival advantages.

In laboratory studies, one prominent instance of one-shot learning is the Bruce effect [7]. Here the female mouse forms an olfactory memory of her mating partner that allows her to terminate the pregnancy if she encounters another male that threatens infanticide. Another form of rapid learning accessible to laboratory experiments is fear conditioning, where a formerly innocuous stimulus gets associated with a painful experience, leading to subsequent avoidance of the stimulus [5, 13]. These learning systems appear designed for special purposes, they perform very specific associations, and govern binary behavioral decisions. They are likely implemented by specialized brain circuits, and indeed great progress has been made in localizing these operations to the accessory olfactory bulb [6] and the cortical amygdala [20].

In the attempt to identify more generalizable mechanisms of learning and decision making, one route has been to train laboratory animals on abstract tasks with tightly specified sensory inputs that are linked to motor outputs via arbitrary contingency rules. Canonical examples are a monkey reporting motion in a visual stimulus by saccading its eyes [26], and a mouse in a box classifying stimuli by moving its forelimbs or the tongue [9, 17]. The tasks are of low complexity, typically a 1-bit decision based on 1 or 2 bits of input. Remarkably they are learned exceedingly slowly: a mouse typically requires many weeks of shaping and thousands of trials to reach asymptotic performance; a monkey may require many months [10].

What is needed therefore is a rodent behavior that involves complex decision making, with many input variables and many possible choices. Ideally the animals would learn to perform this task without excessive intervention by human shaping, so we may be confident that they employ innate brain mechanisms rather than circuits created by the training. Obviously the behavior should be easy to measure in the laboratory. Finally, it would be satisfying if this behavior showed a glimpse of rapid learning.

Navigation through space is a complex behavior displayed by many animals. It typically involves integrating multiple cues to decide among many possible actions.

It relies intimately on rapid learning. For example a pigeon or desert ant leaving its shelter acquires the information needed for the homing path in a single episode. Major questions remain about how the brain stores this information and converts it to a policy for decisions during the homing path. One way to formalize the act of decision-making in the laboratory is to introduce structure in the environment in the form of a maze that defines straight paths and decision points. A maze of tunnels is in fact a natural environment for a burrowing rodent. Early studies of rodent behavior did place the animals into true labyrinths [35], but their use gradually declined in favor of linear tracks or boxes with a single choice point.

We report here on the behavior of laboratory mice in a complex labyrinth of tunnels. A single mouse is placed in a home cage from which it has free access to the maze for one night. No handling, shaping, or training by the investigators is involved. By continuous video-recording and automated tracking we observe the animal's entire life experience within the labyrinth. Some of the mice are water-deprived, and a single location deep inside the maze offers water. We find that these animals learn to navigate to the water port after just a few reward experiences. In many cases one can identify unique moments of "insight" when the animal's behavior changes discontinuously. This all happens within ~1 hour. Underlying the rapid learning is an efficient mode of exploration driven by simple navigation rules. Mice that do not lack water show the same patterns of exploration. This laboratory-based navigation behavior may form a suitable substrate for studying the neural mechanisms that implement few-shot learning.

7.2 Results

Adaptation to the maze

At the start of the experiment a single mouse was placed in a conventional mouse cage with bedding and food. A short tunnel offered free access to a maze consisting of a warren of corridors (Figure 7.1A-B). The bottom and walls of the maze were constructed of black plastic that is transparent in the infrared. A video camera placed below the maze captured the animal's actions continuously using infrared illumination (Figure 7.1B). The recordings were analyzed offline to track the movements of the mouse, with keypoints on the nose, mid-body, tail base, and the four feet (Figure 7.1D). All observations were made in darkness during the animal's subjective night.

The logical structure of the maze is a binary tree, with 6 levels of branches, leading from the single entrance to 64 endpoints (Figure 7.1C). A total of 63 T-junctions are

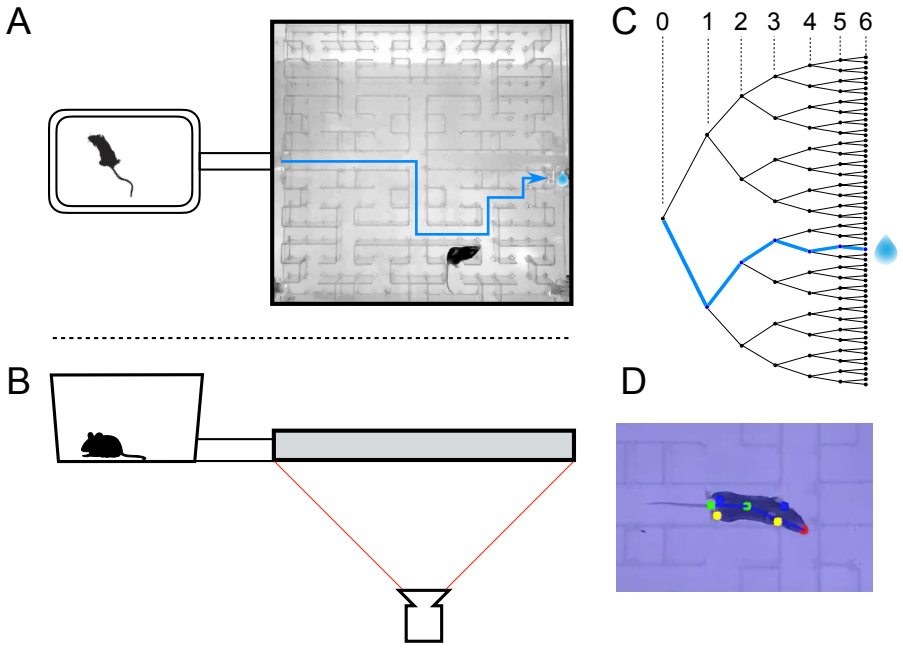

Figure 7.1: **The maze environment.** Top (A) and side (B) views of a home cage, connected via an entry tunnel to an enclosed labyrinth. The animal's actions in the maze are recorded via video from below using infrared illumination. (C) The maze is structured as a binary tree with 63 branch points (in levels numbered 0,...,5) and 64 end nodes. One end node has a water port that dispenses a drop when it gets poked. Blue line in A and C: path from maze entry to water port. (D) A mouse considering the options at the maze's central intersection. Colored keypoints are tracked by DeepLabCut: nose, mid body, tail base, 4 feet.

connected by straight corridors in a design with maximal symmetry (Figure 7.1A, Figure 7.6–Figure 7.7), such that all the nodes at a given level of the tree have the same local geometry. One of the 64 endpoints of the maze is outfitted with a water port. After activation by a brief nose poke, the port delivers a small drop of water, followed by a 90-s time-out period.

After an initial period of exploratory experiments, we settled on a frozen protocol that was applied to 20 animals. Ten of these mice had been mildly water-deprived for up to 24 hours; they received food in the home cage and water only from the port hidden in the maze. Another ten mice had free access to food and water in the cage, and received no water from the port in the maze. Each animal's behavior in the maze was recorded continuously for 7 h during the first night of its experience with the

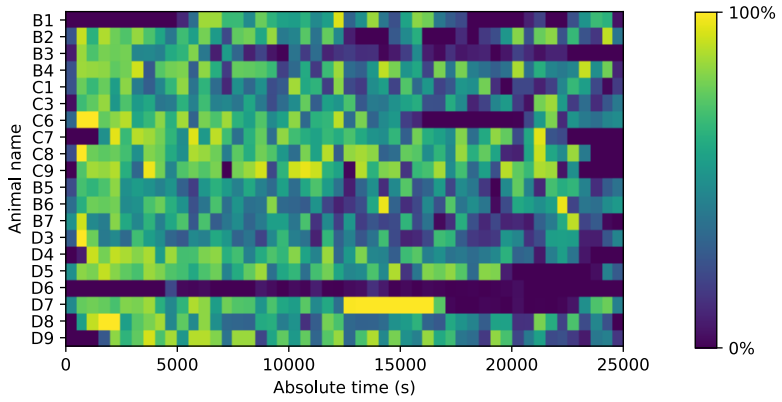

Figure 7.2: **Fraction of time spent in the maze.** Mice could move freely between the home cage and the maze. For each animal (vertical), the fraction of time in the maze (color scale) is plotted as a function of time since start of the experiment. Time bins are 500 s. Note that mouse D6 hardly entered the maze; it never progressed beyond the first junction. This animal was excluded from all subsequent analysis steps.

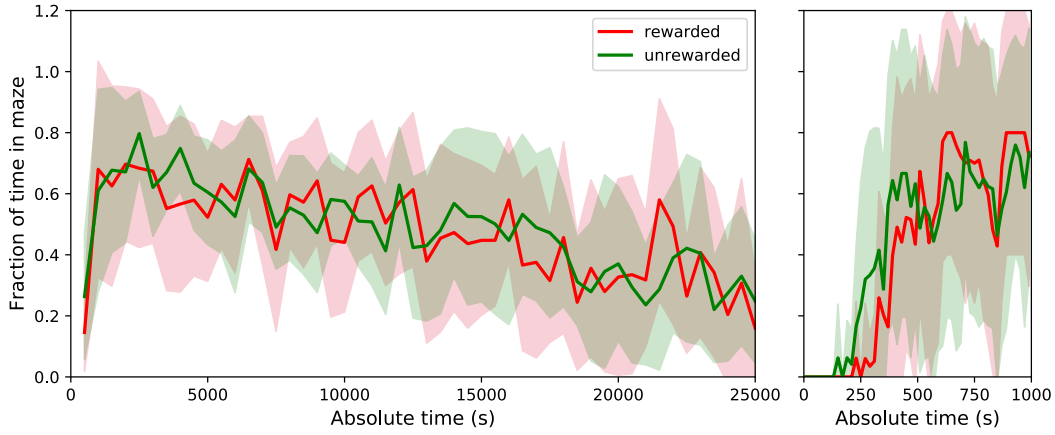

Figure 7.3: **Average fraction of time spent in the maze by group.** This shows the average fraction of time in the maze as Mean ± SD over the population of 10 rewarded and 9 unrewarded animals. Right: expanded axis for early times. The tunnel to the maze opens at time 0. Rewarded and unrewarded animals used the maze in remarkably similar ways. Exploration of the maze began around 250 s after tunnel opening. Within the next 250 s the maze occupancy rose quickly to ~70%, then declined gradually over 7 h to ~30%.

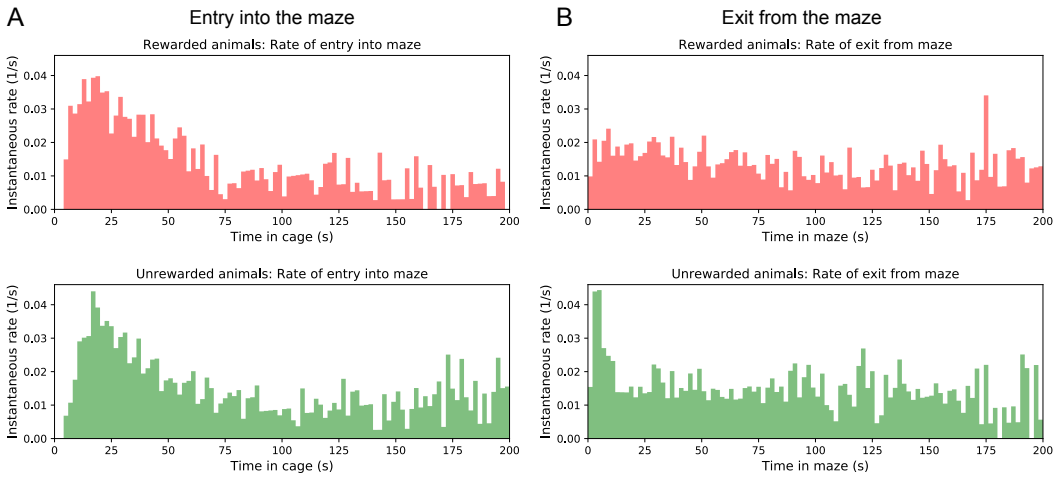

Figure 7.4: **Rates of transition between cage and maze.** (A) The instantaneous probability per unit time $r_{\mathrm{m}}(t)$ of entering the maze after having spent time t in the cage. Note this rate is highest immediately upon entering the cage, then declines by a large factor. (B) The instantaneous probability per unit time $r_{\mathrm{c}}(t)$ of exiting the maze after having spent time t in the maze.

maze, starting the moment the connection tunnel was opened (sample videos here). The investigator played no role during this period, and the animal was free to act as it wished including travel between the cage and the maze.

All of the mice except one passed between the cage and the maze readily and frequently (Figure 7.1–Figure 7.2). The single outlier animal barely entered the maze and never progressed past the first junction; we excluded this mouse's data from subsequent analysis. On average over the entire period of study the animals spent 46% of the time in the maze (Figure 7.1–Figure 7.3). This fraction was similar whether or not the animal was motivated by water rewards (47% for rewarded vs 44% for unrewarded animals). Over time the animals appeared increasingly comfortable in the maze, taking breaks for grooming and the occasional nap. When the investigator lifted the cage lid at the end of the night, some animals were seen to escape into the safety of the maze.

We examined the rate of transitions from the cage to the maze and how it depends on time spent in the cage (Figure 7.1–Figure 7.4A). Surprisingly the rate of entry into the maze is highest immediately after the animal returns to the cage. Then it declines gradually by a factor of 4 over the first minute in the cage and remains steady thereafter. This is a large effect, observed for every individual animal in both the rewarded and unrewarded groups. By contrast the opposite transition, namely exit from the maze, occurs at an essentially constant rate throughout the visit (Figure

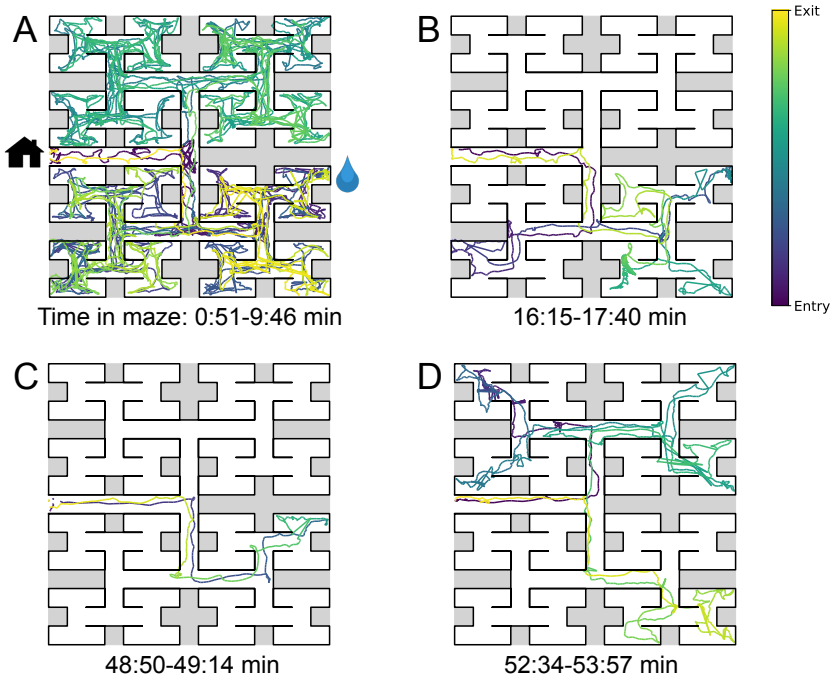

Figure 7.5: **Sample trajectories during adaptation to the maze.** Four sample bouts from one mouse (B3) into the maze at various times during the experiment (time markings at bottom). The trajectory of the animal's nose is shown; time is encoded by the color of the trace. The entrance from the home cage and the water port are indicated in Panel A.

7.1–Figure 7.4B).

The nature of the animal's forays into the maze changed over time. We call each foray from entrance to exit a "bout." After a few hesitant entries into the main corridor, the mouse engaged in one or more long bouts that dove deep into the binary tree to most or all of the leaf nodes (Figure 7.5A). For a water-deprived animal, this typically led to discovery of the reward port. After ~10 bouts, the trajectories became more focused, involving travel to the reward port and some additional exploration (Figure 7.5B). At a later stage still, the animal often executed perfect exploitation bouts that led straight to the reward port and back with no wrong turns (Figure 7.5C). Even at this late stage, however, the animal continued to explore other parts of the maze (Figure 7.5D). Similarly the unrewarded animals explored the maze throughout the night (Figure 7.1–Figure 7.3). While the length and structure of the animal's trajectories changed over time, the speed remained remarkably constant after ~50 s of adaptation (Figure 7.5–Figure 7.10).

Whereas Figure 7.5 illustrates the trajectory of a mouse's nose in full spatio-temporal

detail, a convenient reduced representation is the "node sequence." This simply marks the events when the animal enters each of the 127 nodes of the binary tree that describes the maze (see Methods and Figure 7.6–Figure 7.7). Among these nodes, 63 are T-junctions where the animal has 3 choices for the next node, and 64 are end nodes where the animal's only choice is to reverse course. We call the transition from one node to the next a "step." The analysis in the rest of the paper was carried out on the animal's node sequence.

Few-shot learning of a reward location

We now examine early changes in the animal's behavior when it rapidly acquires and remembers information needed for navigation. First we focus on navigation to the water port.

The ten water-deprived animals had no indication that water would be found in the maze. Yet, all 10 discovered the water port in less than 2000 s and fewer than 17 bouts (Figure 7.6A). The port dispensed only a drop of water followed by a 90-s timeout before rearming. During the timeout the animals generally left the port location to explore other parts of the maze or return home, even though they were not obliged to do so. For each of the water-deprived animals, the frequency at which it consumed rewards in the maze increased rapidly as it learned how to find the water port, then settled after a few reward experiences (Figure 7.6A).

How many reward experiences are sufficient to teach the animal reliable navigation to the water port? To establish a learning curve one wants to compare performance on the identical task over successive trials. Recall that this experiment has no imposed trial structure. Yet the animals naturally segmented their behavior through discrete visits to the maze. Thus we focused on all the instances when the animal started at the maze entrance and walked to the water port (Figure 7.6B).

On the first few occasions these paths to water can involve hundreds of steps between nodes and their length scatters over a wide range. However, after a few rewards, the animals began taking the perfect path without detours (6 steps, Figure 7.6–Figure 7.7), and soon that became the norm. Note the path length plotted here is directly related to the number of "turning errors": every time the mouse turns away from the shortest path to the water port, that adds two steps to the path length (Equation 7.7). The rate of these errors declined over time, by a factor of e after ~10 rewards consumed (Figure 7.6B). Late in the night ~75% of the paths to water were perfect. The animals executed them with increasing speed; eventually these fast "water runs"

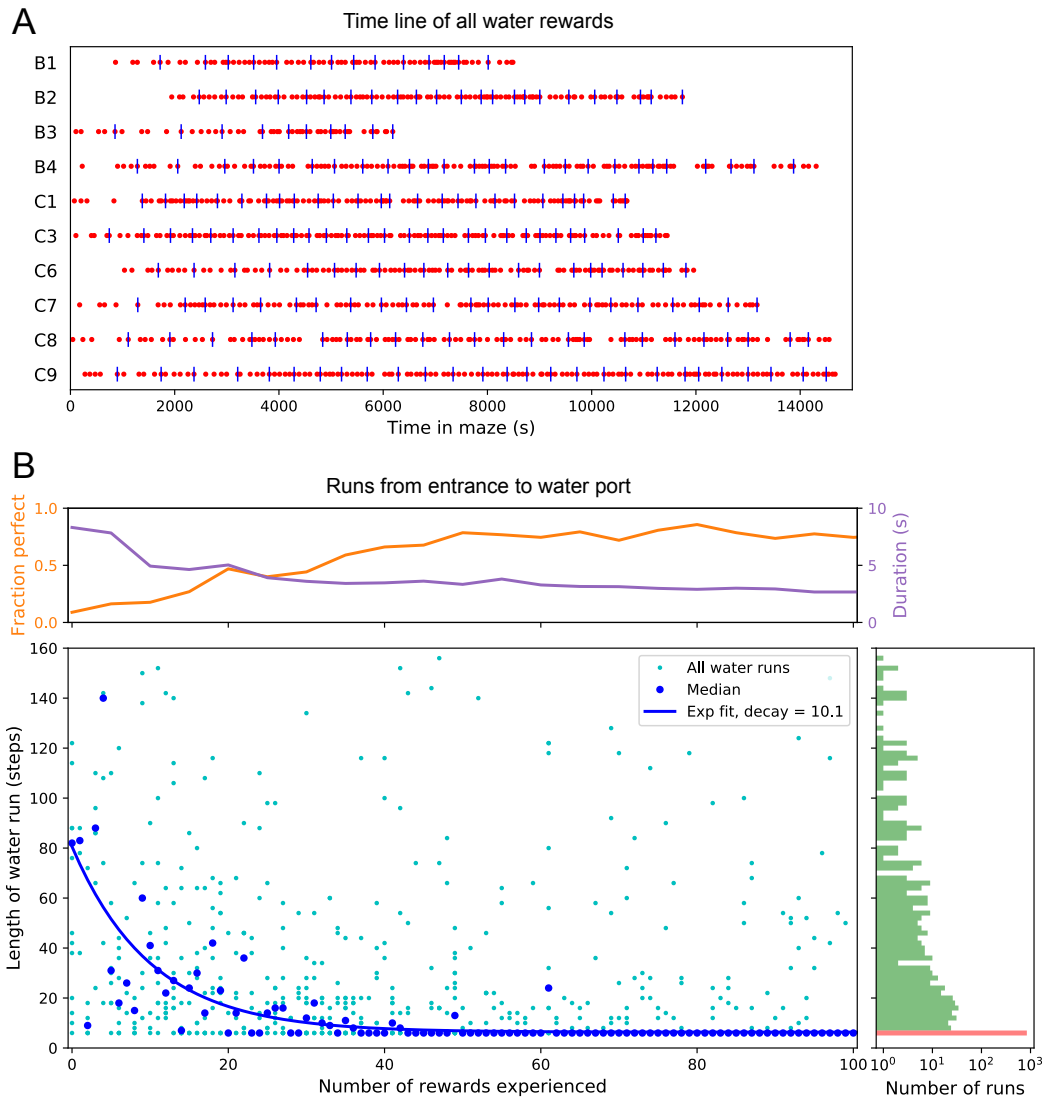

Figure 7.6: **Few-shot learning of path to water.** (A) Time line of all water rewards collected by 10 water-deprived mice (red dots, every fifth reward has a blue tick mark). (B) The length of runs from the entrance to the water port, measured in steps between nodes, and plotted against the number of rewards experienced. Main panel: All individual runs (cyan dots) and median over 10 mice (blue circles). Exponential fit decays by $1/e$ over 10.1 rewards. Right panel: Histogram of the run length, note log axis. Red: perfect runs with the minimum length 6; green: longer runs. Top panel: The fraction of perfect runs (length 6) plotted against the number of rewards experienced, along with the median duration of those perfect runs.

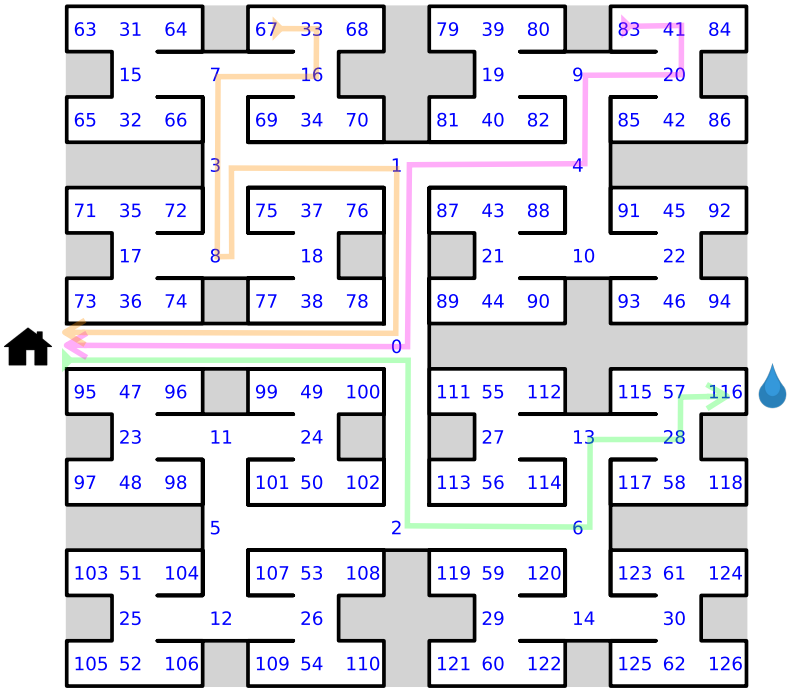

Figure 7.7: **Definition of node trajectories.** A numbering scheme for all 127 nodes of the maze. Green: a direct path from the entrance to the water port ("water run") with the node sequence $(s_i) = (0, 2, 6, 13, 28, 57, 116)$, involving 6 decisions. Magenta: a direct path from end node 83 to the exit ("home run"). Orange: a path from end node 67 to the exit that includes a reversal. Here the home run starts only from node 8, namely $(8, 3, 1, 0)$.

took as little as 2 s (Figure 7.6B). Many of these visits went unrewarded owing to the 90-s timeout period on the water port.

In summary, after ~10 reward experiences on average the mice learn to navigate efficiently to the water port, which requires making 6 correct decisions, each among 3 options. Note that even at late times, long after they have perfected the "water run," the animals continue to take some extremely long paths: a subject for a later section (Figure 7.15).

The role of cues attached to the maze

These observations of rapid learning raise the question, "How do the animals navigate?" In particular, does the mouse build an internal representation that guides its action at every junction? Or does it place marks in the external environment that signal the route to the water port? In an extreme version of externalized cognition, the mouse leaves behind a trail of urine marks or other secretions as it walks away

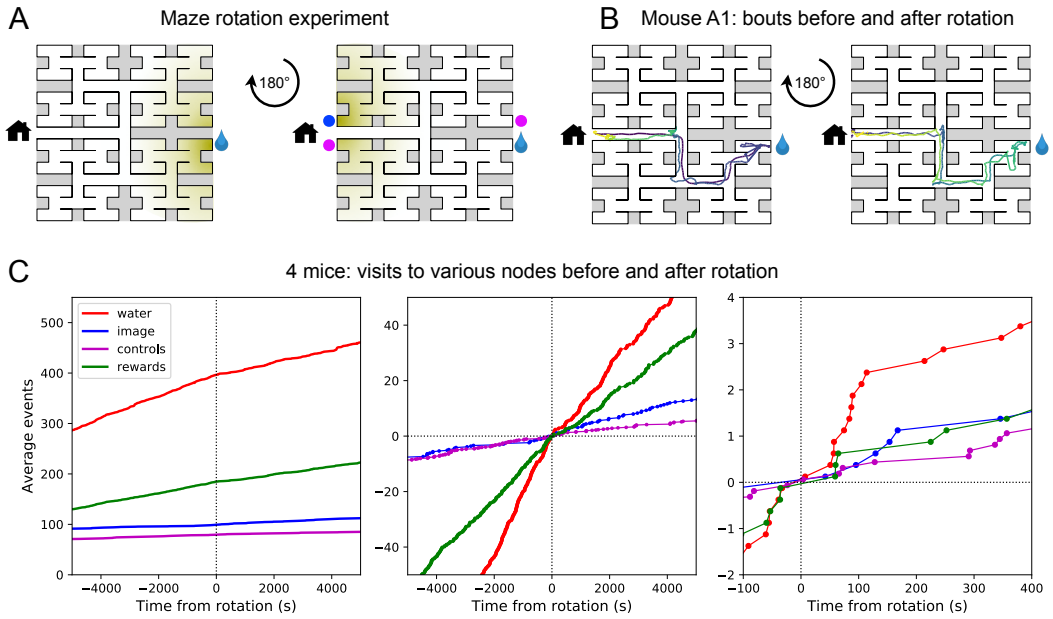

Figure 7.8: **Navigation is robust to rotation of the maze.** (A) Logic of the experiment: The animal may have deposited an odorant in the maze (shading) that is centered on the water port. After 180-degree rotation of the maze, that gradient would lead to the image of the water port (blue dot). We also measure how often the mouse goes to two control nodes (magenta dots) that are related by symmetry. (B) Trajectory of mouse 'A1' in the bouts immediately before and after maze rotation. Time coded by color from dark to light as in Figure 7.5. (C) Left: Cumulative number of rewards as well as visits to the water port, the image of the water port, and the control nodes. All events are plotted vs time before and after the maze rotation. Average over 4 animals. Middle and right: Same data with the counts centered on zero and zoomed in for better resolution.

from the water port, and on a subsequent bout simply sniffs its way up the odor gradient (Figure 7.8A). This would require no internal representation.

The following experiment offers some partial insights. Owing to the design of the labyrinth one can rotate the entire apparatus by 180 degrees, open one wall and close another, and obtain a maze with the same structure (Figure 7.8A). Alternatively one can also rotate only the floor. After such a modification, all the physical cues attached to the rotated parts now point in the wrong direction, namely to the end node 180 degrees opposite the water port (the "image location"). If the animal navigated to the goal following cues previously deposited in the maze, it should end up at that image location.

We performed a maze rotation on four animals after several hours of exposure, when they had acquired the perfect route to water. Immediately after rotation, 3 of the

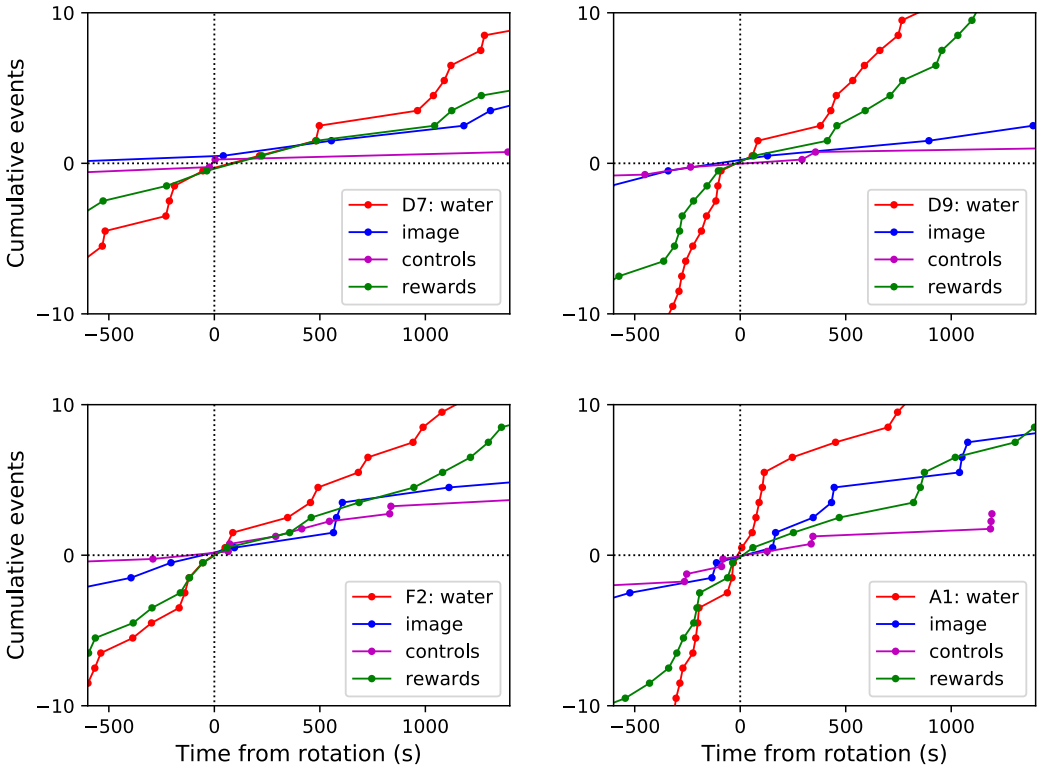

Figure 7.9: **Navigation before and after maze rotation.** Cumulative number of rewards, visits to the water port, the image of the water port, and the control nodes, plotted vs time before and after the maze rotation. Display as in Figure 7.8C, but split for each of 4 animals.

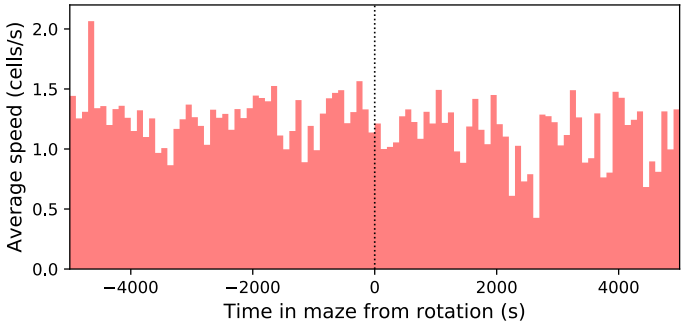

Figure 7.10: **Speed of the mouse vs time in the maze.** Average over 4 animals. Time is plotted relative to the maze rotation.

4 animals went to the correct water port on their first entry into the maze, and before ever visiting the image location (e.g. Figure 7.8B). The fourth mouse visited the image location once and then the correct water port (Figure 7.8, Figure 7.9). The mice continued to collect water rewards efficiently even immediately after the rotation.

Nonetheless, the maze rotation did introduce subtle changes in behavior that lasted for an hour or more (Figure 7.8C). Visits to the image location were at chance levels prior to rotation, then increased by a factor of 1.8. Visits to the water port declined in frequency, although they still exceeded visits to the image location by a factor of 5. The reward rate declined by a factor of 0.7. These effects could be verified for each animal (Figure 7.8, Figure 7.9). The speed of the mice was not disturbed (Figure 7.8, Figure 7.10).

In summary, for navigation to the water port the experienced animals do not strictly depend on physical cues that are attached to the maze. This includes any material they might have deposited, but also pre-existing construction details by which they may have learned to identify locations in the maze. The mice clearly notice a change in these cues, but continue to navigate effectively to the goal. This conclusion applies to the time point of the rotation, a few hours into the experiment. Conceivably the animal's navigation policy and its use of sensory cues changes in the course of learning. This and many other questions regarding the mechanisms of cognition will be taken up in a separate study.

Discontinuous learning

While an average across animals shows evidence of rapid learning (Figure 7.6) one wonders whether the knowledge is acquired gradually or discontinuously, through moments of "sudden insight." To explore this we scrutinized more closely the time line of individual water-deprived animals in their experience with the maze. The discovery of the water port and the subsequent collection of water drops at a regular rate is one clear change in behavior that relies on new knowledge. Indeed, the rate of water rewards can increase rather suddenly (Figure 7.6A), suggesting an instantaneous step in knowledge.

Over time, the animals learned the path to water not only from the entrance of the maze but from many locations scattered throughout the maze. The largest distance between the water port and an end node in the opposite half of the maze involves 12 steps through 11 intersections (Figure 7.11A). Thus we included as another behavioral

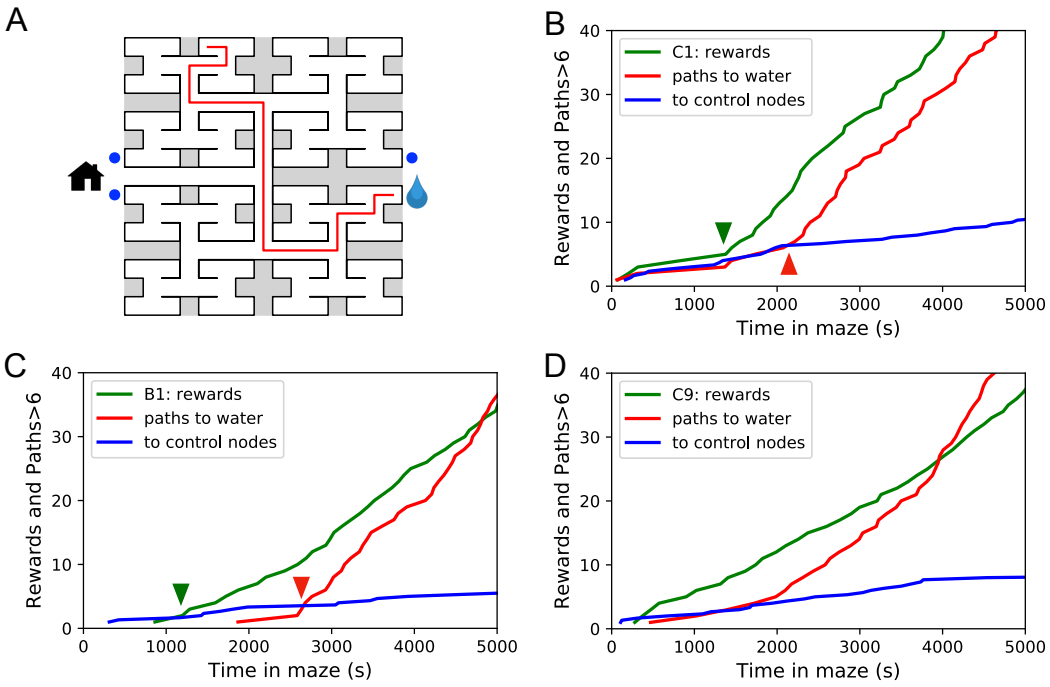

Figure 7.11: **Sudden changes in behavior.** (A) An example of a long uninterrupted path through 11 junctions to the water port (drop icon). Blue circles mark control nodes related by symmetry to the water port to assess the frequency of long paths occurring by chance. (B) For one animal (named C1) the cumulative number of rewards (green); of long paths (>6 junctions) to the water port (red); and of similar paths to the 3 control nodes (blue, divided by 3). All are plotted against the time spent in the maze. Arrowheads indicate the time of sudden changes, obtained from fitting a step function to the rates. (C) Same as B for animal B1. (D) Same as B for animal C9, an example of more continuous learning.

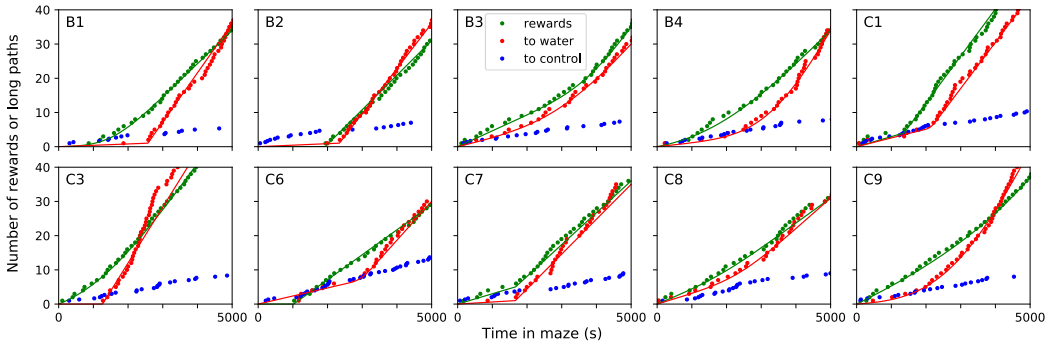

Figure 7.12: **Sudden changes in behavior for all rewarded animals.** For each of the 10 water-deprived animals this shows the cumulative rate of rewards, of long direct paths (>6 steps) to the water port, and of similar paths to 3 control nodes. Display as in Figure 7.11; panels B-D of that figure are included again here. Dots are data, lines are fits using a 4-parameter sigmoid function for the rate of occurrence of the events.

Animal	Time of step (s)	Ratio of rates after/before
B1	2580 ± 110	36.4
B2	2350 ± 220	30.3
C1	2070 ± 310	5.49
C3	1280 ± 80	1640
C7	1680 ± 280	16.9

Figure 7.13: **Statistics of sudden changes in behavior.** Summary of the steps in the rate of long paths to water detected in 5 of the 10 rewarded animals. Mean and standard deviation of the step time are derived from maximum likelihood fits of a step model to the data.

variable the occurrence of long direct paths to the water port which reflects how directedly the animals navigate within the maze.

Figure 7.11B shows for one animal the cumulative occurrence of water rewards and that of long direct paths to water. The animal discovers the water port early on at 75 s, but at 1380 s the rate of water rewards jumps suddenly by a factor of 5. The long paths to water follow a rather different time line. At first they occur randomly, at the same rate as the paths to the unrewarded control nodes. At 2070 s the long paths suddenly increase in frequency by a factor of 5. Given the sudden change in rates of both kinds of events, there is little ambiguity about when the two steps happen, and they are well separated in time (Figure 7.11B).

The animal behaves as though it gains a new insight at the time of the second step that allows it to travel to the water port directly from elsewhere in the maze. Note that the two behavioral variables are independent: the long paths don't change when the reward rate steps up, and the reward rate doesn't change when the rate of long paths steps up. Another animal (Figure 7.11C) similarly showed an early step in the reward rate (at 860 s) and a dramatic step in the rate of long paths (at 2580 s). In this case the emergence of long paths coincided with a modest increase (factor of 2) in the reward rate.

Similar discontinuities in behavior were seen in at least 5 of the 10 water-deprived animals (Figure 7.11B, Figure 7.11–Figure 7.12, Figure 7.11–Figure 7.13), and their timing could be identified to a precision of ~200 s. More gradual performance change was observed for the remaining animals (Figure 7.11 D). We varied the criterion of performance by asking for even longer error-free paths, and the results were largely unchanged and no additional discontinuity appeared. These observations suggest that mice can acquire a complex decision-making skill rather suddenly. A mouse may have multiple moments of sudden insight that affect different aspects

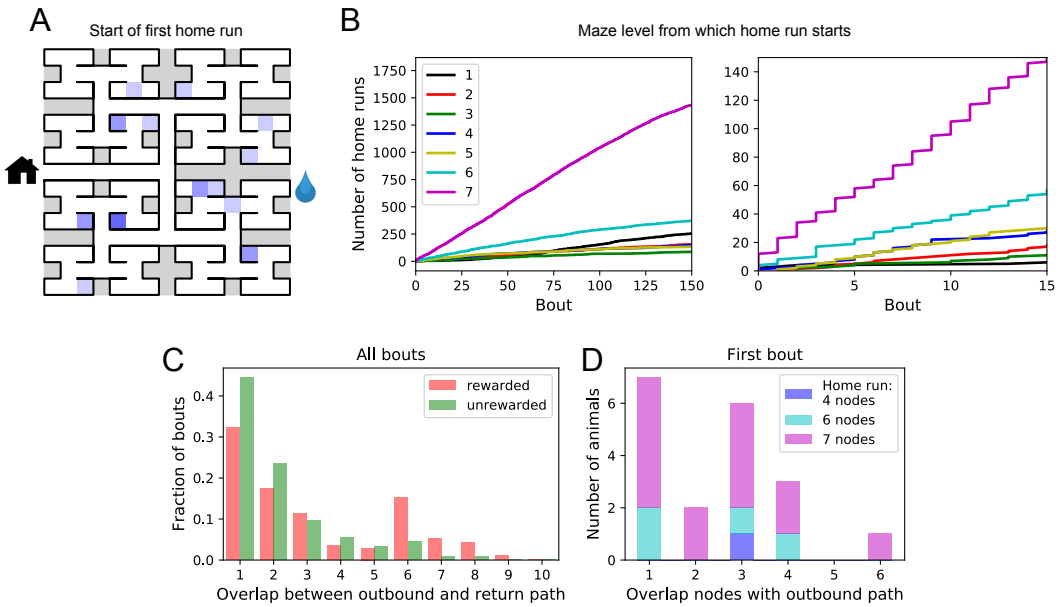

Figure 7.14: **Homing succeeds on first attempt.** (A) Locations in the maze where the 19 animals started their first return to the exit (home run). Some locations were used by 2 or 3 animals (darker color). (B) Left: The cumulative number of home runs from different levels in the maze, summed over all animals, and plotted against the bout number. Level 1 = first T-junction, level 7 = end nodes. Right: Zoom of (Left) into early bouts. (C) Overlap between the outbound and the home path. Histogram of the overlap for all bouts of all animals. (D) Same analysis for just the first bout of each animal. The length of the home run is color-coded as in Panel B.

of its behavior. The exact time of the insight cannot be predicted but is easily identified post-hoc. Future neurophysiological studies of the phenomenon will face the interesting challenge of capturing these singular events.

One-shot learning of the home path

For an animal entering an unfamiliar environment, the most important path to keep in memory may be the escape route. In the present case that is the route to the maze entrance, from which the tunnel leads home to the cage. We expected that the mice would begin by penetrating into the maze gradually and return home repeatedly so as to confirm the escape route, a pattern previously observed for rodents in an open arena [14, 36]. This might help build a memory of the home path gradually level-by-level into the binary tree. Nothing could be further from the truth.

At the end of any given bout into the maze, there is a "home run," namely the direct path without reversals that takes the animal to the exit (see Figure 7.6–Figure 7.7). Figure 7.14 A shows the nodes where each animal started its first home run, following

the first penetration into the maze. With few exceptions, that first home run began from an end node, as deep into the maze as possible. Recall that this involves making the correct choice at six successive 3-way intersections, an outcome that is unlikely to happen by chance.

The above hypothesis regarding gradual practice of home runs would predict that short home runs should appear before long ones in the course of the experiment. The opposite is the case (Figure 7.14 B). In fact, the end nodes (level 7 of the maze) are by far the favorite place from which to return to the exit, and those maximal-length home runs systematically appear before shorter ones. This conclusion was confirmed for each individual animal, whether rewarded or unrewarded.

Clearly the animals do not practice the home path or build it up gradually. Instead they seem to possess an Ariadne's thread [29] starting with their first excursion into the maze, long before they might have acquired any general knowledge of the maze layout. On the other hand the mouse does not follow the strategy of Theseus, namely to precisely retrace the path that led it into the labyrinth. In that case, the animal's home path should be the reverse of the path into the maze that started the bout. Instead the entry path and the home path tend to have little overlap (Figure 7.14C). Note the minimum overlap is 1, because all paths into and out of the maze have to pass through the central junction (node 0 in Figure 7.6–Figure 7.7). This is also the most frequent overlap. The peak at overlaps 6-8 for rewarded animals results from the frequent paths to the water port and back, a sequence of at least 7 nodes in each direction. The separation of outbound and return path is seen even on the very first home run (Figure 7.14D). Many home runs from the deepest level (7 nodes) have only the central junction in common with the outbound path (overlap = 1).

In summary it appears that the animal acquires a homing strategy over the course of a single bout, and in a manner that allows a direct return home even from locations not previously encountered.

Structure of behavior in the maze

Here we focus on rules and patterns that govern the animal's activity in the maze on both large and small scales.

Behavioral states

Once the animal has learned to perform long uninterrupted paths to the water port, one can categorize its behavior within the maze by three states: (1) walking to the

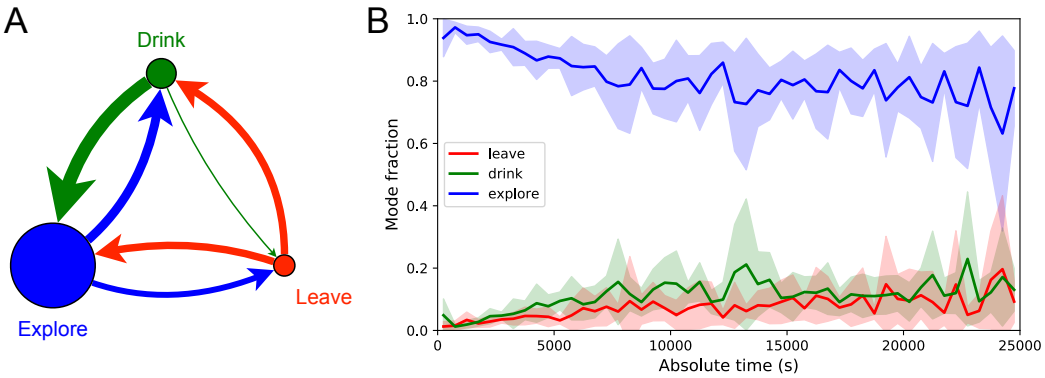

Figure 7.15: **Exploration is a dominant and persistent mode of behavior.** (A) Ethogram for rewarded animals. Area of the circle reflects the fraction of time spent in each behavioral mode averaged over animals and duration of the experiment. Width of the arrow reflects the probability of transitioning to another mode. "Drink" involves travel to the water port and time spent there. Transitions from "Leave" represent what the animal does at the start of the next bout into the maze. (B) The fraction of time spent in each mode as a function of absolute time throughout the night. Mean ± SD across the 10 rewarded animals.

A Fraction of time in modes

Mode	rewarded	unrewarded
leave	0.053 ± 0.014	0.054 ± 0.013
drink	0.103 ± 0.026	
explore	0.844 ± 0.032	0.946 ± 0.013

B Transition probability between modes: rewarded animals

from / to:	leave	drink	explore
leave		0.51 ± 0.14	0.49 ± 0.14
drink	0.10 ± 0.05		0.90 ± 0.05
explore	0.40 ± 0.11	0.60 ± 0.11	

Figure 7.16: **Three modes of behavior.** (A) The fraction of time mice spent in each of the three modes while in the maze. Mean ± SD for 10 rewarded and 9 unrewarded animals. (B) Probability of transitioning from the mode on the left to the mode at the top. Transitions from 'leave' represent what the animal does at the start of the next bout into the maze.

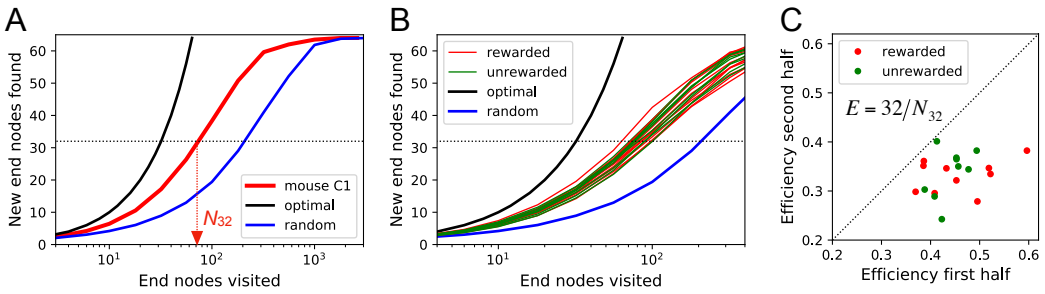

Figure 7.17: **Exploration covers the maze efficiently.** (A) The number of distinct end nodes encountered as a function of the number of end nodes visited for: mouse C1 (red); the optimal explorer agent (black); an unbiased random walk (blue). Arrowhead: The value $N_{32} = 76$ by which mouse C1 discovered half of the end nodes. (B) An expanded section of the graph in A including curves from 10 rewarded (red) and 9 unrewarded (green) animals. The efficiency of exploration, defined as $E = 32/N_{32}$, is 0.385 ± 0.050 (SD) for rewarded and 0.384 ± 0.039 (SD) for unrewarded mice. (C) The efficiency of exploration for the same animals, comparing the values in the first and second halves of the time in the maze. The decline is a factor of 0.74 ± 0.12 (SD) for rewarded and 0.81 ± 0.13 (SD) for unrewarded mice.

water port; (2) walking to the exit; and (3) exploring the maze. Operationally we define exploration as all periods in which the animal is in the maze but not on a direct path to water or to the exit. For the ten sated animals this includes all times in the maze except for the walks to the exit.

Figure 7.15 illustrates the occupancies and transition probabilities between these states. The animals spent most of their time by far in the exploration state: 84% for rewarded and 95% for unrewarded mice. Across animals there was very little variation in the balance of the 3 modes (Figure 7.15–Figure 7.16). The rewarded mice began about half their bouts into the maze with a trip to the water port and the other half by exploring (Figure 7.15A). After a drink, the animals routinely continued exploring, about 90% of the time.

For water-deprived animals the dominance of exploration persisted even at a late stage of the night when they routinely executed perfect exploitation bouts to and from the water port: Over the duration of the night the "explore" fraction dropped slightly from 0.92 to 0.75, with the balance accrued to the "drink" and "leave" modes as the animals executed many direct runs to the water port and back. The unrewarded group of animals also explored the maze throughout the night even though it offered no overt rewards (Figure 7.15–Figure 7.16). One suspects that the animals derive some intrinsic reward from the act of patrolling the environment itself.

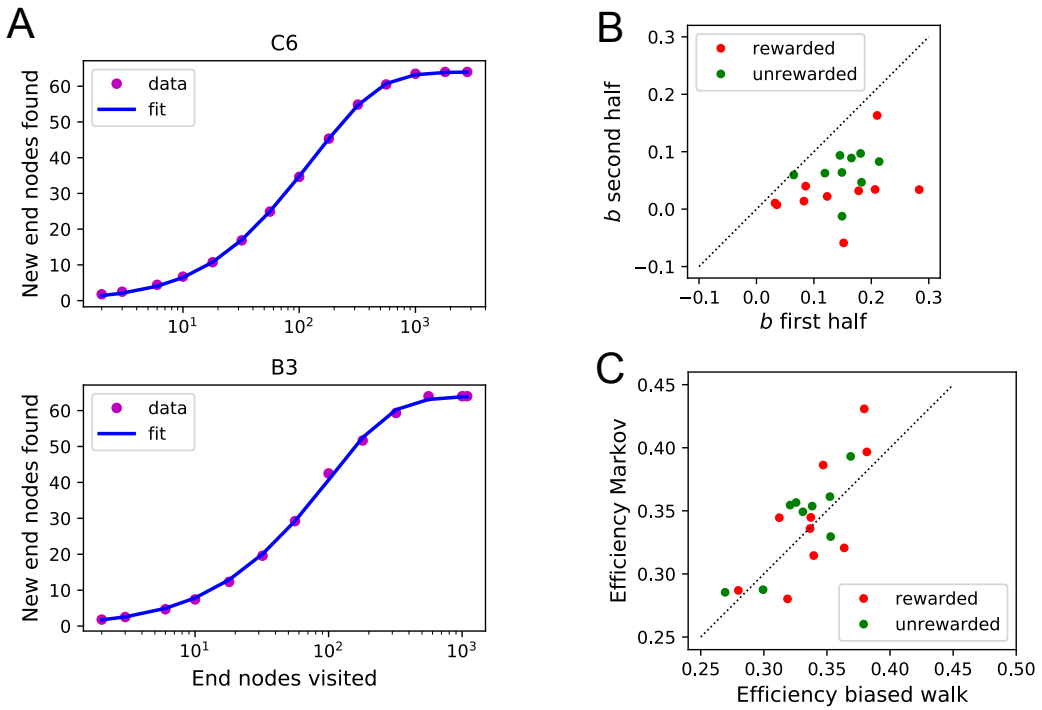

Figure 7.18: **Functional fits to measure exploration efficiency.** (A) Fitting Equation 7.12 to the data from the mouse's exploration. Animals with best fit (top) and worst fit (bottom). The relative uncertainty in the two fit parameters a and b was only 0.0038 ± 0.0020 (mean ± SD across animals). (B) The fit parameter b for all animals, comparing the first to the second half of the night. (C) The efficiency E (Equation 7.1) predicted from two models of the mouse's trajectory: The 4-bias random walk (Figure 7.22D) and the optimal Markov chain (Figure 7.22C).

Efficiency of exploration

During the direct paths to water and to the exit the animal behaves deterministically, whereas the exploration behavior appears stochastic. Here we delve into the rules that govern the exploration component of behavior.

One can presume that a goal of the exploratory mode is to rapidly survey all parts of the environment for the appearance of new resources or threats. We will measure the efficiency of exploration by how rapidly the animal visits all end nodes of the binary maze, starting at any time during the experiment. The optimal agent with perfect memory and complete knowledge of the maze—including the absence of any loops—could visit the end nodes systematically one after another without repeats, thus encountering all of them after just 64 visits. A less perfect agent, on the other hand, will visit the same node repeatedly before having encountered all of them. Figure 7.17A plots for one exploring mouse the number of distinct end nodes it

encountered as a function of the number of end nodes visited. The number of new nodes rises monotonically; 32 of the end nodes have been discovered after the mouse checked 76 times; then the curve gradually asymptotes to 64. We will characterize the efficiency of the search by the number of visits N_{32} required to survey half the end nodes, and define

$$E = \frac{32}{N_{32}}. \tag{7.1}$$

This mouse explores with efficiency $E = 32/76 = 0.42$. For comparison, Figure 7.17A plots the performance of the optimal agent ($E = 1.0$) and that of a random walker that makes random decisions at every 3-way junction ($E = 0.23$). Note the mouse is about half as efficient as the optimal agent, but twice as efficient as a random walker.

The different mice were remarkably alike in this component of their exploratory behavior (Figure 7.17B): across animals the efficiency varied by only 11% of the mean (0.387 ± 0.044 SD). Furthermore there was no detectable difference in efficiency between the rewarded animals and the sated unrewarded animals. Over the course of the night the efficiency declined significantly for almost every animal—whether rewarded or not—by an average of 23% (Figure 7.17C).

Rules of exploration

What allows the mice to search much more efficiently than a random walking agent? We inspected more closely the decisions that the animals make at each 3-way junction. It emerged that these decisions are governed by strong biases (Figure 7.19). The probability of choosing each arm of a T-junction depends crucially on how the animal entered the junction. The animal can enter a T-junction from 3 places and exit it in 3 directions (Figure 7.19A). By tallying the frequency of all these occurrences across all T-junctions in the maze, one finds clear deviations from an unbiased random walk (Figure 7.19B, Figure 7.19–Table 7.20).

First, the animals have a strong preference for proceeding through a junction rather than returning to the preceding node (P_{SF} and P_{BF} in Figure 7.19B). Second there is a bias in favor of alternating turns left and right rather than repeating the same direction turn (P_{SA}). Finally, the mice have a mild preference for taking a branch off the straight corridor rather than proceeding straight (P_{BS}). A comparison across animals again revealed a remarkable degree of consistency even in these local rules of behavior: the turning biases varied by only 3% across the population and even

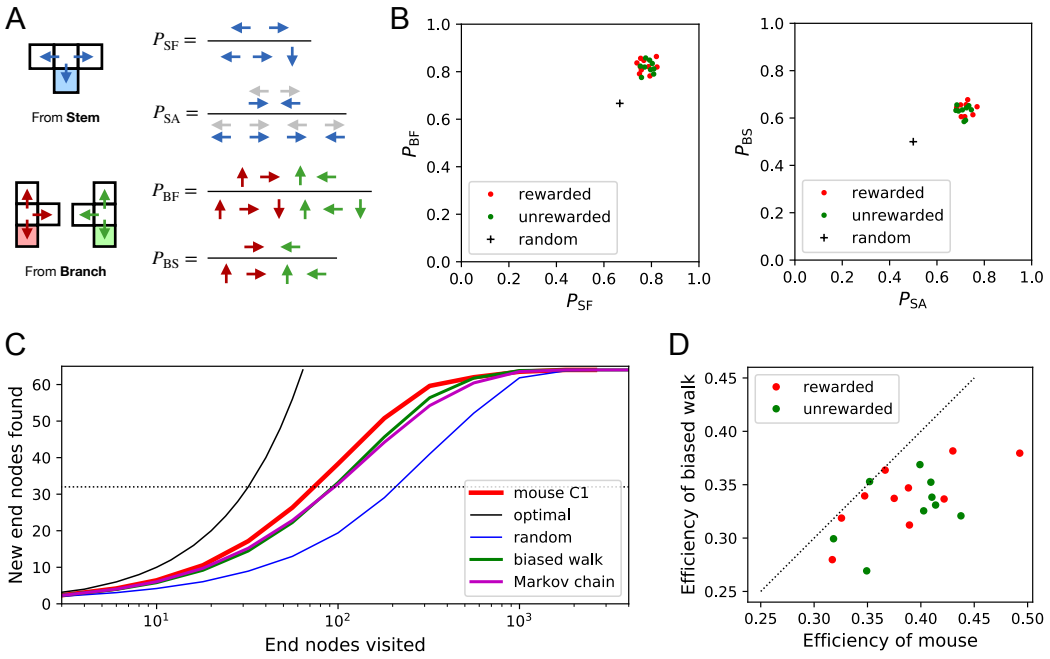

Figure 7.19: **Turning biases favor exploration.** (A) Definition of four turning biases at a T-junction based on the ratios of actions taken. Top: An animal arriving from the stem of the T (shaded) may either reverse or turn left or right. P_{SF} is the probability that it will move forward rather than reversing. Given that it moves forward, P_{SA} is the probability that it will take an alternating turn from the preceding one (gray), i.e. left-right or right-left. Bottom: An animal arriving from the bar of the T may either reverse or go straight, or turn into the stem of the T. P_{BF} is the probability that it will move forward through the junction rather than reversing. Given that it moves forward, P_{BS} is the probability that it turns into the stem. (B) Scatter graph of the biases P_{SF} and P_{BF} (left) and P_{SA} and P_{BS} (right). Every dot represents a mouse. Cross: values for an unbiased random walk. (C) Exploration curve of new end nodes discovered vs end nodes visited, displayed as in Figure 7.17A, including results from a biased random walk with the 4 turning biases derived from the same mouse, as well as a more elaborate Markov-chain model (see Figure 7.22C). (D) Efficiency of exploration (Equation 7.1) in 19 mice compared to the efficiency of the corresponding biased random walk.

Bias	rewarded	unrewarded
P_{SF}	0.77 ± 0.03	0.78 ± 0.02
P_{SA}	0.72 ± 0.02	0.71 ± 0.02
P_{BF}	0.82 ± 0.03	0.81 ± 0.03
P_{BS}	0.64 ± 0.02	0.63 ± 0.02

Figure 7.20: **Statistics of the four turning biases.** Mean and standard deviation of the 4 biases of Figure 7.19A-B across animals in the rewarded and unrewarded groups.

between the rewarded and unrewarded groups (Figure 7.19B, Figure 7.19–Table 7.20).

Qualitatively, one can see that these turning biases will improve the animal's search strategy. The forward biases P_{SF} and P_{BF} keep the animal from re-entering territory it has covered already. The bias P_{BS} favors taking a branch that leads out of the maze. This allows the animal to rapidly cross multiple levels during an outward path and then enter a different territory. By comparison, the unbiased random walk tends to get stuck in the tips of the tree and revisits the same end nodes many times before escaping. To test this intuition we simulated a biased random agent whose turning probabilities at a T-junction followed the same biases as measured from the animal (Figure 7.19C). These biased agents did in fact search with much higher efficiency than the unbiased random walk. They did not fully explain the behavior of the mice (Figure 7.19D), accounting for ~87% of the animal's efficiency (compared to 60% for the random walk). A more sophisticated model of the animal's behavior—involving many more parameters (Figure 7.22C)—failed to get any closer to the observed efficiency (Figure 7.19C, Figure 7.17–Figure 7.18C). Clearly some components of efficient search in these mice remain to be understood.

Systematic node preferences

A surprising aspect of the animals' explorations is that they visit certain end nodes of the binary tree much more frequently than others (Figure 7.21). This effect is large: more than a factor of 10 difference between the occupancy of the most popular and least popular end nodes (Figure 7.21A-B). This was surprising given our efforts to design the maze symmetrically, such that in principle all end nodes should be equivalent. Furthermore the node preferences were very consistent across animals and even across the rewarded and unrewarded groups. Note that the standard

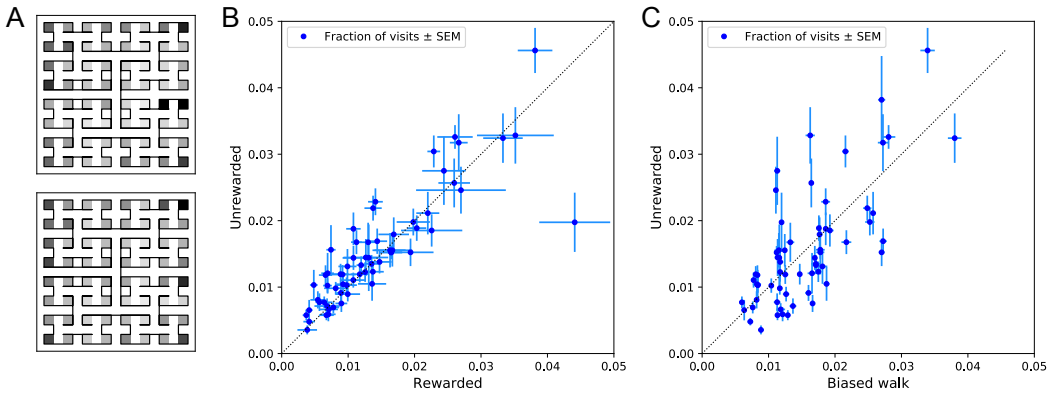

Figure 7.21: **Preference for certain end nodes during exploration.** (A) The number of visits to different end nodes encoded by a gray scale. Top: rewarded, bottom: unrewarded animals. Gray scale spans a factor of 12 (top) or 13 (bottom). (B) The fraction of visits to each end node, comparing the rewarded vs unrewarded group of animals. Each data point is for one end node, the error bar is the SEM across animals in the group. The outlier on the bottom right is the neighbor of the water port, a frequently visited end node among rewarded animals. The water port is off scale and not shown. (C) As in Panel B but comparing the unrewarded animals to their simulated 4-bias random walks. These biases explain 51% of the variance in the observed preference for end nodes.

error across animals of each node's occupancy is much smaller than the differences between the nodes (Figure 7.21B).

The nodes on the periphery of the maze are systematically preferred. Comparing the outermost ring of 26 end nodes (excluding the water port and its neighbor) to the innermost 16 end nodes, the outer ones are favored by a large factor of 2.2. This may relate to earlier reports of a "centrifugal tendency" among rats patrolling a maze [41].

Interestingly, the biased random walk using four bias numbers (Figure 7.19, Figure 7.22D) replicates a good amount of the pattern of preferences. For unrewarded animals, where the maze symmetry is not disturbed by the water port, the biased random walk predicts 51% of the observed variance across nodes (Figure 7.21C), and an outer/inner node preference of 1.97, almost matching the observed ratio of 2.20. The more complex Markov-chain model of behavior (Figure 7.22C) performed slightly better, explaining 66% of the variance in port visits and matching the outer/inner node preference of 2.20.

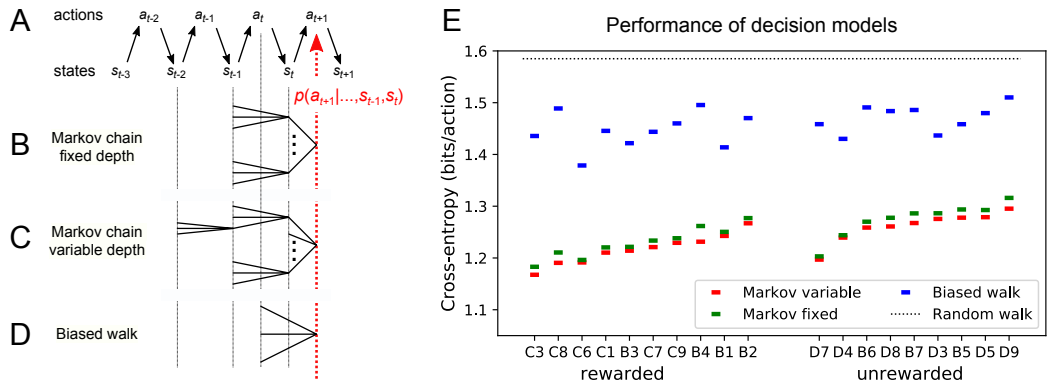

Figure 7.22: **Recent history constrains the mouse's decisions.** (A) The mouse's trajectory through the maze produces a sequence of states s_t = node occupied after step t. From each state, up to 3 possible actions lead to the next state (end nodes allow only one action). We want to predict the animal's next action, a_{t+1}, based on the prior history of states or actions. (B-D) Three possible models to make such a prediction. (B) A fixed-depth Markov chain where the probability of the next action depends only on the current state s_t and the preceding state s_{t-1}. The branches of the tree represent all 3×127 possible histories (s_{t-1}, s_t). (C) A variable-depth Markov chain where only certain branches of the tree of histories contribute to the action probability. Here one history contains only the current state, some others reach back three steps. (D) A biased random walk model, as defined in Figure 7.19, in which the probability of the next action depends only on the preceding action, not on the state. (E) Performance of the models in (B,C,D) when predicting the decisions of the animal at T-junctions. In each case we show the cross-entropy between the predicted action probability and the real actions of the animal (lower values indicate better prediction, perfect prediction would produce zero). Dotted line represents an unbiased random walk with 1/3 probability of each action.

Models of maze behavior

Moving beyond the efficiency of exploration, one may ask more broadly: How well do we really understand what the mouse does in the maze? Can we predict its action at the next junction? Once the predictable component is removed, how much intrinsic randomness remains in the mouse's behavior? Here we address these questions using more sophisticated models that predict the probability of the mouse's future actions based on the history of its trajectory.

At a formal level, the mouse's trajectory through the maze is a string of numbers standing for the nodes the animal visited (Figure 7.22A and Figure 7.6–Figure 7.7). We want to predict the next action of the mouse, namely the step that takes it to the next node. The quality of the model will be assessed by the cross-entropy between the model's predictions and the mouse's observed actions, measured in bits per

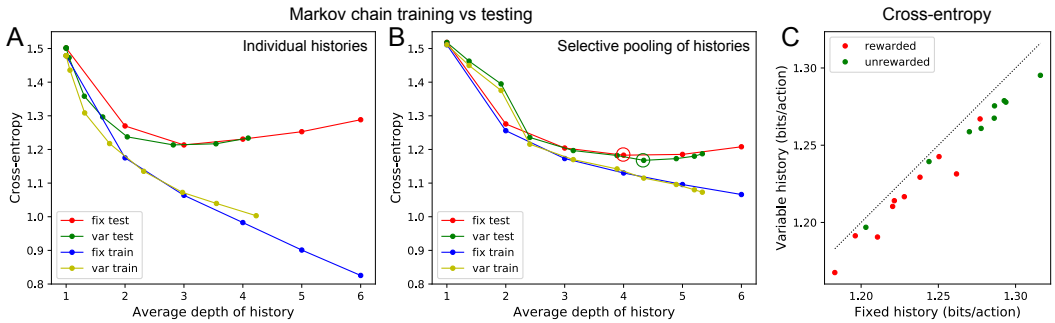

Figure 7.23: **Fitting Markov models of behavior.** (A) Results of fitting the node sequence of a single animal (C3) with Markov models having a fixed depth ("fix") or variable depth ("var"). The cross-entropy of the model's prediction is plotted as a function of the average depth of history. In both cases we compare the results obtained on the training data ("train") vs those on separate testing data ("test"). Note that at larger depth the "test" and "train" estimates diverge, a sign of over-fitting the limited data available. (B) As in (A) but to combat the data limitation we pooled the counts obtained at all nodes that were equivalent under the symmetry of the maze (see Methods). Note considerably less divergence between "train" and "test" results, and a slightly lower cross-entropy during "test" than in (A). (C) The minimal cross-entropy (circles in (B)) produced by variable vs fixed history models for each of the 19 animals. Note the variable history model always produces a better fit to the behavior.

action. This is the uncertainty that remains about the mouse's next action given the prediction from the model. The ultimate lower limit is the true source entropy of the mouse, namely that component of its decisions that cannot be explained by the history of its actions.

One family of models we considered are fixed-depth Markov chains (Figure 7.22B). Here the probability of the next action a_{t+1} is specified as a function of the history stretching over the k preceding nodes $(s_{t-k+1}, \ldots, s_t)$. In fitting the model to the mouse's actual node sequence, one tallies how often each history leads to each action, and uses those counts to estimate the conditional probabilities $p(a_{t+1}|s_{t-k+1}, \ldots, s_t)$. Given a new node sequence, the model will then use the history strings $(s_{t-k+1}, \ldots, s_t)$ to predict the outcome of the next action. In practice we trained the model on 80% of the animal's trajectory and tested it by evaluating the cross-entropy on the remaining 20%.

Ideally, the depth k of these action trees would be very large, so as to take as much of the prior history into account as possible. However, one soon runs into a problem of over-fitting: because each T-junction in the maze has 3 neighboring junctions, the

number of possible histories grows as 3^k. As k increases, this quickly exceeds the length of the measured node sequence, so that every history appears only zero or one times in the data. At this point one can no longer estimate any probabilities, and cross-validation on a different segment of data fails catastrophically. In practice we found that this limitation sets in already beyond $k = 2$ (Figure 7.22–Figure 7.23A). To address this issue of data-limitation we developed a variable-depth Markov chain (Figure 7.22C). This model retains longer histories, but only if they occur frequently enough to allow a reliable probability estimate (see Methods, Figure 7.22–Figure 7.23B-C). In addition, we explored different schemes of pooling the counts across certain T-junctions that are related by the symmetry of the maze (see Methods).

With these methods we focused on the portions of trajectory when the mouse was in "explore" mode, because the segments in "drink" and "leave" mode are fully predictable. Furthermore, we evaluated the models only at nodes corresponding to T-junctions, because the decision from an end node is again fully predictable. Figure 7.22E compares the performance of various models of mouse behavior. The variable-depth Markov chains routinely produced the best fits, although the improvement over fixed-depth models was modest. Across all 19 animals in this study, the remaining uncertainty about the animal's action at a T-junction is 1.237 ± 0.035 (SD) bits/action, compared to the prior uncertainty of $\log_2 3 = 1.585$ bits. The rewarded animals have slightly lower entropy than the unrewarded ones (1.216 vs 1.261 bits/action). The Markov chain models that produced the best fits to the behavior used history strings with an average length of ~4.

We also evaluated the predictions obtained from the simple biased random walk model (Figure 7.22D). Recall that this attempts to capture the history-dependence with just 4 bias parameters (Figure 7.19A). As expected this produced considerably higher cross-entropies than the more sophisticated Markov chains (by about 18%, Figure 7.22E). Finally we used several professional file compression routines to try and compress the mouse's node sequence. In principle, this sets an upper bound on the true source entropy of the mouse, even if the compression algorithm has no understanding of animal behavior. The best such algorithm (bzip2 compression [33]) far under-performed all the other models of mouse behavior, giving 43% higher cross-entropy on average, and thus offered no additional useful bounds.

We conclude that during exploration of the maze, the mouse's choice behavior is strongly influenced by its current location and ~3 locations preceding it. There are minor contributions from states further back. By knowing the animal's history one

can narrow down its action plan at a junction from the a priori 1.59 bits (one of three possible actions) to just ~1.24 bits. This finally is a quantitative answer to the question, "How well can one predict the animal's behavior?" Whether the remainder represents an irreducible uncertainty—akin to "free will" of the mouse—remains to be seen. Readers are encouraged to improve on this number by applying their own models of behavior to our published data set.

7.3 Discussion

Summary of contributions

We present a new approach to the study of learning and decision-making in mice. We give the animal access to a complex labyrinth and leave it undisturbed for a night while monitoring its movements. The result is a rich data set that reveals new aspects of learning and the structure of exploratory behavior. With these methods we find that mice learn a complex task that requires 6 correct 3-way decisions after only ~10 experiences of success (Figure 7.5, Figure 7.6). Along the way the animal gains task knowledge in discontinuous steps that can be localized to within a few minutes of resolution (Figure 7.11). Underlying the learning process is an exploratory behavior that occupies 90% of the animal's time in the maze and persists long after the task has been mastered, even in complete absence of an extrinsic reward (Figure 7.15). The decisions the animal makes at choice points in the labyrinth are constrained in part by the history of its actions (Figure 7.19, Figure 7.22), in a way that favors efficient searching of the maze (Figure 7.17). This microstructure of behavior is surprisingly consistent across mice, with variation in parameters of only a few percent (Figure 7.19). Our most expressive models to predict the animal's choices still leave a remaining uncertainty of ~1.24 bits per decision (Figure 7.22), a quantitative benchmark by which competing models can be tested. Finally, some of the observations constrain what algorithms the animals might use for learning and navigation (Figure 7.8).

Historical context

Mazes have been a staple of animal psychology for well over 100 years. The early versions were true labyrinths. For example, Small [35] built a model of the maze in Hampton Court gardens scaled to rat size. Subsequent researchers felt less constrained by Victorian landscapes and began to simplify the maze concept. Most commonly the maze offered one standard path from a starting location to a food reward box. A few blind alleys would branch from the standard path, and researchers

would tally how many errors the animal committed by briefly turning into a blind [39]. Later on, the design was further reduced to a single T-junction. After all, the elementary act of maze navigation is whether to turn left or right at a junction [38], so why not study that process in isolation? And reducing the concept even further, one can ask the animal to refrain from walking altogether, and instead poke its nose into a hole on the left or the right side of a box [40]. This led to the popular behavior boxes now found in rodent neuroscience laboratories everywhere. Each of these reductions of the "maze" concept enabled a new type of experiment to study learning and decision-making, for example limiting the number of choice points allows one to better sample neural activity at each one. However, the essence of a "confusing network of paths" has been lost along the way, and with it the behavioral richness of the animals navigating those decisions.

Owing in part to the dissemination of user-friendly tools for animal tracking, one sees a renaissance of experiments that embrace complex environments, including mazes with many choice points [1, 21, 24, 30, 32, 44, 46], 3-dimensional environments [16], and infinite mazes [34]. The labyrinth in the present study is considerably more complex than Hampton Court or most of the mazes employed by Tolman and others [8, 22, 39]. In those mazes the blind alleys are all short and unbranched; when an animal strays from the target path it receives feedback quickly and can correct. By contrast our binary tree maze has 64 equally deep branches, only one of which contains the reward port. If the animal makes a mistake at any level of the tree, it can find out only after traveling all the way to the last node.

Another crucial aspect of our experimental design is the absence of any human interference. Most studies of animal navigation and learning involve some kind of trial structure. For example the experimenter puts the rat in the start box, watches it make its way through the maze, coaxes it back on the path if necessary, and picks it up once it reaches the target box. Then another trial starts. In modern experiments with two-alternative-forced-choice (2-AFC) behavior boxes, the animal doesn't have to be picked up, but a trial starts with the appearance of a cue, and then proceeds through some strict protocol through delivery of the reward. The argument in favor of imposing a trial structure is that it creates reproducible conditions, so that one can gather comparable data and average them suitably over many trials.

Our experiments had no imposed structure whatsoever; in fact it may be inappropriate to call them experiments. The investigator opened the entry to the maze in the evening and did not return until the morning. A potential advantage of leaving the

animals to themselves is that they are more likely to engage in mouse-like behavior, rather than constantly responding to the stress of human interference or the alienation from being a cog in a behavior machine. The result was a rich data set, with the typical animal delivering ~15,000 decisions in a single night, even if one only counts the nodes of the binary tree as decision points. Since the mice made all the choices, the scientific effort lay primarily in adapting methods of data analysis to the nature of mouse trajectories. Somewhat surprisingly, the absence of experimental structure was no obstacle to making precise and reproducible measurements of the animal's behavior.

How fast do animals learn?

Among the wide range of phenomena of animal learning, one can distinguish easy and hard tasks by some measure of task complexity. In a simple picture of a behavioral task the animal needs to recognize several different contexts, and based on that, express one of several different actions. One can draw up a contingency table between contexts and actions, and measure the complexity of the task by the mutual information in that table. This ignores any task difficulties associated with sensing the context at all or with producing the desired actions. However, in all the examples discussed here the stimuli are discriminated easily and the actions come naturally, thus the learning difficulty lies only in forming the associations, not in sharpening the perceptual mechanisms or practicing complex motor output.

Many well-studied behaviors have a complexity of 1 bit or less, and often animals can learn these associations after a single experience. For example, in the Bruce effect [7] the female maps two different contexts (smell of mate vs non-mate) onto two kinds of pregnancy outcomes (carry to term vs abort). The mutual information in that contingency table is at most 1 bit, and may be considerably lower, for example if non-mate males are very rare or very frequent. Mice form the correct association after a single instance of mating, although proper memory formation requires several hours of exposure to the mate odor [31].

Similarly, fear learning under the common electroshock paradigm establishes a mapping between two contexts (paired with shock vs innocuous) and two actions (freeze vs proceed), again with an upper bound of 1 bit of complexity. Rats and mice will form the association after a single experience lasting only seconds, and alter their behavior over several hours [5, 13]. This is an adaptive warning system to deal with life-threatening events, and rapid learning here has a clear survival value.

Animals are particularly adept at learning a new association between an odor and food. For example bees will extend their proboscis in response to a new odor after just one pairing trial where the odor appeared together with sugar [4]. Similarly rodents will start digging for food in a scented bowl after just a few pairings with that odor [11]. Again, these are 1-bit tasks learned rapidly after one or a few experiences.

By comparison the tasks that a mouse performs in the labyrinth are more complex. For example, the path from the maze entrance to the water port involves 6 junctions, each with 3 options. At a minimum 6 different contexts must be mapped correctly into one of 3 actions each, which involves $6 \cdot \log_2 3 = 9.5$ bits of complexity. The animals begin to execute perfect paths from the entrance to the water port well within the first hour (Figure 7.5C, Figure 7.6B). At a later stage during the night the animal learns to walk direct paths to water from many different locations in the maze (Figure 7.11); by this time it has consumed 10-20 rewards. In the limit, if the animal could turn correctly towards water from each of 63 junctions in the maze, it would have learned $63 \cdot \log_2 3 = 100$ bits. Conservatively we estimate that the animals have mastered 10-20 bits of complexity based on 10-20 reward experiences within an hour of time spent in the maze. Note this considers only information about the water port and ignores whatever else the animals are learning about the maze during their incessant exploratory forays. These numbers align well with classic experiments on rats in diverse mazes and problem boxes [22]. Although those tasks come in many varieties, a common theme is that ~10 successful trials are sufficient to learn ~10 decisions [45].

In a different corner of the speed-complexity space are the many 2-alternative-forced-choice (2AFC) tasks in popular use today. These tend to be 1-bit tasks, for example the monkey should flick its eyes to the left when visual motion is to the left [26], or the mouse should turn a steering wheel to the right when a light appears on the left [9]. Yet, the animals take a long time to learn these simple tasks. For example, the mouse with the steering wheel requires about 10,000 experiences before performance saturates. It never gets particularly good, with a typical hit rate only 2/3 of the way from random to perfect. All this training takes 3-6 weeks; in the case of monkeys several months. The rate of learning, measured in task complexity per unit time, is surprisingly low: < 1 bit/month compared to ~10 bits/h observed in the labyrinth. The difference is a factor of 6,000. Similarly when measured in complexity learned per reward experience, the 2-AFC mouse may need 5,000 rewards to learn a contingency table with 1 bit complexity, whereas the mouse in the maze needs ~10 rewards

to learn 10 bits. Given these enormous differences in learning rate, one wonders whether the ultra-slow mode of learning has any relevance for an animal's natural condition. In the month that the 2AFC mouse requires to finally report the location of a light, its relative in the wild has developed from a baby to having its own babies. Along the way, that wild mouse had to make many decisions, often involving high stakes, without the benefit of 10,000 trials of practice.

Sudden insight

The dynamics of the learning process are often conceived as a continuously growing association between stimuli and actions, with each reinforcing experience making an infinitesimal contribution. The reality can be quite different. When a child first learns to balance on a bicycle, performance goes from abysmal to astounding within a few seconds. The timing of such a discontinuous step in performance seems impossible to predict but easy to recognize after the fact.

From the early days of animal learning experiments there have been warnings against the tendency to average learning curves across subjects [12, 19]. The average of many discontinuous curves will certainly look continuous and incremental, but that reassuring shape may miss the essence of the learning process. A recent reanalysis of many Pavlovian conditioning experiments suggested that discontinuous steps in performance are the rule rather than the exception [15]. Here we found that the same applies to navigation in a complex labyrinth. While the average learning curve presents like a continuous function (Figure 7.6B), the individual records of water rewards show that each animal improves rather quickly but at different times (Figure 7.6A).

Owing to the unstructured nature of the experiment, the mouse may adopt different policies for getting to the water port. In at least half the animals, we observed a discontinuous change in that policy, namely when the animal started using efficient direct paths within the maze (Figure 7.11, Figure 7.11–Figure 7.13). This second switch happened considerably after the animal started collecting rewards, and did not greatly affect the reward rate. Furthermore, the animals never reverted to the less efficient policy, just as a child rarely unlearns to balance a bicycle.

Presumably this switch in performance reflects some discontinuous change in the animal's internal model of the maze, what Tolman called the "cognitive map" [2, 37]. In the unrewarded animals we could not detect any discontinuous change in the use of long paths. However, as Tolman argued, those animals may well

acquire a sophisticated cognitive map that reveals itself only when presented with a concrete task, like finding water. Future experiments will need to address this. The discontinuous changes in performance pose a challenge to conventional models of reinforcement learning, in which reward events are the primary driver of learning and each event contributes an infinitesimal update to the action policy. It will also be important to model the acquisition of distinct kinds of knowledge that contribute to the same behavior, like the location of the target and efficient routes to approach it.

Exploratory behavior

By all accounts the animals spent a large fraction of the night exploring the maze (Figure 7.1–Figure 7.3). The water-deprived animals continued their forays into the depths of the maze long after they had found the water port and learned to exploit it regularly. After consuming a water reward they wandered off into the maze 90% of the time (Figure 7.15B) instead of lazily waiting in front of the port during the timeout period. The sated animals experienced no overt reward from the maze, yet they likewise spent nearly half their time exploring that environment. As has been noted many times, animals—like humans—derive some form of intrinsic reward from exploration [3]. Some have suggested that there exists a homeostatic drive akin to hunger and thirst that elicits the information-seeking activity, and that the drive is in turn sated by the act of exploration [18]. If this were the case, then the drive to explore should be weakest just after an episode of exploration, much as the drive for food-seeking is weaker after a big meal.

Our observations are in conflict with this notion. The animal is most likely to enter the maze within the first minute of its return to the cage (Figure 7.1–Figure 7.4), a strong trend that runs opposite to the prediction from satiation of curiosity. Several possible explanations come to mind: (1) On these very brief visits to the cage, the animal may just want to certify that the exit route to the safe environment still exists, before continuing with exploration of the maze. (2) The temporal contrast between the boredom of the cage and the mystery of the maze is highest right at the moment of exit from the maze, and that may exert pressure to re-enter the maze. Understanding this in more detail will require dedicated experiments. For example, one could deliberately deprive the animals of access to the maze for some hours, and test whether that results in an increased drive to explore, as observed for other homeostatic drives around eating, drinking, and sleeping.

When left to their own devices, mice choose to spend much of their time engaged in

exploration. One wonders how that affects their actions when they are strapped into a rigid behavior machine, like a 2AFC choice box. Presumably the drive to explore persists, perhaps more so because the forced environment is so unpleasant. And within the confines of the two alternatives, the only act of exploration the mouse has left is to give the wrong answer. This would manifest as an unexpectedly high error rate on unambiguous stimuli, sometimes called the "lapse rate" [10, 28]. The fact that the lapse rate decreases only gradually over weeks to months of training [9] suggests that it is difficult to crush the animal's drive to explore.

The animals in our experiments had never been presented with a maze environment, yet they quickly settled into a steady mode of exploration. Once a mouse progressed beyond the first intersection it typically entered deep into the maze to one or more end nodes (Figure 7.14). Within 50 s of the first entry, the animals adopted a steady speed of locomotion that they would retain throughout the night (Figure 7.5–Figure 7.10). Within 250 s of first contact with the maze, the average animal already spent 50% of its time there (Figure 7.1–Figure 7.3). Contrast this with a recent study of "free exploration" in an exposed arena: Those animals required several hours before they even completed one walk around the perimeter [14]. Here the drive to explore is clearly pitted against fear of the open space, which may not be conducive to observing exploration per se.

The persistence of exploration throughout the entire duration of the experiment suggests that the animals are continuously surveying the environment, perhaps expecting new features to arise. These surveys are quite efficient: the animals cover all parts of the maze much faster than expected from a random walk (Figure 7.17). Effectively they avoid re-entering territory they surveyed just recently. It is often assumed that this requires some global memory of places visited in the environment [24, 27]. Such memory would have to persist for a long time: surveying half of the available end nodes typically required 450 turning decisions. However, we found that a global long-term memory is not needed to explain the efficient search. The animals seem to be governed by a set of local turning biases that require memory only of the most recent decision and no knowledge of location (Figure 7.19). These local biases alone can explain most of the character of exploration without any global understanding or long-term memory. Incidentally, they also explain other seemingly global aspects of the behavior, for example the systematic preference that the mice have for the outer rather than the inner regions of the maze (Figure 7.21). Of course, this argument does not exclude the presence of a long-term memory, which may

reveal itself in some other feature of the behavior.

Perhaps the most remarkable aspect of these biases is how similar they are across all 19 mice studied here, regardless of whether the animal experienced water rewards or not (Figure 7.19B, Figure 7.19–Table 7.20), and independent of the sex of the mouse. The four decision probabilities were identical across individuals to within a standard deviation of <0.03. We cannot think of a trivial reason why this should be so. For example the two biases for forward motion (Figure 7.19B left) are poised halfway between the value for a random walk ($p = 2/3$) and certainty ($p = 1$). At either of those extremes, simple saturation might lead to a reproducible value, but not in the middle of the range. Why do different animals follow the exact same decision rules at an intersection between tunnels? Given that tunnel systems are part of the mouse's natural ecology, it is possible that those rules are innate and determined genetically. Indeed the rules by which mice build tunnels have a strong genetic component [42], so the rules for using tunnels may be written in the genes as well. The high precision with which one can measure those behaviors even in a single night of activity opens the way to efficient comparisons across genotypes, and also across animals with different developmental experience.

Finally, after mice discover the water port and learn to access it from many different points in the maze (Figure 7.11) they are presumably eager to discover other things. In ongoing work we installed three water ports (visible in the videos accompanying this article) and implemented a rule that activates the three ports in a cyclic sequence. Mice discovered all three ports rapidly and learned to visit them in the correct order. Future experiments will have to raise the bar on what the mice are expected to learn in a night.

Mechanisms of navigation

How do the animals navigate when they perform direct paths to the water port or to the exit? The present study cannot resolve that, but one can gain some clues based on observations so far. Early workers already concluded that rodents in a maze will use whatever sensory cues and tricks are available to accomplish their tasks [23]. Our maze was designed to restrict those options somewhat.

To limit the opportunity for visual navigation, the floor and walls of the maze are visually opaque. The ceiling is transparent, but the room is kept dark except for infrared illuminators. Even if the animal finds enough light, the goals (water port or exit) are invisible within the maze except from the immediately adjacent corridor.

There are no visible beacons that would identify the goal.

With regard to the sense of touch and kinesthetics, the maze was constructed for maximal symmetry. At each level of the binary tree all the junctions have locally identical geometry, with intersecting corridors of the same length. In practice the animals may well detect some inadvertent cues, like an unusual drop of glue, that could identify one node from another. The maze rotation experiment suggests that such cues are not essential for the animal's sense of location in the maze, at least in the expert phase.

The role of odors deserves particular attention because the mouse may use them both passively and actively. Does the animal first find the water port by following the smell of water? Probably not. For one, the port only emits a single drop of water when triggered by a nose poke. Second, we observed many instances where the animal is in the final corridor adjacent to the water port yet fails to discover it. The initial discovery seems to occur via touch. The reader can verify this in the videos accompanying this article. Regarding active use of odor markings in the maze, the maze rotation experiment suggests that such cues are not required for navigation, at least once the animals have adopted the shortest path to the water port (Figure 7.8).

Another algorithm that is often invoked for animals moving in an open arena is vector-based navigation [43]. Once the animal discovers a target, it keeps track of that target's heading and distance using a path integrator. When it needs to return to the target it follows the heading vector and updates heading and distance until it arrives. Such a strategy has limited appeal inside a labyrinth because the vectors are constantly blocked by walls. Consider, for example, the "home runs" back to the exit at the end of a bout. Here the target, namely the exit, is known from the start of the bout, because the animal enters through the same hole. At the end of the bout, when the mouse decides to exit from the maze, can it follow the heading vector to the exit? Figure 7.14A shows the 13 locations from which mice returned in a direct path to the exit on their very first foray. None of these locations is compatible with heading-based navigation: In each case an animal following the heading to the exit would get stuck in a different end node first and would have to reverse from there, quite unlike what really happened.

Finally, a partial clue comes from errors the animals make. We found that the rotation image of the water port, an end node diametrically across the entire maze, is one of the most popular destinations for rewarded animals (Figure 7.21A). These errors would be highly unexpected if the animals navigated from the entrance to the water

by odor markings, or if they used an absolute representation of heading and distance. On the other hand, if the animal navigates via a remembered sequence of turns, then it will end up at that image node if it makes a single mistake at just the first T-junction.

Future directed experiments will serve to narrow down how mice learn to navigate this environment, and how their policy might change over time. Since the animals get to perfection within an hour or so, one can test a new hypothesis quite efficiently. Understanding what mechanisms they use will then inform thinking about the algorithm for learning, and about the neuronal mechanisms that implement it.

7.4 Methods and Materials

Experimental design

The goal of the study was to observe mice as they explored a complex environment for the first time, with little or no human interference and no specific instructions. In preliminary experiments we tested several labyrinth designs and water reward schedules. Eventually we settled on the protocol described here, and tested 20 mice in rapid succession. Each mouse was observed only over a 7-hour period during the first night it encountered the labyrinth.

Maze construction

The maze measured ~24 x 24 x 2 inches; for manufacture we used materials specified in inches, so dimensions are quoted in those non-SI units where appropriate. The ceiling was made of 0.5-inch clear acrylic. Slots of 1/8-inch width were cut into this plate on a 1.5-inch grid. Pegged walls made of 1/8-inch infrared-transmitting acrylic (opaque in the visible spectrum, ePlastics) were inserted into these slots and secured with a small amount of hot glue. The floor was a sheet of infrared-transmitting acrylic, supported by a thicker sheet of clear acrylic. The resulting corridors (1–1/8-inches wide) formed a 6-level binary tree with T-junctions and progressive shortening of each branch, ranging from ~12-inch to 1.5-inch (Figure 7.1 and Figure 7.5). A single end node contained a 1.5-cm circular opening with a water delivery port (described below). The maze included provision for two additional water ports not used in the present report. Once per week the maze was submerged in cage cleaning solution. Between different animals the floor and walls were cleaned with ethanol.

Reward delivery system

The water reward port was controlled by a Matlab script on the main computer through an interface (Sanworks Bpod State Machine r1). Rewards were triggered when the animal's nose broke the IR beam in the water port (Sanworks Port interface + valve). The interface briefly opened the water valve to deliver ~30 µL of water and flashed an infrared LED mounted outside the maze for 1 s. This served to mark reward events on the video recording. Following each reward, the system entered a time-out period for 90 s, during which the port did not provide further reward. In experiments with sated mice the water port was turned off.

Cage and connecting passage

The entrance to the maze was connected to an otherwise normal mouse cage by red plastic tubing (3 cm diameter, 1 m long). The cage contained food, bedding, nesting material, and in the case of unrewarded experiments, also a normal water bottle.

Animals and treatments

All mice were C57BL/6J animals (Jackson Labs) between the ages of 45 and 98 days (mean 62 days). Both sexes were used: 4 males and 6 females in the rewarded experiments, 5 males and 4 females in the unrewarded experiments. For water deprivation, the animal was transferred from its home cage (generally group-housed) to the maze cage ~22 h before the start of the experiment. Non-deprived animals were transferred minutes before the start. All procedures were performed in accordance with institutional guidelines and approved by the Caltech IACUC.

Video recording

All data reported here were collected over the course of 7 hours during the dark portion of the animal's light cycle. Video recording was initiated a few seconds prior to connecting the tunnel to the maze. Videos were recorded by an OpenCV python script controlling a single webcam (Logitech C920) located ~1 m below the floor of the maze. The maze and access tube were illuminated by multiple infrared LED arrays (center wavelength 850 nm). Three of these lights illuminated the maze from below at a 45-degree angle, producing contrast to resolve the animal's foot pads. The remaining lights pointed at the ceiling of the room to produce backlight for a sharp outline of the animal.

Animal tracking

A version of DeepLabCut [25] modified to support gray-scale processing was used to track the animal's trajectory, using key points at the nose, feet, tail base, and mid-body. All subsequent analysis was based on the trajectory of the animal's nose, consisting of positions $x(t)$ and $y(t)$ in every video frame.

Rates of transition between cage and maze

This section relates to Figure 7.1–Figure 7.4. We entertained the hypothesis that the animals become "thirsty for exploration" as they spend more time in the cage. In that case one would predict that the probability of entering the maze in the next second will increase with time spent in the cage. One can compute this probability from the

distribution of residency times in the cage, as follows:

Say $t = 0$ when the animal enters the cage. The probability density that the animal will next leave the cage at time t is

$$p(t) = e^{-\int_0^t r(t')dt'} r(t) \tag{7.2}$$

where $r(t)$ is the instantaneous rate for entering the maze. So

$$\int_0^t p(t')\,dt' = 1 - e^{-\int_0^t r(t')dt'} \tag{7.3}$$

$$\int_0^t r(t')\,dt' = -\ln\left(1 - \int_0^t p(t')\,dt'\right) \tag{7.4}$$

This relates the cumulative of the instantaneous rate function to the cumulative of the observed transition times. In this way we computed the rates

$$r_m(t) = \text{rate of entry into the maze as a function of time spent in the cage} \tag{7.5}$$

$$r_c(t) = \text{rate of entry into the cage as a function of time spent in the maze} \tag{7.6}$$

The rate of entering the maze is highest at short times in the cage (Figure 7.1–Figure 7.4A). It peaks after ~15 s in the cage and then declines gradually by a factor of 4 over the first minute. So the mouse is most likely to enter the maze just after it returns from there. This runs opposite to the expectation from a homeostatic drive for exploration, which should be sated right after the animal returns. We found no evidence for an increase in the rate at late times. These effects were very similar in rewarded and unrewarded groups and in fact the tendency to return early was seen in every animal.

By contrast the rate of exiting the maze is almost perfectly constant over time (Figure 7.1–Figure 7.4B). In other words the exit from the maze appears like a constant rate Poisson process. There is a slight elevation of the rate at short times among rewarded animals (Figure 7.1–Figure 7.4B top). This may come from the occasional brief water runs they perform. Another strange deviation is an unusual number of very

short bouts (duration 2-12 s) among unrewarded animals (Figure 7.1–Figure 7.4B bottom). These are brief excursions in which the animal runs to the central junction, turns around, and runs to the exit. Several animals exhibited these, often several bouts in a row, and at all times of the night.

Reduced trajectories

From the raw nose trajectory we computed two reduced versions. First we divided the maze into discrete "cells," namely the squares the width of a corridor that make up the grid of the maze. At any given time the nose is in one of these cells and that time series defines the **cell trajectory**.

At a coarser level still one can ask when the animal passes through the nodes of the binary tree, which are the decision points in the maze. The special cells that correspond to the nodes of the tree are those at the center of a T-junction and those at the leaves of the tree. We marked all the times when the trajectory $(x(t), y(t))$ entered a new node cell. If the animal leaves a node cell and returns to it before entering a different node cell, that is not considered a new node. This procedure defines a discrete **node sequence** s_i and corresponding arrival times at those nodes t_i. We call the transition between two nodes a "step." Much of the analysis in this paper is derived from the animal's node sequence. The median mouse performed 16,192 steps in the 7 h period of observation (mean = 15,257; SD = 3,340).

In Figure 7.11 and Figure 7.14 we count the occurrence of **direct paths** leading to the water port (a "water run") or to the exit (a "home run"). A direct path is a node sequence without any reversals. Figure 7.6–Figure 7.7 illustrates some examples.

If the animal makes one wrong step from the direct path, that step needs to be backtracked, adding a total of two steps to the length of the path. If further errors occur during backtracking they need to be corrected as well. The binary maze contains no loops, so the number of errors is directly related to the length of the path:

$$\text{Errors} = (\text{Length of path} - \text{Length of direct path})/2. \tag{7.7}$$

Maze rotation

The maze rotation experiment (Figure 7.8) was performed on 4 mice, all water-deprived. Two of the animals ("D7" and "D9") had experienced the maze before, and are part of the "rewarded" group in other sections of the report. Two additional animals ("F2" and "A1") had had no prior contact with the maze.

The maze rotation occurred after at least 6 hours of exposure, by which time the animals had all perfected the direct path to the water port.

For animals "D7" and "D9" we rotated only the floor of the maze, leaving the walls and ceiling in the original configuration. For "F2" and "A1" we rotated the entire maze, moving one wall segment at the central junction and the water port to attain the same shape. Navigation remained intact for all animals. Note that "A1" performed a perfect path to the water port and back immediately before and after a full maze rotation (Figure 7.8B).

The visits to the 4 locations in the maze (Figure 7.8C, Figure 7.8–Figure 7.9) were limited to direct paths of length at least 2 steps. This avoids counting rapid flickers between two adjacent nodes. In other words, the animal has to move at least 2 steps away from the target node before another visit qualifies.

Statistics of sudden insight

In Figure 7.11 one can distinguish two events: first the animal finds the water port and begins to collect rewards at a steady rate, this is when the green curve rises up. At a later time the long direct paths to the water port become much more frequent than to the comparable control nodes: this is when the red and blue curves diverge. For almost all animals these two events are well separated in time (Figure 7.11–Figure 7.12). In many cases the rate of long paths seems to change discontinuously: a sudden change in slope of the curve.

Here we analyze the degree of "sudden change," namely how rapidly the rate changes in a time series of events. We modeled the rate as a sigmoid function of time during the experiment:

$$r\,(t) = r_{\mathrm{i}} + \frac{r_{\mathrm{f}} - r_{\mathrm{i}}}{2}\,\mathrm{erf}\left(\frac{t - t_{\mathrm{s}}}{w}\right) \tag{7.8}$$

where

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}\mathrm{d}x$$

The rate begins at a low initial level r_{i}, reflecting chance occurrence of the event, and saturates at a high final level r_{f}, limited for example by the animal's walking speed. The other two parameters are the time t_{s} of half-maximal rate change, and the width

w over which that rate change takes place. A sudden change in the event rate would correspond to $w = 0$.

The data are a set of n event times t_i in the observation interval $[0, T]$. We model the event train as an inhomogeneous Poisson point process with instantaneous rate $r(t)$. The likelihood of the data given the rate function $r(t)$ is

$$L\left[r\left(t\right)\right] = e^{-\int_0^T r(t)dt} \prod_i r\left(t_i\right) \tag{7.9}$$

and the log likelihood is

$$\ln L = \sum_i \ln r\left(t_i\right) - \int_0^T r\left(t\right) dt \tag{7.10}$$

For each of the 10 rewarded mice, we maximized $\ln L$ over the 4 parameters of the rate model, both for the reward events and the long paths to water. The resulting fits are plotted in Figure 7.11–Figure 7.12.

Focusing on the learning of long paths to water, for 6 of the 10 animals the optimal width parameter w was less than 300 s: B1, B2, C1, C3, C6, C7. These are the same animals one would credit with a sudden kink in the cumulative event count based on visual inspection (Figure 7.11–Figure 7.12).

To measure the uncertainty in the timing of this step, we refit the data for this subgroup of mice with a model involving a sudden step in the rate,

$$r\left(t\right) = \begin{cases} r_{\mathrm{i}}, t < t_{\mathrm{s}} \\ r_f, t > t_{\mathrm{s}} \end{cases} \tag{7.11}$$

and computed the likelihood of the data as a function of the step time t_{s}. We report the mean and standard deviation of the step time over its likelihood in Figure 7.11–Figure 7.13. Animal C6 was dropped from this "sudden step" group, because the uncertainty in the step time was too large (~900 s).

Efficiency of exploration

The goal of this analysis is to measure how effectively the animal surveys all the end nodes of the maze. The specific question is: in a string of n end nodes that the

animal samples, how many of these are distinct? On average how does the number of distinct nodes d increase with n? This was calculated as follows:

We restricted the animal's node trajectory (s_i) to clips of exploration mode, excluding the direct paths to the water port or the exit. All subsequent steps were applied to these clips, then averaged over clips. Within each clip we marked the sequence of end nodes (e_i). We slid a window of size n across this sequence and counted the number of distinct nodes d in each window. Then we averaged d over all windows in all clips. Then we repeated that for a wide range of n. The resulting $d(n)$ is plotted in the figures reporting new nodes vs nodes visited (Figure 7.17A,B and Figure 7.19C).

For a summary analysis we fitted the curves of $d(n)$ with a 2-parameter function:

$$d(n) \approx 64 \left(1 - \frac{1}{1 + \frac{z+bz^3}{1+b}}\right) \tag{7.12}$$

where

$$z = n\,/a\ . \tag{7.13}$$

The parameter a is the number of visits n required to survey half of the end nodes, whereas b reflects a relative acceleration in discovering the last few end nodes. This function was found by trial and error and produces absurdly good fits to the data (Figure 7.17–Figure 7.18). The values quoted in the text for efficiency of exploration are $E = 32\,/a$ (Equation 7.1).

The value of b was generally small (~0.1) with no difference between rewarded and unrewarded animals. It declined slightly over the night (Figure 7.17–Figure 7.18B), along with the decline in a (Figure 7.17C).

Biased random walk

For the analysis of Figure 7.19 we considered only the parts of the trajectory during "exploration" mode. Then we parsed every step between two nodes in terms of the type of action it represents. Note that every link between nodes in the maze is either a "left branch" or a "right branch", depending on its relationship to the parent T-junction. Therefore there are 4 kinds of action:

- $a = 0$: "in left", take a left branch into the maze
- $a = 1$: "in right", take a right branch into the maze

- $a = 2$: "out left", take a left branch out of the maze
- $a = 3$: "out right", take a right branch out of the maze

At any given node some actions are not available, for example from an end node one can only take one of the 'out' actions.

To compute the turning biases we considered every T-junction along the trajectory and correlated the action a_0 that led into that node with the subsequent action a_1. By tallying the action pairs (a_0, a_1) we computed the conditional probabilities $p(a_1|a_0)$. Then the 4 biases are defined as

$$P_{\mathrm{SF}} = \frac{p\,(0\,|0) + p\,(0\,|1) + p\,(1\,|0) + p\,(1\,|1)}{p\,(0\,|0) + p\,(0\,|1) + p\,(1\,|0) + p\,(1\,|1) + p\,(2\,|0) + p\,(3\,|1)} \tag{7.14}$$

$$P_{\mathrm{SA}} = \frac{p\,(0\,|1) + p\,(1\,|0)}{p\,(0\,|0) + p\,(0\,|1) + p\,(1\,|0) + p\,(1\,|1)} \tag{7.15}$$

$$P_{\mathrm{BF}} = \frac{p\,(0\,|3) + p\,(1\,|2) + p\,(2\,|2) + p\,(2\,|3) + p\,(3\,|2) + p\,(3\,|3)}{p\,(0\,|3) + p\,(1\,|2) + p\,(2\,|2) + p\,(2\,|3) + p\,(3\,|2) + p\,(3\,|3) + p\,(0\,|2) + p\,(1\,|3)} \tag{7.16}$$

$$P_{\mathrm{BS}} = \frac{p\,(2\,|2) + p\,(2\,|3) + p\,(3\,|2) + p\,(3\,|3)}{p\,(0\,|3) + p\,(1\,|2) + p\,(2\,|2) + p\,(2\,|3) + p\,(3\,|2) + p\,(3\,|3)} \tag{7.17}$$

For the simulations of random agents (Figure 7.17, Figure 7.19) we used trajectories long enough so the uncertainty in the resulting curves was smaller than the line width.

Models of decisions during exploration

The general approach is to develop a model that assigns probabilities to the animal's next action, namely which node it will move to next, based on its recent history of actions. All the analysis was restricted to the animal's "exploration" mode and to the 63 nodes in the maze that are T-junctions. During the "drink" and "leave" modes the animal's next action is predictable. Similarly when it finds itself at one of the 64 end nodes it only has one action available.

For every mouse trajectory we split the data into 5 segments, trained the model on 80% of the data, and tested it on 20%, averaging the resulting cross-entropy over the 5 possible splits. Each segment was in turn composed of parts of the trajectory sampled evenly throughout the 7-h experiment, so as to average over the small changes in the course of the night. The model was evaluated by the cross-entropy between the predictions and the animal's true actions. If one had an optimal model of behavior, the result would reveal the animal's true source entropy.

Fixed-depth Markov chain

To fit a model with fixed-history depth k to a measured node sequence (s_t), we evaluated all the substrings in that sequence of length $(k+1)$. At any given time t, the k-string $\mathbf{h}_t = (s_{t-k+1}, \ldots, s_t)$ identifies the history of the animal's k most recent locations. The current state s_t is one of 63 T-junctions. Each state is preceded by one of 3 possible states. So the number of history strings is $63 \cdot 3^{k-1}$. The 2-string (s_t, s_{t+1}) identifies the next action a_{t+1}, which can be "in left", "in right", or "out", corresponding to the 3 branches of the T junction. Tallying the history strings with the resulting actions leads to a contingency table of size $63 \cdot 3^{k-1} \times 3$, containing

$$n(\mathbf{h}, a) = \text{number of times history } \mathbf{h} \text{ leads to action } a \tag{7.18}$$

Based on these sample counts we estimated the probability of each action a conditional on the history $\mathbf{h}$ as

$$p\left(a \,|\mathbf{h}\right) = \frac{n\left(\mathbf{h}, a\right) + 1}{\sum\limits_{a'} n\left(\mathbf{h}, a'\right) + 3} \tag{7.19}$$

This amounts to additive smoothing with a pseudocount of 1, also known as "Laplace smoothing." These conditional probabilities were then used in the testing phase to predict the action at time t based on the preceding history $\mathbf{h}_t$. The match to the actually observed actions a_t was measured by the cross-entropy

$$H = \left\langle -\log_2 p\left(a_t \,|\mathbf{h}_t\right)\right\rangle_t \tag{7.20}$$

Variable-depth Markov chain

As one pushes to longer histories, i.e. larger k, the analysis quickly becomes data-limited, because the number of possible histories grows exponentially with k. Soon one finds that the counts for each history-action combination drop to where one can no longer estimate probabilities correctly. In an attempt to offset this problem we pruned the history tree such that each surviving branch had more than some minimal number of counts in the training data. As expected, this model is less prone to over-fitting and degrades more gently as one extends to longer histories (Figure 7.22–Figure 7.23A). The lowest cross-entropy was obtained with an average history length of ~4.0 but including some paths of up to length 6. Of all the algorithms we

tested, this produced the lowest cross-entropies, although the gains relative to the fixed-depth model were modest (Figure 7.22–Figure 7.23C).

Pooling across symmetric nodes in the maze

Another attempt to increase the counts for each history involved pooling counts over multiple T-junctions in the maze that are closely related by symmetry. For example, all the T-junctions at the same level of the binary tree look locally similar, in that they all have corridors of identical length leading from the junction. If one supposes that the animal acts the same way at each of those junctions, one would be justified in pooling across these nodes, leading to a better estimate of the action probabilities, and perhaps less over-fitting. This particular procedure was unsuccessful, in that it produced higher cross-entropy than without pooling.

However, one may want to distinguish two types of junctions within a given level: L-nodes are reached by a left branch from their parent junction one level lower in the tree, R-nodes by a right branch. For example, in Figure 7.6–Figure 7.7, node 1 is L-type and node 2 is R-type. When we pooled histories over all the L-nodes at a given level and separately over all the R-nodes, the cross-entropy indeed dropped, by about 5% on average. This pooling greatly reduced the amount of over-fitting (Figure 7.22–Figure 7.23B), which allowed the use of longer histories, which in turn improved the predictions on test data. The benefit of distinguishing L- and R-nodes probably relates to the animal's tendency to alternate left and right turns.

All the Markov model results we report are obtained using pooling over L-nodes and R-nodes at each maze level.

Data availability

All data and code needed to reproduce the figures and quoted results are available in this public repository: https://github.com/markusmeister/Rosenberg-2021-Repository.

References

[1] Alejandra Alonso, Jacqueline van der Meij, Dorothy Tse, and Lisa Genzel. "Naïve to Expert: Considering the Role of Previous Knowledge in Memory". en. In: *Brain and Neuroscience Advances* 4 (Jan. 2020), pp. 1–17. ISSN: 2398-2128. DOI: 10.1177/2398212820948686.

[2] Timothy E. J. Behrens, Timothy H. Muller, James C. R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior". eng. In: *Neuron* 100.2 (Oct. 2018), pp. 490–509. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2018.10.002.

[3] D. E. Berlyne. *Conflict, Arousal, and Curiosity*. New York, NY, US: McGraw-Hill Book Company, 1960, pp. xii, 350. DOI: 10.1037/11164-000.

[4] M. E. Bitterman, R. Menzel, Andrea Fietz, and Sabine Schäfer. "Classical Conditioning of Proboscis Extension in Honeybees (Apis Mellifera)". In: *Journal of Comparative Psychology* 97.2 (1983), pp. 107–119. ISSN: 1939-2087(Electronic),0735-7036(Print). DOI: 10.1037/0735-7036.97.2.107.

[5] R. Bourtchuladze, B. Frenguelli, J. Blendy, D. Cioffi, G. Schutz, and A. J. Silva. "Deficient Long-Term Memory in Mice with a Targeted Mutation of the cAMP-Responsive Element-Binding Protein". eng. In: *Cell* 79.1 (Oct. 1994), pp. 59–68. ISSN: 0092-8674. DOI: 10.1016/0092-8674(94)90400-6.

[6] P. A. Brennan and E. B. Keverne. "Neural Mechanisms of Mammalian Olfactory Learning". eng. In: *Progress in Neurobiology* 51.4 (Mar. 1997), pp. 457–481. ISSN: 0301-0082. DOI: 10.1016/s0301-0082(96)00069-x.

[7] H. M. Bruce. "An Exteroceptive Block to Pregnancy in the Mouse". eng. In: *Nature* 184 (July 1959), p. 105. ISSN: 0028-0836. DOI: 10.1038/184105a0.

[8] J. Buel. "The Linear Maze. I. "Choice-Point Expectancy," "Correctness," and the Goal Gradient". In: *Journal of Comparative Psychology* 17.2 (1934), pp. 185–199. ISSN: 0093-4127(Print). DOI: 10.1037/h0072346.

[9] Christopher P. Burgess, Armin Lak, Nicholas A. Steinmetz, Peter Zatka-Haas, Charu Bai Reddy, Elina A. K. Jacobs, Jennifer F. Linden, Joseph J. Paton, Adam Ranson, Sylvia Schröder, Sofia Soares, Miles J. Wells, Lauren E. Wool, Kenneth D. Harris, and Matteo Carandini. "High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice". In: *Cell Reports* 20.10 (Sept. 2017), pp. 2513–2524. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.08.047.

[10] M. Carandini and Anne K Churchland. "Probing Perceptual Decisions in Rodents." In: *Nature Neuroscience* 16 (July 2013), pp. 824–31. DOI: 10.1038/nn.3410.

[11] Thomas A. Cleland, Venkata Anupama Narla, and Karim Boudadi. “Multiple Learning Parameters Differentially Regulate Olfactory Generalization”. In: *Behavioral Neuroscience* 123.1 (Feb. 2009), pp. 26–35. ISSN: 0735-7044. DOI: 10.1037/a0013991.

[12] WK Estes. “The Problem of Inference from Curves Based on Group Data”. In: *Psychological Bulletin* 53.2 (1956), pp. 134–140. ISSN: 0033-2909. DOI: 10.1037/h0045156.

[13] Michael Fanselow and Robert Bolles. “Naloxone and Shock-Elicited Freezing in the Rat”. In: *Journal of comparative and physiological psychology* 93 (Sept. 1979), pp. 736–44. DOI: 10.1037/h0077609.

[14] Ehud Fonio, Yoav Benjamini, and Ilan Golani. “Freedom of Movement and the Stability of Its Unfolding in Free Exploration of Mice”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.50 (Dec. 2009), pp. 21335–21340. ISSN: 0027-8424. DOI: 10.1073/pnas.0812513106.

[15] C. R. Gallistel, S. Fairhurst, and P. Balsam. “The Learning Curve: Implications of a Quantitative Analysis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.36 (Sept. 2004), pp. 13124–13131. ISSN: 0027-8424. DOI: 10.1073/pnas.0404965101.

[16] Marie-Claude Grobéty and Françoise Schenk. “Spatial Learning in a Three-Dimensional Maze”. en. In: *Animal Behaviour* 43.6 (June 1992), pp. 1011–1020. ISSN: 0003-3472. DOI: 10.1016/S0003-3472(06)80014-X.

[17] Zengcai V. Guo, Nuo Li, Daniel Huber, Eran Ophir, Diego Gutnisky, Jonathan T. Ting, Guoping Feng, and Karel Svoboda. “Flow of Cortical Activity Underlying a Tactile Decision in Mice”. In: *Neuron* 81.1 (Jan. 2014), pp. 179–194. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2013.10.020.

[18] R. N. Hughes. “Intrinsic Exploration in Animals: Motives and Measurement”. In: *Behavioural Processes* 41.3 (Dec. 1997), pp. 213–226. ISSN: 0376-6357. DOI: 10.1016/S0376-6357(97)00055-7.

[19] I. Krechevsky. “"Hypotheses" in Rats”. In: *Psychological Review* 39.6 (1932), pp. 516–532. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0073500.

[20] J. E. LeDoux. “Emotion Circuits in the Brain”. eng. In: *Annual Review of Neuroscience* 23 (2000), pp. 155–184. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.23.1.155.

[21] Colin G. McNamara, Álvaro Tejero-Cantero, Stéphanie Trouche, Natalia Campo-Urriza, and David Dupret. “Dopaminergic Neurons Promote Hippocampal Reactivation and Spatial Memory Persistence”. en. In: *Nature Neuroscience* 17.12 (Dec. 2014), pp. 1658–1660. ISSN: 1546-1726. DOI: 10.1038/nn.3843.

[22] Norman L. Munn. “The Learning Process”. In: *Handbook of Psychological Research on the Rat; an Introduction to Animal Psychology*. Oxford, England: Houghton Mifflin, 1950, pp. 226–288.

[23] Norman L. Munn. “The Role of Sensory Processes in Maze Behavior”. In: *Handbook of Psychological Research on the Rat; an Introduction to Animal Psychology*. Oxford, England: Houghton Mifflin, 1950, pp. 181–225.

[24] Máté Nagy, Attila Horicsányi, Enikő Kubinyi, Iain D. Couzin, Gábor Vásárhelyi, Andrea Flack, and Tamás Vicsek. “Synergistic Benefits of Group Search in Rats”. eng. In: *Current Biology* (Sept. 2020). ISSN: 1879-0445. DOI: 10.1016/j.cub.2020.08.079.

[25] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. “Using DeepLabCut for 3D Markerless Pose Estimation across Species and Behaviors”. en. In: *Nature Protocols* 14.7 (July 2019), pp. 2152–2176. ISSN: 1750-2799. DOI: 10.1038/s41596-019-0176-0.

[26] W. T. Newsome and E. B. Pare. “A Selective Impairment of Motion Perception Following Lesions of the Middle Temporal Visual Area (MT)”. en. In: *Journal of Neuroscience* 8.6 (June 1988), pp. 2201–2211. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.08-06-02201.1988.

[27] DS Olton. “Mazes, Maps, and Memory”. In: *American Psychologist* 34.7 (1979), pp. 583–596. ISSN: 0003-066X. DOI: 10.1037/0003-066X.34.7.583.

[28] Sashank Pisupati, Lital Chartarifsky-Lynn, Anup Khanal, and Anne K Churchland. “Lapses in Perceptual Decisions Reflect Exploration”. In: *eLife* 10 (Jan. 2021). Ed. by Daeyeol Lee, Joshua I Gold, Long Ding, and Alex C Kwan, e55490. ISSN: 2050-084X. DOI: 10.7554/eLife.55490.

[29] Pseudo-Apollodorus. “Epitome”. In: *Library and Epitome*. I-II Century AD, Ch 1 Sec 9.

[30] Laure Rondi-Reig, Géraldine H. Petit, Christine Tobin, Susumu Tonegawa, Jean Mariani, and Alain Berthoz. “Impaired Sequential Egocentric and Allocentric Memories in Forebrain-Specific-NMDA Receptor Knock-out Mice during a New Task Dissociating Strategies of Navigation”. eng. In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 26.15 (Apr. 2006), pp. 4071–4081. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.3408-05.2006.

[31] A. E. Rosser and E. B. Keverne. “The Importance of Central Noradrenergic Neurones in the Formation of an Olfactory Memory in the Prevention of Pregnancy Block”. eng. In: *Neuroscience* 15.4 (Aug. 1985), pp. 1141–1147. ISSN: 0306-4522. DOI: 10.1016/0306-4522(85)90258-1.

[32] Nobuya Sato, Chihiro Fujishita, and Atsuhito Yamagishi. “To Take or Not to Take the Shortcut: Flexible Spatial Behaviour of Rats Based on Cognitive Map in a Lattice Maze”. eng. In: *Behavioural Processes* 151 (June 2018), pp. 39–43. ISSN: 1872-8308. DOI: 10.1016/j.beproc.2018.03.010.

[33] Julian Seward. *Bzip2*. July 2019.

[34] Takaharu Shokaku, Toru Moriyama, Hisashi Murakami, Shuji Shinohara, Nobuhito Manome, and Kazuyuki Morioka. “Development of an Automatic Turntable-Type Multiple T-Maze Device and Observation of Pill Bug Behavior”. In: *Review of Scientific Instruments* 91.10 (Oct. 2020), p. 104104. ISSN: 0034-6748. DOI: 10.1063/5.0009531.

[35] Willard S. Small. “Experimental Study of the Mental Processes of the Rat. II”. In: *The American Journal of Psychology* 12.2 (1901), pp. 206–239. ISSN: 0002-9556. DOI: 10.2307/1412534.

[36] Ofer Tchernichovski, Yoav Benjamini, and Ilan Golani. “The Dynamics of Long-Term Exploration in the Rat”. en. In: *Biological Cybernetics* 78.6 (July 1998), pp. 423–432. ISSN: 1432-0770. DOI: 10.1007/s004220050446.

[37] E. C. Tolman. “Cognitive Maps in Rats and Men”. In: *Psychological Review* 55.4 (1948), pp. 189–208. ISSN: 0033-295X. DOI: 10.1037/h0061626.

[38] E. C. Tolman. “The Determiners of Behavior at a Choice Point”. In: *Psychological Review* 45 (1938), pp. 1–41. ISSN: 0033-295X. DOI: 10.1037/h0062733.

[39] EC Tolman and CH Honzik. “Degrees of Hunger, Reward and Non-Reward, and Maze Learning in Rats”. In: *University of California Publications in Psychology* 4 (1930), pp. 241–256.

[40] Naoshige Uchida and Zachary F. Mainen. “Speed and Accuracy of Olfactory Discrimination in the Rat”. eng. In: *Nature Neuroscience* 6.11 (Nov. 2003), pp. 1224–1229. ISSN: 1097-6256. DOI: 10.1038/nn1142.

[41] H. J. Uster, K. Bättig, and H. H. Nägeli. “Effects of Maze Geometry and Experience on Exploratory Behavior in the Rat”. en. In: *Animal Learning & Behavior* 4.1 (Mar. 1976), pp. 84–88. ISSN: 1532-5830. DOI: 10.3758/BF03211992.

[42] Jesse N. Weber, Brant K. Peterson, and Hopi E. Hoekstra. “Discrete Genetic Modules Are Responsible for Complex Burrow Evolution in Peromyscus Mice”. en. In: *Nature* 493.7432 (Jan. 2013), pp. 402–405. ISSN: 1476-4687. DOI: 10.1038/nature11816.

[43] R. Wehner, B. Michel, and P. Antonsen. “Visual Navigation in Insects: Coupling of Egocentric and Geocentric Information”. en. In: *Journal of Experimental Biology* 199.1 (Jan. 1996), pp. 129–140. ISSN: 0022-0949, 1477-9145.

[44] Ruth A. Wood, Marius Bauza, Julija Krupic, Stephen Burton, Andrea Delekate, Dennis Chan, and John O'Keefe. "The Honeycomb Maze Provides a Novel Test to Study Hippocampal-Dependent Spatial Navigation". en. In: *Nature* 554.7690 (Feb. 2018), pp. 102–105. ISSN: 1476-4687. DOI: 10.1038/nature25433.

[45] H. Woodrow. "The Problem of General Quantitative Laws in Psychology". In: *Psychological Bulletin* 39.1 (1942), pp. 1–27. ISSN: 1939-1455(Electronic),0033-2909(Print). DOI: 10.1037/h0058275.

[46] Ryan M. Yoder, Benjamin J. Clark, Joel E. Brown, Mignon V. Lamia, Stephane Valerio, Michael E. Shinder, and Jeffrey S. Taube. "Both Visual and Idiothetic Cues Contribute to Head Direction Cell Stability during Navigation along Complex Routes". In: *Journal of Neurophysiology* 105.6 (Mar. 2011), pp. 2989–3001. ISSN: 0022-3077. DOI: 10.1152/jn.01041.2010.

Part IV

Theory and Computation

Chapter 8

ENDOTAXIS: A UNIVERSAL ALGORITHM FOR MAPPING, GOAL-LEARNING, AND NAVIGATION

[1] Tony Zhang, Matthew Rosenberg, Pietro Perona, and Markus Meister. "Endotaxis: A Universal Algorithm for Mapping, Goal-Learning, and Navigation". In: *bioRxiv* (2021). URL: https://www.biorxiv.org/content/10.1101/2021.09.24.461751v1.

An animal entering a new environment typically faces three challenges: explore the space for resources, memorize their locations, and navigate towards those targets as needed. Experimental work on exploration, mapping, and navigation has mostly focused on simple environments—such as an open arena, a pond [32], or a featureless desert [36]—and much has been learned about neural signals in diverse brain areas under these conditions [12, 42]. However, many natural environments are highly constrained, such as a system of burrows, or of paths through the underbrush. More generally, many cognitive tasks are equally constrained, allowing only a small set of actions at a given time in the process. Here we propose an algorithm that learns the structure of an arbitrary environment, discovers useful targets during exploration, and navigates back to those targets by the shortest path. It makes use of a behavioral module common to all motile animals, namely the ability to follow an odor to its source [6]. We show how the brain can learn to generate internal "virtual odors" that guide the animal to any location of interest. The *endotaxis* algorithm can be implemented with a simple 3-layer neural circuit using only biologically realistic structures and learning rules. Several neural components of this scheme are found in brains from insects to humans. Nature may have evolved a general mechanism for search and navigation on the ancient backbone of chemotaxis.

8.1 Introduction

Efficient navigation requires knowing the structure of the environment: which locations are connected to which others [50]. One would like to understand how the brain acquires that knowledge, what neural representation it adopts for the resulting map, how it tags significant locations in that map, and how that knowledge gets read out for decision-making during navigation. Here we propose a mechanism that solves all these problems and operates reliably in diverse and complex environments.

One algorithm for finding a valuable resource is common to all animals: chemotaxis. Every motile species has a way to track odors through the environment, either to find the source of the odor or to avoid it [6]. This ability is central to finding food, connecting with a mate, and avoiding predators. It is believed that brains originally evolved to organize the motor response in pursuit of chemical stimuli. Indeed some of the oldest regions of the mammalian brain, including the hippocampus, seem organized around an axis that processes smells [2, 24].

The specifics of chemotaxis, namely the methods for finding an odor and tracking it, vary by species, but the toolkit always includes a random trial-and-error scheme: try various actions that you have available, then settle on the one that makes the odor stronger [6]. For example a rodent will weave its head side-to-side, sampling the local odor gradient, then move in the direction where the smell is stronger. Worms and maggots follow the same strategy. Dogs track a ground-borne odor trail by casting across it side-to-side. Flying insects perform similar casting flights. Bacteria randomly change direction every now and then, and continue straight as long as the odor improves [8]. We propose that this universal behavioral module for chemotaxis can be harnessed to solve general problems of search and navigation in a complex environment.

For concreteness, consider a mouse exploring a labyrinth of tunnels (Figure 8.1A). The maze may contain a source of food that emits an odor (Figure 8.1A top). That odor will be strongest at the source and decline with distance along the tunnels of the maze. The mouse can navigate to the food location by simply following the odor gradient uphill. Suppose that the mouse discovers some other interesting locations that do not emit a smell, like a source of water, or the exit from the labyrinth (Figure 8.1A). It would be convenient if the mouse could tag such a location with an odorous material, so it may be found easily on future occasions. Ideally the mouse would carry with it multiple such odor tags, so it can mark different targets each with its specific recognizable odor (Figure 8.1A mid and bottom).

Here we show that such tagging does not need to be physical. Instead we propose a mechanism by which the mouse's brain may compute a "virtual odor" signal that declines with distance from a chosen target. That neural signal can be made available to the chemotaxis module as though it were a real odor, enabling navigation up the gradient towards the target. Because this goal signal is computed in the brain rather than sensed externally, we call this hypothetical process *endotaxis*.

8.2 A Circuit to Implement Endotaxis

In Figure 8.1B we present a neural circuit model that implements three goals: mapping the connectivity of the environment; tagging of goal locations with a virtual odor; and navigation towards those goals. The model includes four types of neurons: feature cells, point cells, map cells, and goal cells.

Feature cells: These cells fire when the animal encounters an interesting feature that may form a target for future navigation. Each feature cell is selective for a specific kind of resource, for example water or food, by virtue of sensory pathways that respond to those stimuli.

Point cells: This layer of cells represents the animal's location.[1] Each neuron in this population has a small response field within the environment. The neuron fires when the animal enters that response field. We assume that these point cells exist from the outset as soon as the animal enters the environment. Each cell's response field is defined by some conjunction of external and internal sensory signals at that location.

Map cells: This layer of neurons learns the structure of the environment, namely how the various locations are connected in space. The map cells get excitatory input from point cells with low convergence: Each map cell should collect input from only one or a few point cells. These input synapses are static. The map cells also excite each other with all-to-all connections. These recurrent synapses are modifiable according to rules of Hebbian plasticity and, after learning, represent the topology of the environment.

Goal cells: These neurons mark the locations of special resources in the map of the environment. A goal cell for a specific feature receives excitatory input from the corresponding feature cell. It also receives Hebbian excitatory synapses from map cells. Those synapses are strengthened when the presynaptic map cell is active at the

[1] We avoid the term "place cell" here because (1) that term has a technical meaning in the rodent hippocampus, whereas the arguments here extend to species that don't have a hippocampus; (2) all the cells in this network have a place field, but it is smallest for the point cells.

same time as the feature cell.

Each of the goal cells carries a virtual odor signal for its assigned feature. That signal increases systematically as the animal moves closer to the target feature. A mode switch selects one among many possible virtual odors (or real odors) to be routed to the chemotaxis module for odor tracking. [2] The animal then pursues its chemotaxis search strategy to maximize that odor, which leads it to the selected tagged feature.

Why does the circuit work?

The key insight is that the output of the goal cell declines systematically with the distance of the animal from that target. This relationship holds even if the environment is a complex graph with constrained connectivity. Here we explain how this comes about, with mathematical details in the supplement.

As the animal explores a new environment, when it moves from one location to an adjacent one, those two point cells briefly fire together. That leads to a Hebbian strengthening of the excitatory synapses between the two corresponding map cells. In this way the recurrent network of map cells learns the connectivity of the graph that describes the environment. To a first approximation, the matrix of synaptic connections among the map cells will converge to the correlation matrix of their inputs [14, 19], which in turn reflects the adjacency matrix of the graph (Equation 8.22). Now the brain can use this adjacency information to find the shortest path to a target.

After this map learning, the output of the map network is a hump of activity, centered on the current location x of the animal and declining with distance along the various paths in the graph (Figure 8.1C top). If the animal moves to a different location y, the map output is another hump of activity, now centered on y (Figure 8.1C bottom). The overlap of the two hump-shaped profiles will be large if nodes x and y are close on the graph, and small if they are distant. Fundamentally the endotaxis network computes that overlap. How is it done?

Suppose the animal visits y and finds water there. Then the profile of map activity $v_i(y)$ gets stored in the synapses G_{gi} onto the goal cell g that responds to water (Figure 8.1B, Equation 8.26). When the animal subsequently moves to a different location x, the goal cell g receives the current map output $v_i(x)$ filtered through the previously stored synaptic template $v_i(y)$. This is the desired measure of overlap

[2] That mode switch is controlled by the *murinculus*: a tiny mouse inside the mouse that tells the mouse what to do. We do not claim to know how that works.

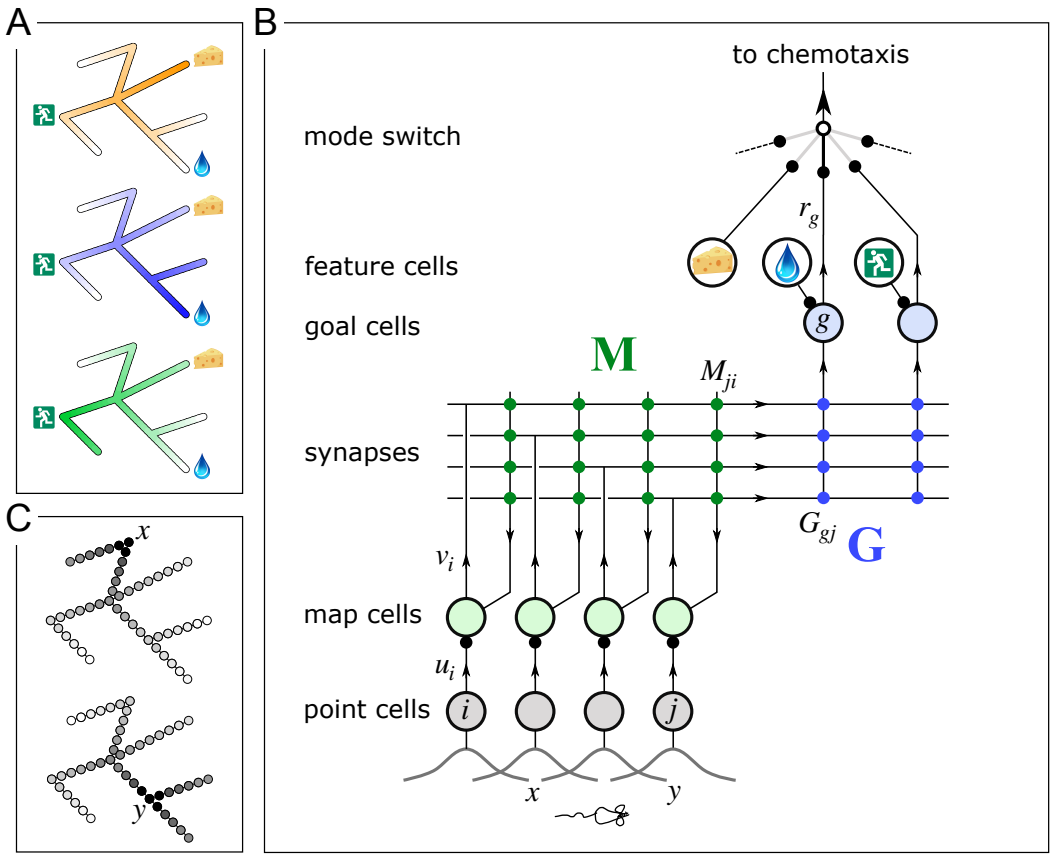

Figure 8.1: **A** mechanism for endotaxis. **A:** A constrained environment of nodes linked by straight corridors, with special locations offering food, water, and the exit. Top: A real odor emitted by the food source decreases with distance (shading). Middle: A virtual odor tagged to the water source. Bottom: A virtual odor tagged to the exit. **B:** A neural circuit to implement endotaxis. Open circles: four populations of neurons that represent "feature," "point," "map," and "goal." Arrows: Signal flow. Solid circles: Synapses. Point cells have small receptive fields localized in the environment and excite map cells. Map cells excite each other by recurrent Hebbian synapses and excite goal cells by another set of Hebbian synapses. A goal cell also receives sensory input from a feature cell indicating the presence of a resource, e.g. water or the exit. The feature cell for cheese responds to a real odor emitted by that target. A "mode" switch selects among various goal signals depending on the animal's need. They may be virtual odors (water, exit) or real odors (cheese). The resulting signal gets fed to the chemotaxis module for gradient ascent. Mathematical symbols used in the text: u_i is the output of a place cell at location i, v_i is the output of the corresponding map cell, $\mathbf{M}$ is the matrix of synaptic weights among map cells, $\mathbf{G}$ are the synaptic weights from the map cells onto goal cells, and r_g is the output of goal cell g. **C:** The output of map cells after the map has been learned; here the animal is located at points x (top) or y (bottom). Black means high activity. For illustration, each map cell is drawn at the center of its place field.

(Equation 8.27), and one can show mathematically that it declines exponentially with the shortest graph-distance between x and y (Equation 8.28).

8.3 Performance of the Endotaxis Algorithm

Some important features of endotaxis can already be appreciated at this level of detail. First, the structure of the environment is acquired separately from the location of resources. The graph that connects different points in the environment is learned by the synapses in the map network. By contrast the location of special goals within that map is learned by the synapses onto the goal cells. The animal can explore and learn the environment regardless of the presence of threats or resources. Once a resource is found, its location can be tagged immediately within the existing map structure. If the distribution of resources changes, the knowledge of the connectivity map remains unaffected. Second, the endotaxis algorithm is "always on." There is no separation of learning and recall into different phases. Both the map network and the goal network get updated continuously based on the animal's trajectory through the environment, and the goal signals are always available for directed navigation via gradient ascent.

Simultaneous acquisition of map and targets during exploration

To illustrate these functions, and to explore capabilities that are less obvious from an analytical inspection, we simulated agents navigating by the endotaxis algorithm (Figure 8.1B) through a range of environments (Figs 8.2-8.3). In each case we assumed that there are point cells that fire at specific locations, owing to a match of their sensory receptive fields with features in the environment. The locations of these point cells define the nodes of the graph that the agent will learn. Both the map synapses and the goal synapses start out tabula rasa with zero synaptic strengths. This is because the animal has no notion of the topology of the environment (which location connects with which other location), and no information on the location of the resources. As the agent explores the environment, for example by a random walk, map synapses get updated based on the simultaneous firing of point cells corresponding to neighboring locations. We used a standard formulation of Hebbian learning, called Oja's rule, which has only two parameters. Similarly the synapses onto goal cells get updated based on the presynaptic map cell and the postsynaptic signal from feature cells. Map cells and goal cells were allowed to learn at different rates (see Section 8.6 for detail).

A simple Gridworld environment (Figure 8.2) serves to observe the dynamics

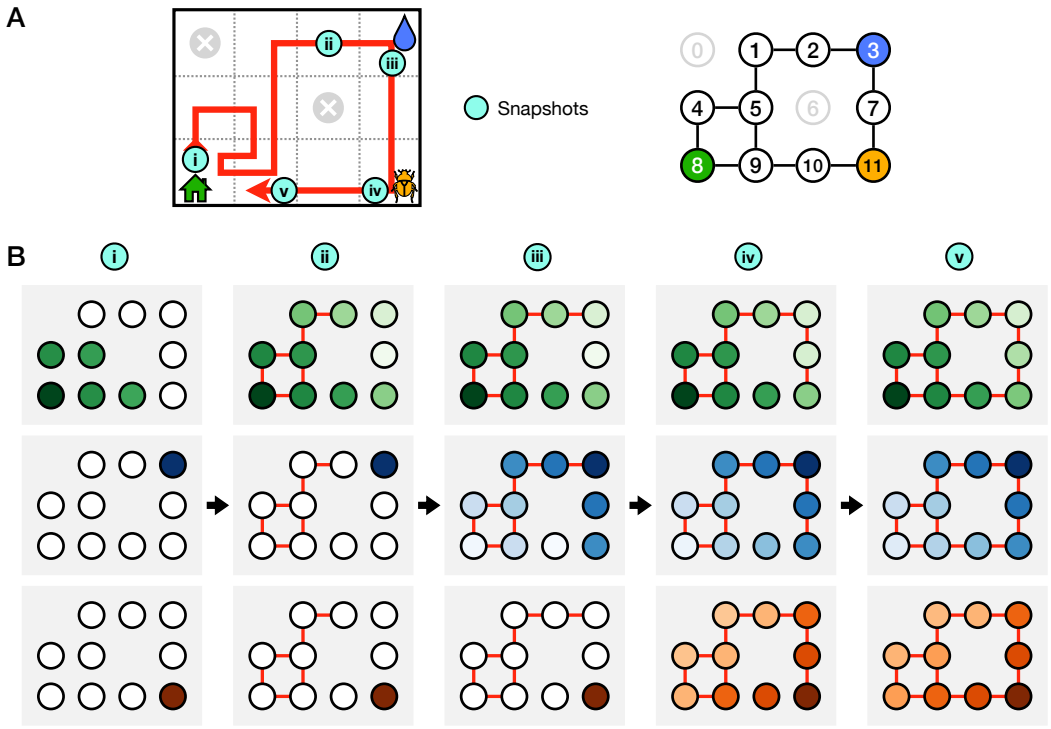

Figure 8.2: **The map and the targets are learned independently. (A)** Left: an agent explores a simple Gridworld with 3 salient goal locations following the red trajectory. Space is discretized into square tiles, each tile represented by one point cell. Circles with crosses represent obstacles, namely tiles that are not reachable. Right: graph of this environment, where each tile becomes a node, and edges represent traversable connections between tiles. **(B)** The response fields of three goal neurons for home (top), water (middle), and bug (bottom) at the 5 instants during the learning process (i-v). Red edges connect previously visited nodes. The response (log color scale) is plotted at each location where the agent could be placed. The agent starts random walking from the entrance (i) and gradually discovers the other two goal locations (water at time iii, bug at time iv). Upon discovery of a goal location, the corresponding goal cell's signal is immediately useful in all previously visited locations (iii, iv) as well as nodes that are ≤ 2 steps away. Any new locations visited subsequently and nodes ≤ 2 steps away are also recruited into the goal cell's response field (v).

of learning in detail. There are three locations of interest: the entrance to the environment, experienced at the very start of exploration; a water source; and a food item. When the agent first enters the novel space, a feature neuron that responds to the entrance excites a goal cell, which leads to the potentiation of synapses onto that neuron. Effectively that tags the entrance, and from now on that goal cell encodes a virtual "entrance odor" that declines with distance from the entrance. With every step the agent takes, the map network gets updated, and the range of the entrance odor spreads further. At all times the agent could decide to follow this virtual odor uphill to the entrance. The water source starts out invisible from anywhere except its special location. However, as soon as the agent reaches the water, the water goal cell gets integrated in the circuit through the potentiation of synapses from map cells. Because the map network is already established along the path that the agent took, that immediately creates a virtual "water odor" that spreads through the environment and declines with distance from the water location (Figure 8.2B-iii).

As the agent explores the environment further, the virtual odors spread accordingly to the new locations visited (Figure 8.2B i-v). After extensive exploration, the map and goal networks reach a steady state. Now the virtual odors are available at every point in the environment, and they decline monotonically with the shortest-path distance to the respective goal location (Figure 8.2B-v). As one might expect, an agent endotaxing uphill on this virtual odor always reaches the goal location, and does so by the shortest possible path (Figure 8.3Bi-Ci).

We performed a similar simulation for a complex labyrinth used in a recent study of mouse navigation [40]. The topology of the maze was a binary tree with a single entrance, 63 T-junctions, and 64 end nodes (Figure 8.3A-ii). A single source of water was located at one of the end nodes. In these experiments mice learned a direct path to the water source after visiting it ~10 times; they also performed error-free paths back to entrance on the first attempt [40]. Again the simulated agent explored the labyrinth with a random walk. The virtual entrance odor allowed it to navigate back to the entrance from any point along the trajectory. The first visit to the water port established a goal cell with virtual water odor. After exploration had covered the entire labyrinth, both the entrance odor and the water odor were available at every location (Figure 8.3B-ii), allowing for flawless navigation to the sources by endotaxis (Figure 8.3C-ii).

It turns out that endotaxis is a useful strategy beyond spatial navigation. For instance, the game "Towers of Hanoi" represents a more complex environment (Figure 8.3A-iii).

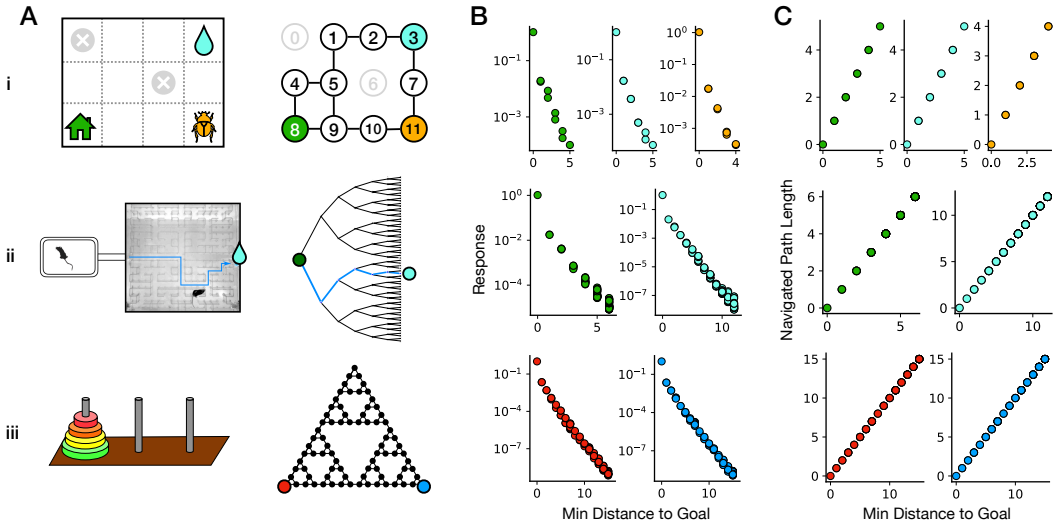

Figure 8.3: **E**ndotaxis can operate in environments with diverse topologies. **(A)** Three tasks and their corresponding graph representations: i) Gridworld of Figure 8.2 with 3 goal nodes (home, water, and food). ii) A binary tree labyrinth used in mouse navigation experiments [40], with 2 goals (home and water). iii) Tower of Hanoi game, with 2 goals (the configurations of disks that solve the game). **(B)** The virtual odors after extensive exploration. For each goal neuron the response at every node is plotted against the shortest graph distance from the node to the goal. **(C)** Navigation by endotaxis: For every starting node in the environment this plots the number of steps to the goal against the shortest distance.

Disks of different sizes are stacked on three pegs, with the constraint that no disk can rest on top a smaller one. The game is solved by rearranging the pile of disks from one peg to another. In any state of the game there are either 2 or 3 possible actions, and they form an interesting graph with many loops (Figure 8.3A-iii). Again the simulated agent explored this graph by random walking. Once it encountered a solution, that state was tagged with a virtual odor. After enough exploration the virtual odor signal was available from every possible game state, and the agent could solve the game from any starting state in the shortest number of moves. This example illustrates that endotaxis is a useful algorithm even for cognitive tasks that don't involve spatial movement. It merely requires the existence of neurons that recognize any given state of the game. To start with, the agent has no internal model of the game, so it must happen on the first solution by chance. However, when prompted to solve the problem again, the agent can use the learned virtual odor to complete the game in the fewest possible moves.

These simulations suggest that the endotaxis algorithm can function perfectly in environments of reasonable complexity, learning both the connectivity of the

environment and the location of multiple resources within that map. How robust is that performance? First, the model did not require careful tuning of parameters. Instead, we found that endotaxis works over several log units of the two parameters in Oja's rule for synaptic plasticity (Figure 8.6). It fails in a predictable fashion: for example, if the agent takes longer to explore the environment than the time constant for synaptic change, then the map is always partially forgotten, and navigation to a target will fail. Second, we considered the effects of noise in neural signals, and found a gradual failure when the signal-to-noise value exceeded 1 (Figure 8.8).

8.4 Adaptation to Change in the Environment

An attractive feature of the endotaxis algorithm is that it separates learning the map from learning the target locations. In many real-world environments the topology of the map (how are locations connected?) is probably more stable than the targets (which locations are interesting?). Separating the two allows the agent to adjust to changes on both fronts using different rules and time-scales. We illustrate an example of each.

Change in connectivity

Suppose that the connectivity of the environment changes. For example, a shortcut appears between two locations that used to be separated, or a blockage separates two previously adjacent locations (Figure 8.4A.i-ii). This alters the correlation in firing among the point cells during the agent's explorations, and over time that will reflect in the synapses of the map network. How will endotaxis adapt to such changes?

To explore these adjustments, we considered navigation on a ring-shaped maze with a single goal location (Figure 8.4A.i). Note that the ring is the simplest graph that offers two routes to a target, and we will evaluate whether the algorithm finds the shorter one. An agent explores the ring by stepping among locations in a random walk, and builds the map cell network from that experience. After a period of ~100 steps, navigation by endotaxis is perfect, in that the agent chooses the shorter route to the goal from every start node (Figure 8.4B-C.i). If the ring gets broken by removing one link, then endotaxis fails from some start nodes because it steers the agent towards the blocked path. Over time, the representation of the former link gets erased from the map network because the corresponding map synapses weaken whenever the link isn't used. Gradually, over several hundred random steps, navigation returns to perfect performance again (Figure 8.4B-C.i).

Conversely, when a new shortcut appears between previously separated locations

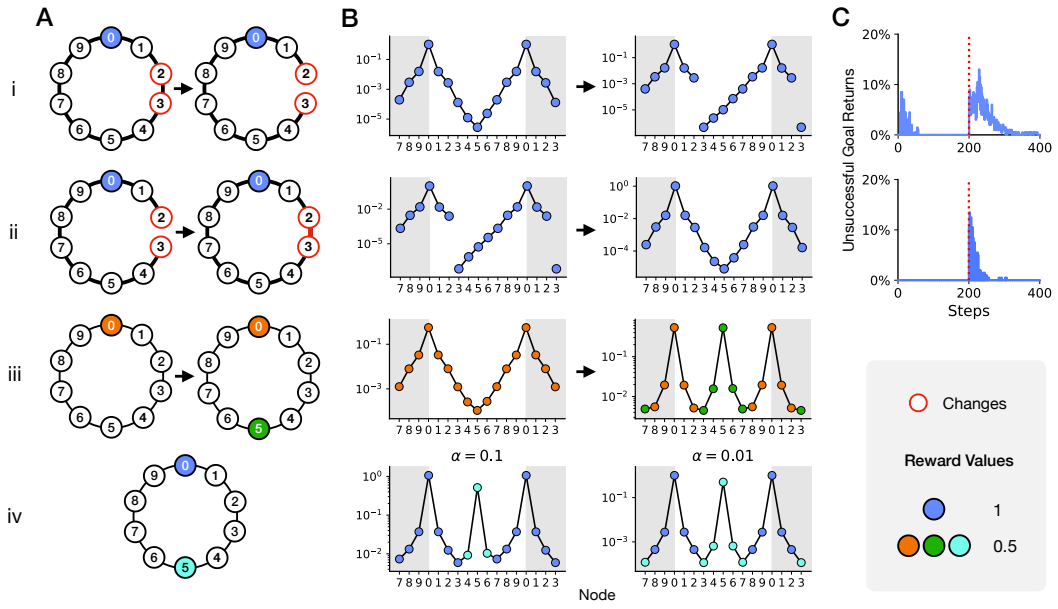

Figure 8.4: **E**ndotaxis adapts quickly to changes in the environment or the target locations. **(A)** A ring environment modified by sudden appearance of a blockage: (i), a shortcut (ii), an additional goal target (iii), or a dual-reward environment with different saliency (iv). Graphs shown before and after modification or constant. Shaded nodes are goal locations. Labels identify positions with point cells in the graph. **(B i-iii)** Response profile of the goal neuron after sufficient exploration, shown for timesteps just before modification (200) and at the end of exploration (400). Color of nodes indicates the goal location the agent will eventually reach by following the virtual odor starting from that node. Note the virtual odor peaks at either one or two target locations depending on the environment, with a higher amplitude at the stronger target. An agent following endotaxis will navigate to the stronger target from a wider domain of attraction. **(B iv)** Varying α in Oja's Rule for map learning adjusts the tradeoff between distance and reward. With a smaller α there is an equal number of starting nodes that reach node 0 and node 5. **(C)** Ability to navigate back to goal over 400 steps of random walk exploration, showing the fraction of successful returns to a goal from the current location at each timestep over 200 random walk explorations. Dotted line marks the time of modification. Note that navigation gets disrupted briefly, then it returns to perfect.

(Figure 8.4A.ii), a similar change takes place. For a brief period endotaxis is suboptimal, because the agent sometimes takes the long route even though a shorter one is available. However, that perturbation gets incorporated into the map much more quickly, after just a few tens of steps of exploration (Figure 8.4B-C.ii). One can understand the asymmetry as follows: As the agent explores the environment, a newly available link is confirmed with certainty the first time it gets traveled. By contrast the loss of a link remains uncertain until the agent has not taken that route many times.

Appearance of new targets

Suppose the agent has discovered one location with a water resource. Some time later water also appears at a second location (Figure 8.4A.iii). When the agent discovers that, the same water goal cell will get activated and therefore receive a potentiation of synapses active at that second location. Now the input network to that goal cell contains the sum of two templates, corresponding to the map outputs from the two target locations. As before, the current map output gets filtered through these synaptic weights to create the virtual odor. One might worry that this goal signal steers the agent to a location half-way between the two targets. Instead, simulations on the ring show that the virtual odor peaks at both targets, and endotaxis takes the agent reliably to the nearest one (Figure 8.4B.iii).

Choice between multiple targets

Suppose one of the targets offering the same resource is more valuable than the other, for example because it gives a larger reward (Figure 8.4A.iv). In the endotaxis model (Figure 8.1B), the larger reward causes higher activity of the feature cell that responds to this resource, and thus stronger potentiation of the synapses onto the associated goal cell (Equation 8.20). Thus the input template of the goal cell becomes a weighted sum of the map outputs from the two target locations, with greater weight for the location with higher reward (Figure 8.4B.iv). The virtual odor still shows two peaks, but the stronger target now has a greater region of attraction; for some starting locations the agent chooses the longer route in favor of the larger reward, a sensible behavior.

What determines the trade-off between the longer distance and the greater reward? In the endotaxis model (Figure 8.1B) this is set by α_{M}, one of the two parameters of the synaptic learning rule in the map network (Equation 8.19). A small α_{M} raises the cost of any additional step traveled and thus diminishes the importance of reward

differences (Figure 8.4B.iv right panel). By contrast a large α_M favors the larger reward regardless of distance traveled. One can show that the role of α_M is directly equivalent to the discount factor in reinforcement learning theory (Equation 8.28).

In summary, endotaxis adapts readily to changes in the environment or in the availability of rewards. Furthermore, it implements a rational choice between multiple targets of the same kind, using a variable weighting of reward versus distance. None of these features required any custom tuning: they all follow directly from the basic formulation in Figure 8.1B.

8.5 Discussion

Summary of claims

We have presented a neural mechanism that can support learning, navigation, and problem solving in complex and changing environments. It is based on chemotaxis, namely the ability to follow an odor signal to its source, which is shared universally by most or all motile animals. The algorithm, called endotaxis, is formulated as a neural network that creates an internal "virtual odor" which the animal can follow to reach any chosen target location (Figure 8.1). When the agent begins to explore the environment, the network learns both the structure of the space, namely how various points are connected, and the location of valuable resources (Figure 8.2). After sufficient exploration the agent can then navigate back to those target locations from any point in the environment (Figure 8.3). The algorithm is *always on* and it adapts flexibly to changes in the structure of the environment or in the locations of targets (Figure 8.4). Furthermore, even in its simplest form, endotaxis can arbitrate among multiple locations with the same resource, by trading off the promised reward against the distance traveled (Figure 8.4). Beyond spatial navigation, endotaxis can also learn the solution to purely cognitive tasks (Figure 8.3), or any problem defined by search on a graph. The neural network model that implements endotaxis has a close resemblance to known brain circuits. We propose that evolution may have built upon the ancient behavioral module for chemotaxis to enable much more general abilities for search and navigation, even in the absence of odor gradients. In the following sections we consider how these findings relate to some well-established phenomena and results on animal navigation.

Animal behavior

The millions of animal species no doubt use a wide range of mechanisms to get around their environment, and it is worth specifying which of those problems endotaxis might

solve. First, the learning mechanism proposed here applies to complex environments, namely those in which discrete paths form sparse connections between points. For a bird, this is less of a concern, because it can get from every point to any other "as the crow flies." For a rodent and many other terrestrial animals, on the other hand, the paths they may follow are constrained by obstacles and by the need to remain under cover. In those conditions the brain cannot assume that the distance between points is given by Euclidean geometry, or that beacons for a goal will be visible in a straight line from far away, or that a target can be reached by following a known heading. Second, we are focusing on the early experience with a new environment. Endotaxis can get an animal from zero knowledge to a cognitive map that allows reliable navigation towards goals encountered on a previous foray. It explains how an animal can return home from inside a complex environment on the first attempt [40], or navigate to a special location after encountering it just once (Figs 8.2,8.3). But it does not implement more advanced routines of spatial learning, such as stringing a habitual sequence of actions together into one, or internal deliberation to plan entire routes. Clearly, expert animals will make use of algorithms other than the beginner's choice proposed here.

A key characteristic of endotaxis, distinct from other forms of navigation, is the reliance on trial-and-error. The agent does not deliberate to plan the shortest path to the goal. Instead, it finds the shortest path by locally sampling the real-world actions available at its current point, and choosing the one that maximizes the virtual odor signal. In fact, there is strong evidence that animals navigate by real-world trial-and-error, at least in the early phase of learning [39]. Rats and mice often stop at an intersection, bend their body halfway along each direction, then choose one corridor to proceed. Sometimes they walk a few steps down a corridor, then reverse and try another one. These actions—called "vicarious trial and error"—look eerily like sniffing out an odor gradient, but they occur even in absence of any olfactory cues. Lashley [27], in his first scientific paper on visual discrimination in the rat, reported that rats at a decision point often hesitate "with a swaying back and forth between the passages." Similar behaviors occur in arthropods [48] and humans [41] when poised at a decision point. We suggest that the animal does indeed sample a gradient, not of an odor, but of an internally generated virtual odor that reflects the proximity to the goal. The animal uses the same policy of spatial sampling that it would apply to a real odor signal, consistent with the idea that endotaxis is built on the ancient behavioral module for chemotaxis.

Frequently a rodent stopped at a maze junction merely turns its head side-to-side, rather than walking down a corridor to sample the gradient. Within the endotaxis model, this could be explained if some of the point cells in the lowest layer (Figure 8.1B) are selective for head direction or for the view down a specific corridor. During navigation, activation of that "direction cell" systematically precedes activation of point cells further down that corridor. Therefore the direction cell gets integrated into the map network. From then on, when the animal turns in that direction, this action takes a step along the graph of the environment without requiring a walk in ultimately fruitless directions. In this way the agent can sample the goal gradient while minimizing energy expenditure.

The vicarious trial and error movements are commonplace early on during navigation in a new environment. Later on the animal performs them more rarely and instead moves smoothly through multiple intersections in a row [39]. This may reflect a transition between different modes of navigation, from the early endotaxis, where every action gets evaluated on its real-world merit, to a mode where many actions are strung together into behavioral motifs. At a late stage of learning the agent may also develop an internal forward model for the effects of its own actions, which would allow for prospective planning of an entire route. An interesting direction for future research is to seek a neuromorphic circuit model for such action planning; perhaps it can be built naturally on top of the endotaxis circuit.

While rodents engaged in early navigation act as though they are sniffing out a virtual odor, we would dearly like to know whether the experience *feels like* sniffing to them. The prospects for having that conversation in the near future are dim, but in the meantime we can talk to humans about the topic. Human language has an intriguing set of metaphors for decision making under uncertainty: "this doesn't smell right," "sniff out a solution," "that idea stinks," "smells fishy to me," "the sweet smell of success." All these sayings apply in situations where we don't yet understand the rules but are just feeling our way into a problem. Going beyond mere correlation, there is also a causal link: fishy smells can change people's decisions on matters entirely unrelated to fish [28]. In the endotaxis model (Figure 8.1B) this might happen if the mode switch is leaky, allowing real smells to interfere with virtual odors. Perhaps this partial synesthesia between smells and decisions results from the evolutionary repurposing of an ancient behavioral module that was intended for olfactory search.

Brain circuits

The proposed circuitry (Fig 8.1) relates closely to some real existing neural networks: the so-called cerebellum-like circuits. They include the insect mushroom body, the mammalian cerebellum, and a host of related structures in non-mammalian vertebrates [7, 16]. The distinguishing features are: a large population of neurons with selective responses (e.g. Kenyon cells, cerebellar granule cells), massive convergence from that population onto a smaller set of output neurons (e.g. Mushroom body output neurons, Purkinje cells), and synaptic plasticity at the output neurons gated by signals from the animal's experience (e.g. dopaminergic inputs to mushroom body, climbing fiber input to cerebellum). It is thought that this plasticity creates an adaptive filter by which the output neurons learn to predict the behavioral consequences of the animal's actions [7, 53]. This is what the goal cells do in the endotaxis model.

The analogy to the insect mushroom body invites a broader interpretation of what purpose that structure serves. In the conventional picture the mushroom body helps with odor discrimination and forms memories of discrete odors that are associated with salient experience [22]. Subsequently the animal can seek or avoid those odors. But insects can also use odors as landmarks in the environment. In this more general form of navigation, the odor is not a goal in itself, but serves to mark a route towards some entirely different goal [26, 44]. In ants and bees, the mushroom body receives massive visual input, and the insect uses discrete panoramic views of the landscape as markers for its location [11, 45, 51]. Our analysis shows how the mushroom body circuitry can tie together these discrete points into a cognitive map that supports navigation towards arbitrary goal locations.

In this picture a Kenyon cell that fires only under a specific pattern of receptor activation becomes selective for a specific location in the environment, and thus would play the role of a map cell in the endotaxis circuit (Figure 8.1). After sufficient exploration of the reward landscape the mushroom body output neurons come to encode the animal's proximity to a desirable goal, and that signal can guide a trial-and-error mechanism for steering. In fact, mushroom body output neurons are known to guide the turning decisions of the insect [5], perhaps through their projections to the central complex [29], an area critical to the animal's turning behavior. Conceivably this is where the insect's basic chemotaxis module is implemented, namely the policy for ascending on a goal signal.

Beyond the cerebellum-like circuits, the general ingredients of the endotaxis model – recurrent synapses, Hebbian learning, many-to-one convergence – are found

commonly in other brain areas including the mammalian neocortex and hippocampus. In the rodent hippocampus, an interesting candidate for map cells are the pyramidal cells in area CA3. Many of these neurons exhibit place fields and they are recurrently connected by synapses with Hebbian plasticity. It was suggested early on that random exploration by the agent produces correlations between nearby place cells, and thus the synaptic weights among those neurons might be inversely related to the distance between their place fields [35, 38]. However, simulations showed that the synapses are substantially strengthened only among immediately adjacent place fields [34, 38] (see also our Equation 8.21), thus limiting the utility for global navigation across the environment. Here we show that a useful global distance function emerges from the *output* of the recurrent network (Equations 8.24, 8.27, 8.28) rather than its synaptic structure. Further, we offer a biologically realistic circuit (Figure 8.1B) that can read out this distance function for subsequent navigation.

Neural signals

The endotaxis circuit proposes three types of neurons—point cells, map cells, and goal cells—and it is instructive to compare their expected signals to existing recordings from animal brains during navigation behavior. Much of that prior work has focused on the rodent hippocampal formation [33], but we do not presume that endotaxis is localized to that structure. The three cell types in the model all have place fields, in that they fire preferentially in certain regions within the graph of the environment. However, they differ in important respects:

Size and location The place field is smallest for a point cell; somewhat larger for a map cell, owing to recurrent connections in the map network; and larger still for goal cells, owing to additional pooling in the goal network. Such a wide range of place field sizes has indeed been observed in surveys of the rodent hippocampus, spanning at least a factor of 10 in diameter [25, 52]. Some place cells show a graded firing profile that fills the available environment. Furthermore one finds more place fields near the goal location of a navigation task, even when that location has no overt markers [23]. Both of those characteristics are expected of the goal cells in the endotaxis model.

Dynamics The endotaxis model assumes that point cells exist from the very outset in any environment. Indeed, many place cells in the rodent hippocampus appear within minutes of the animal's entry into an arena [18, 52]. Furthermore, any given

environment activates only a small fraction of these neurons. Most of the "potential place cells" remain silent, presumably because their sensory trigger feature doesn't match any of the locations in the current environment [3, 15]. In the endotaxis model, each of these sets of point cells is tied into a different map network, which would allow the circuit to maintain multiple cognitive maps in memory [35]. Finally a small change in the environment, such as appearance of a local barrier (Figure 8.4), can indeed lead to disappearance and appearance of nearby place cells [4].

Goal cells, on the other hand, are expected to appear suddenly when the animal first arrives at a memorable location. At that moment the goal cell's input synapses from the map network are activated and the neuron immediately develops a place field. This prediction is reminiscent of a startling experimental observation in recordings from hippocampal area CA1: a neuron can suddenly start firing with a fully formed place field that may be located anywhere in the environment [9]. This event appears to be triggered by a calcium plateau potential in the dendrites of the place cell, which potentiates the excitatory synaptic inputs the cell receives. A surprising aspect of this discovery was the large extent of the resulting place field, which would require the animal several seconds to cover. This was interpreted as a signature of a new plasticity mechanism that extends over several seconds [30]. Our endotaxis model has a different explanation for this phenomenon: the goal cell's place field extends far in space because it taps into the map network, which has already prepared a large place field prior to the agent finding the goal location. In this picture all the synaptic changes are local in time and space, and there is no need to invoke an extended time scale for plasticity.

Learning theories

Endotaxis has similarities with *reinforcement learning* (RL) [47]. In both cases the agent explores a number of locations in the environment. In RL these are called *states* and every state has an associated *value* representing how close the agent is to rewards. In endotaxis, this is the role of the virtual odor, represented by the activity of a goal neuron. The value function gets modified through the experience of reward when the agent reaches a valuable resource; in endotaxis this happens via update of the synapses in the goal network (**G** in Figure 8.1B). In both RL and endotaxis, when the animal wishes to exploit a given resource, it navigates so as to maximize the value function. Over time that value function converges to a form that allows the agent to find the goal directly from every starting state. The exponential decay of the virtual odor with increasing distance from the target (Equation 8.28) is reminiscent

of the exponential decay of the value function in RL, controlled by the discount factor, γ [47].

In endotaxis much of the learning happens independent of any reinforcement. During exploration, the circuit learns the topology of the environment, specifically by updating the synapses in the map network (**M** in Figure 8.1B). The presence of rewards is not necessary for map learning: until a resource is found for the first time, the value function remains zero because the **G** synapses have not yet been established (Equation 8.18). Eventually, when the goal is encountered, **G** is updated in one shot and the value function becomes nonzero throughout the known portion of the environment. Thus the agent learns how to navigate to the goal location from a single reinforcement (Figure 8.2). This is possible because the ground has been prepared, as it were, by learning a map. In animal behavior this phenomenon is called *latent learning*. Early debates in animal psychology pitched latent learning and reinforcement learning as alternative explanations [49]. Instead, in the endotaxis algorithm, neither can function without the other (see Equation 8.18). In *model-based* reinforcement learning, the agent could learn a forward model of the environment and use it to update a value function. A key difference is that endotaxis learns the distances between all pairs of states, and can then establish a value function after a single reinforcement, whereas RL typically requires an iterative method to establish the value function [21, 31, 46].

The neural signals in endotaxis bear some similarity to the so-called *successor representation* [13, 43]. This is a proposal for how the brain might encode the current state of the agent, intended to simplify the mathematics of time-difference reinforcement learning. Each neuron stands for a possible state of the agent. The activity of neuron j is proportional to the time-discounted probability that the agent will find itself at state j in the future. Thus, the output of the endotaxis map network (Equations 8.6, 8.24) qualitatively resembles a successor representation. However there are some important differences. First, the successor representation depends not only on the structure of the environment, but on the optimal policy of the agent, which in turn depends on the distribution of rewards. Thus the successor representation must itself be learned through a reinforcement algorithm. There is agreement in the literature that the successor representation would be more useful if the model of the environment were independent of reward structure [20]; however, it is believed that "it is more difficult to learn" [13]. By contrast, the map matrix in the endotaxis mechanism is built from a policy of random exploration independent of the reward

landscape. Second, no plausible biomorphic mechanism for learning the successor representation has been proposed yet, whereas the endotaxis circuit is made entirely from biologically realistic components.

Outlook

In summary, we have proposed a simple model for spatial learning and navigation in an unknown environment. It includes an algorithm, as well as a fully-specified neural circuit implementation, that makes quantitative and testable predictions about behavior, anatomy, and physiology, from insects to rodents (Section 8.5). Of course the same observables may be consistent with other models, and in fact multiple navigation mechanisms may be at work in parallel or during successive stages of learning. Perhaps the most distinguishing features of the endotaxis algorithm are its reliance on trial-and-error sampling, and the close relationship to chemotaxis. To explore these specific ingredients, future research could work backwards: first find the neural circuit that controls the random trial-and-error sampling of odors. Then test if that module receives a convergence of goal signals from other circuits that process non-olfactory information. If so, that could lead to the mode switch which routes one or another goal signal to the decision-making module. Finally, upstream of that mode switch lies the soul [37] of the animal that tells the navigation machinery what goal to pursue. Given recent technical developments we believe that such a program of module-tracing is almost within reach, at least for the insect brain.

8.6 Supplement

The core function of the endotaxis network is to learn the distance between any two points in the environment starting from purely local connectivity. As the agent explores the graph of the environment, the point cells for two adjacent locations briefly fire together. This is the local event that drives synaptic learning in the map population. Eventually the map network learns the global structure of the graph. In particular, for any chosen goal node on the graph, the network computes a virtual odor signal that varies with the agent's location and declines monotonically with the distance from the goal. Using that distance function the agent can navigate to the goal node by the shortest path. In this section we explain how this global distance measure comes about. We start with an analytical result about computing distances on a graph, continue with a formal analysis of how the endotaxis network functions, and proceed to numerical experiments that supplement results in the text.

A neuromorphic function to compute the shortest distance on a graph

Finding the shortest path between all pairs of nodes on a graph is a central problem of graph theory, known as "all pairs shortest path" (APSP) [54]. Generally an APSP algorithm delivers a matrix containing the distances D_{ij} for all pairs of nodes. That matrix can then be used to construct the actual sequence corresponding to the shortest path iteratively. The Floyd-Warshall algorithm [17] is simple and works even for the more general case of weighted edges between nodes. Unfortunately we know of no plausible way to implement Floyd-Warshall's three nested loops of comparison statements with neurons.

There is, however, a simple function for APSP that operates directly on the adjacency matrix and can be solved by a recurrent neural network. Specifically: if a connected, directed graph has adjacency matrix A_{ij},

$$A_{ij} = \begin{cases} 1, \text{ if node } i \text{ can be reached from node } j \text{ in one step} \\ 0, \text{otherwise, including the } i = j \text{ case} \end{cases} \tag{8.1}$$

then with a suitably small positive value of γ the shortest path distances are given by

$$D_{ij} = \left\lceil \frac{\log \left[(\mathbf{1} - \gamma \mathbf{A})^{-1} \right]_{ij}}{\log \gamma} \right\rceil \tag{8.2}$$

where $\mathbf{1}$ is the identity matrix, and the half-square brackets mean "round up to the nearest integer."

Proof: The powers of the adjacency matrix represent the effects of taking multiple steps on the graph, namely,

$$\left[\mathbf{A}^k\right]_{ij} = N_{ij}^{(k)} = \text{number of distinct paths to get from node } j \text{ to node } i \text{ in } k \text{ steps}$$

where a path is an ordered sequence of edges on the graph. This can be seen by induction as follows. By definition,

$$N_{ij}^{(1)} = A_{ij}$$

Suppose we know $N_{ij}^{(k)}$ and want to compute $N_{ij}^{(k+1)}$. Every path from j to i of length $k+1$ steps has to reach a neighbor of node i in k steps. Therefore,

$$N_{ij}^{(k+1)} = \sum_l A_{il} N_{lj}^{(k)} \tag{8.3}$$

The RHS corresponds to multiplication by $\mathbf{A}$, so the solution is

$$N_{ij}^{(k)} = \left[\mathbf{A}^k\right]_{ij}$$

We are particularly interested in the shortest path from node j to node i. If the shortest distance D_{ij} from j to i is k steps then there must exist a path of length k but not of any length $< k$. Therefore,

$$D_{ij} = \min_k N_{ij}^{(k)} > 0 \tag{8.4}$$

Now consider the Taylor series

$$\begin{aligned} \mathbf{Y} &= (\mathbf{1} - \gamma\mathbf{A})^{-1} \\ &= \mathbf{1} + \gamma\mathbf{A} + \gamma^2\mathbf{A}^2 + \ldots \end{aligned} \tag{8.5}$$

Then

$$Y_{ij} = \sum_{k=0}^{\infty} N_{ij}^{(k)} \gamma^k = N_{ij}^{(D_{ij})} \gamma^{D_{ij}} + N_{ij}^{(D_{ij}+1)} \gamma^{D_{ij}+1} + \ldots \tag{8.6}$$

We will show that if γ is chosen positive but small enough then the growth of $N_{ij}^{(k)}$ with increasing k gets eclipsed by the decay of γ^k such that

$$\gamma^{D_{ij}} < Y_{ij} < \gamma^{D_{ij}-1} \tag{8.7}$$

The left inequality is obvious from Equation 8.6 because $N_{ij}^{(D_{ij})} \geq 1$ by Equation 8.4.

To understand the right inequality, note first that $N_{ij}^{(k)}$ is bounded by a geometric series. From Equation 8.3 it follows that

$$N_{ij}^{(k)} < q^k$$

where q is the largest number of neighbors of any node on the graph. So from Equation 8.6

$$Y_{ij} < (q\gamma)^{D_{ij}} + (q\gamma)^{D_{ij}+1} + \cdots = \frac{(q\gamma)^{D_{ij}}}{1 - q\gamma} \tag{8.8}$$

This expression is $< \gamma^{D_{ij}-1}$ (Equation 8.7) as long as

$$\gamma < \frac{1}{q + q^{D_{ij}}} \tag{8.9}$$

In addition, because

$$D_{ij} < n \equiv \text{ number of nodes on the graph}$$

this is satisfied if one chooses γ such that

$$\gamma < \frac{1}{q + q^n} \tag{8.10}$$

With that condition on γ, the inequality 8.7 holds, and taking the logarithm on both sides leads to the desired result:

$$D_{ij} = \left\lceil \frac{\log Y_{ij}}{\log \gamma} \right\rceil$$

The goal signal in endotaxis

In later sections we show that Y_{ij} can be computed by the endotaxis network, and how the required synaptic weights can be learned from exploration on the graph. For reasons of practical implementation, the network does not operate on Y_{ij} directly but on the scalar products of the column-vectors in $\mathbf{Y}$, namely

$$E_{ij} = \text{“goal signal from node } j \text{ to } i\text{”} = \sum_k Y_{ki} Y_{kj} \tag{8.11}$$

To understand how that goal signal E_{ij} varies with distance one can follow arguments parallel to those that led to Equation 8.6. Using the upper bound by the geometric series (Equation 8.8) and inserting in Equation 8.11 one finds again that it is possible to choose a γ small enough to satisfy

$$\gamma^{D_{ij}} < E_{ij} < \gamma^{D_{ij}-1} \tag{8.12}$$

Under those conditions the goal signal E_{ij} decays exponentially with the graph distance D_{ij}.

Regime of validity of the goal signal

The analytical arguments above all relied on choosing a very small γ. In numerical experiments we found that the exponential dependence of the goal signal E_{ij} on distance (Equation 8.12) actually holds over a wide range of γ (Figure 8.5A).

As γ increases, one enters a regime where the systematic relationship to graph distance (Figure 8.5B) breaks down and the goal signal becomes non-monotonic: comparing all node pairs throughout the graph one now finds many instances where the pair with a larger distance produces a stronger goal signal (Figure 8.5C). This happens because Equation 8.12 is no longer satisfied. Nonetheless, it is still possible that an agent ascending on the goal signal gets all the correct local instructions to find the shortest path. To test this we asked whether the goal signal recommends the correct successor node: for every start node j and goal node i one finds the node connected to j with the highest goal signal. If that neighbor is always one step closer to i then navigation will be perfect.

Indeed we found an extended range of values for γ where the goal signal worked flawlessly for navigation between all pairs of nodes (Figure 8.5C). In this range the goal signal gives the correct turning instructions on a local level, even if it is not

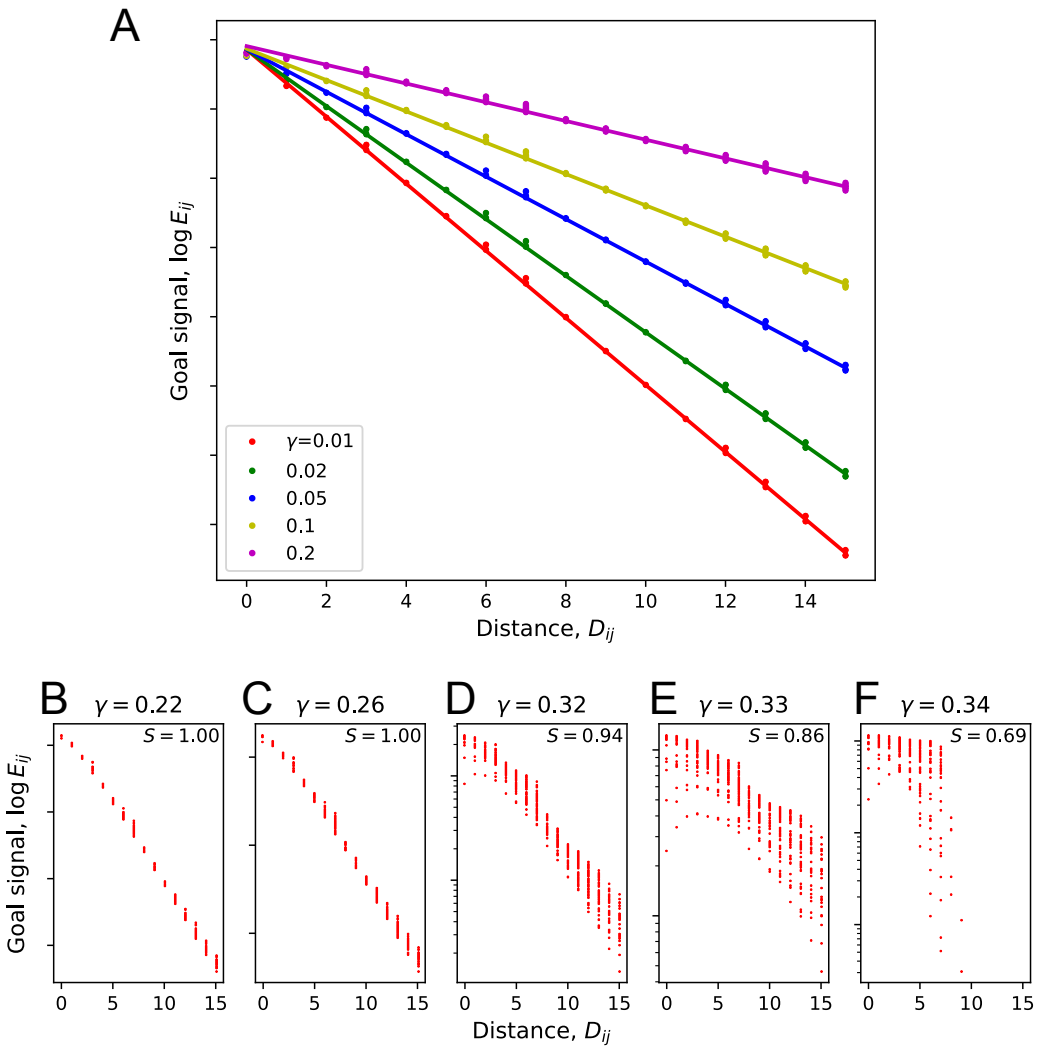

Figure 8.5: **T**he goal signal and the choice of γ. **A:** The goal signal declines exponentially with graph distance (the tower of Hanoi graph with 4 levels was used for these simulations). Data points indicate the goal signal between all pairs of nodes, computed with different values of γ, and plotted against the distance on the graph between the nodes. Lines are exponential fits to the data. **B-F:** Detailed plot of goal signal vs distance as γ approaches the critical value γ_c, which for this graph is 0.335 (Equation 8.13). The fraction of correct successors S is listed in each panel; as S drops below 1, the goal signal becomes less useful for navigation.

globally monotonic with distance across the entire graph. This behavior can also be seen in some of the simulations of random exploration (Figure 8.3B).

At higher γ values navigation begins to fail (Figure 8.5D–E). For an increasing number of start/goal pairs the agent gets trapped in a local maximum of signal before arriving at the goal.

Finally above a certain critical value γ_c, the goal signal fails catastrophically (Figure 8.5F). There is a simple mathematical reason for this: recall that the Taylor expansion (8.5) has a convergence radius of 1. That means all the eigenvalues of $\gamma\mathbf{A}$ must have absolute value < 1, which requires

$$\gamma < \gamma_c \equiv \frac{1}{\text{largest absolute eigenvalue of } \mathbf{A}} \tag{8.13}$$

Outside of that convergence radius the expression $(\mathbf{1} - \gamma\mathbf{A})^{-1}$ can no longer be interpreted as counting paths on the graph and therefore loses any connection to graph distance.

Model formulation

We formalized the endotaxis mechanism of Figure 8.1B as follows:

The environment is parcelled into a set of discrete locations in space that are sparsely connected to each other. The locations and connectors form a graph that is fully specified by the adjacency matrix A_{ij} (Equation 8.1).

We treat neural processing using a textbook linear rate model [14]. Each node on the graph has a point cell corresponding to that location. The point cell fires at a rate of 1 when the agent's position j is at that node, and at a lower level w, with $0 < w < 1$, at the neighboring nodes. Thus the firing fields of neighboring point cells overlap somewhat; this produces correlations among point cells along the agent's trajectory which will drive synaptic plasticity.

$$u_i(x) = \text{firing rate of point cell } i \text{ with the agent at node } x \tag{8.14}$$

$$= \delta_{ix} + w\, A_{ix} \tag{8.15}$$

where δ_{ix} is the Kronecker delta. The output of the map network (Figure 8.1B) is

$$\mathbf{v} = \mathbf{u} + \mathbf{M}\mathbf{v} = (\mathbf{1} - \mathbf{M})^{-1}\mathbf{u} \tag{8.16}$$

where $\mathbf{u}$ is the vector of point cell outputs, $\mathbf{v}$ is the vector of map cell outputs, and $\mathbf{M}$ is the matrix of recurrent synapses among map cells.

A goal cell g receives sensory input s_g from neurons that signal the goal resource available to the agent at the current node:

$$s_g(y) = \text{amount of resource } g \text{ present when the agent is at node } y \tag{8.17}$$

In addition, the goal cell gets input from the map neurons via the network of goal synapses. Thus the vector of goal cell activities with the agent at node x is

$$\mathbf{r}(x) = \mathbf{s}(x) + \mathbf{G}\,\mathbf{v}(x) = \mathbf{s}(x) + \mathbf{G}(\mathbf{1} - \mathbf{M})^{-1}\mathbf{u}(x) \tag{8.18}$$

The recurrent synapses among map cells undergo Hebbian plasticity. To keep the synaptic strengths bounded, some normalization rule is needed. We adopted the standard Oja's Rule [14]:

$$\frac{\mathrm{d}M_{ij}}{\mathrm{d}t} = \beta_{\mathrm{M}}(\alpha_{\mathrm{M}} v_i v_j - M_{ij} {v_i}^2) \tag{8.19}$$

where β sets the speed of synaptic plasticity and α its strength. The map network has no self-synapses: $M_{ii} = 0$.

The synapses from map cells to goal cells also undergo Hebbian plasticity, again via Oja's Rule

$$\frac{\mathrm{d}G_{gi}}{\mathrm{d}t} = \beta_{\mathrm{G}}(\alpha_{\mathrm{G}} r_g v_i - G_{gi} {r_g}^2) \tag{8.20}$$

Because learning about targets is conceptually different from learning the map of the environment, we allowed $\alpha_{\mathrm{G}}, \beta_{\mathrm{G}}$ to differ from $\alpha_{\mathrm{M}}, \beta_{\mathrm{M}}$. Including the spatial overlap w, the model has 5 parameters.

How the endotaxis network learns the goal signal

Consider the linear rate model of the map network in Figure 8.1B and Equations 8.16–8.19. It is well known that a Hebbian recurrent network of this type will

learn the correlation structure of its inputs [14, 19]. Evaluating Equation 8.19 after synapses have equilibrated leads to

$$M_{ij} = \alpha \frac{\langle v_i v_j \rangle}{\langle {v_j}^2 \rangle} \tag{8.21}$$

In the limit of small M_{ij}, i.e. if the inputs from point cells dominate, then $v_i \approx u_i$ and one gets to lowest order

$$M_{ij} \approx \alpha \frac{\langle u_i u_j \rangle}{\langle {u_i}^2 \rangle} = \alpha\, w\, A_{ij} \equiv \gamma\, A_{ij} \tag{8.22}$$

where

$$\gamma = \alpha\, w \tag{8.23}$$

In this approximation, the recurrent synapses M_{ij} directly reflect the connections among point cells and thus the adjacency matrix of the graph.

The output of the map network (Equation 8.16) is

$$\mathbf{v} = (\mathbf{1} - \mathbf{M})^{-1}\mathbf{u} = (\mathbf{1} - \gamma\mathbf{A})^{-1}\mathbf{u} \tag{8.24}$$

So the recurrent network of map cells effectively computes the all-pairs distance function derived above (Equation 8.5). If the agent is at node x then the map output $\mathbf{v}(x)$ equals the x-th column vector of the matrix $\mathbf{Y}$ (in the limit of small w and γ):

$$v_i(x) \approx Y_{ix} \tag{8.25}$$

which declines exponentially with the graph distance D_{ix} (Equation 8.7). These distance-dependent humps of activity are schematized in Figure 8.1C.

The remaining problem is how to use the map output to encode the distance to a specific remembered goal location. Suppose goal g has a rewarding resource only at node y, specifically $s_g(x) = \delta_{xy}$ (Equation 8.17). When the agent first arrives at location y, the synaptic plasticity rule (Equation 8.20) updates the goal synapses G_{gi} from zero to a profile proportional to the current map output:

$$G_{gi} \sim v_i(y) \tag{8.26}$$

Subsequent visits will strengthen that profile. From then on, when the agent is at a location $x \neq y$ the virtual odor varies according to Equation 8.18:

$$\begin{aligned} r_g(x) &= \mathbf{s}(x) + \mathbf{G}\,\mathbf{v}(x) \\ &\sim 0 + \mathbf{v}(y) \cdot \mathbf{v}(x) \equiv E_{xy} \end{aligned} \tag{8.27}$$

This corresponds to the goal signal E analyzed above (Equations 8.11, 8.12, Figure 8.5). Thus the virtual odor computed by the endotaxis network decays exponentially with the agent's distance from the goal

$$E_{xy} \sim \gamma^{D_{xy}} \tag{8.28}$$

where $\gamma = \alpha\, w$.

The explanation here relied on multiple small-signal approximations. However, our simulations show that navigation based on the virtual odor signal is robust in realistic scenarios that include fully non-linear synaptic update rules and stochastic exploration by a random walk (Figs 8.2,8.3,8.4).

In this framework, the factor γ has an interesting interpretation. Its neural meaning is the strength of recurrent synapses in the map network compared to the feed-forward synapses from point cells (Equation 8.22). Ultimately it determines the distance-dependence of the goal signal: for every step along the graph the goal signal declines by a factor of γ (Equation 8.28). By analogy to the value function in reinforcement learning [47], one can identify γ as a discount factor or cost that the agent assigns for every step it has to take. This becomes relevant when the agent trades off two goal locations that offer rewards of different magnitude (Figure 8.4C): an additional step to one of the goals gets compensated if the reward is larger by a factor of $1/\gamma$. If the agent can manipulate γ, for example by varying α in Oja's plasticity rule (Equations 8.19,8.22), that allows it to assign different costs on distance traveled (Figure 8.4C).

Limits and extensions of the endotaxis model

To help illuminate the remarkable phenomenon of rapid learning in a complex environment we sought an explanation in terms of biologically realistic processes. This informed the choice of modeling language, using concrete circuits of neurons and synapses, rather than abstract cognitive functions. Furthermore we kept the

model as simple as possible: the cells are single-compartment neurons without elaborate biophysics. The synapses are of a simple Hebbian type. All the input-output functions are linear. Free parameters are kept to the minimum: two each for the synaptic learning rules in the two networks. This simplicity allowed us to understand how and why the model works in analytical detail (Section 8.6).

Surprisingly this simplest possible model also learns very robustly in simulations over a range of environments. The parameters do not require careful tuning; in fact a single set of 4 numbers works fine for the conditions we studied. In some ways the simulations perform better than real animals. For example in the binary maze the agent can navigate to a reward location flawlessly after discovering it the first time (Figure 8.3B), whereas real mice solve that problem after ~10 experiences [40]. This inspires confidence that as one adds realistic "bells and whistles" to the model the additional degrees of freedom will not break its operation. A number of extensions seem interesting for future work.

The distance function computed by the network fundamentally relies on the decay of neural activation over multiple synaptic links. In a large environment, and operating with a small γ, the virtual odor signal will span many orders of magnitude (Equation 8.28). Real neurons cannot function reliably over such a large dynamic range, but some plausible additions could counteract the decay. A more realistic activation function with a compressive nonlinearity can amplify the signal locally in each neuron. Second, a short-term adaptive gain control might adjust the strength of synapses. In this way map cells far from the animal's current location could become more sensitive and continue to respond to the local trial-and-error movements of the agent.

Another desirable feature would be long-term memory. Animals can learn a cognitive map within minutes, and then retain it for days. Clearly there are multiple time scales for learning and forgetting. In complex brains one supposes that long-term consolidation is handled by transfer of the information between brain areas, for example hippocampus and cortex. Small insect brains don't offer that luxury, but perhaps the goal can be achieved within the endotaxis circuit itself, by endowing synapses with more complex dynamics [1].

A hierarchical extension of the model could be formulated such that an additional set of feedforward weights could read out from the goal signals in the current model formulation, which would allow for weighted preferences of desired goal features. Such a system could be useful for returning to locations with multiple properties

that are desirable to the animal, or remembering a unique set of properties that characterize certain goal locations.

Simulations

Figures 8.2, 8.3, and 8.4 report the results of endotaxis learning while an agent explores the environment. We gave the agent a trajectory, either chosen by design (Figure 8.2) or as an unbiased random walk through the graph (Figs 8.3, 8.4). After every step of the random walk, we computed the cell activities in a forward pass from point cells to goal cells. Then we updated the synaptic weights in the two networks $\mathbf{M}$ and $\mathbf{G}$ via a Hebbian learning rule. See Algorithm 1 for details. Matrix operations were implemented in JAX [10], but for the task complexity explored in this paper there was no need for GPU acceleration.

Learning and subsequent navigation worked robustly over a range of the α_{M} and β_{M} parameters in Oja's Rule (Figure 8.6). α_{M} has an absolute upper bound of γ_{c}/w (Equations 8.13, 8.24) which depends on the eigenspectrum of the graph. In practice the Tower of Hanoi graph posed the strongest challenge, presumably because of its size and the large number of loops. For simplicity, we selected model parameters that allow for perfect navigation on that graph and applied the same model without modifications across all the tasks reported here. Note that this is not an exclusive set: smaller values for α_{M} and β_{M} would work as well.

Change in connectivity

To analyze changes in connectivity (Figure 8.4A.i-ii) we simulated an agent performing a random walk on a ring. At each time step we asked if the agent could navigate to the goal by the shortest path. We assumed that the appearance of a block or a shortcut between two adjacent nodes will alter the sensory cues around both locations (2 and 3 in Figure 8.4A.i-ii). Therefore the point cells that used to encode those locations drop silent, and the respective map cells lose their afferent input, while still remaining in the recurrent network. At the same time two new point cells appear at those locations, because the new cues match their selectivity. Their map cells now receive afferent input from the respective locations, but their recurrent synapses start at zero weight. The agent then continues a random walk around the ring, subject to the new constraints, and the learning algorithm proceeds as usual.

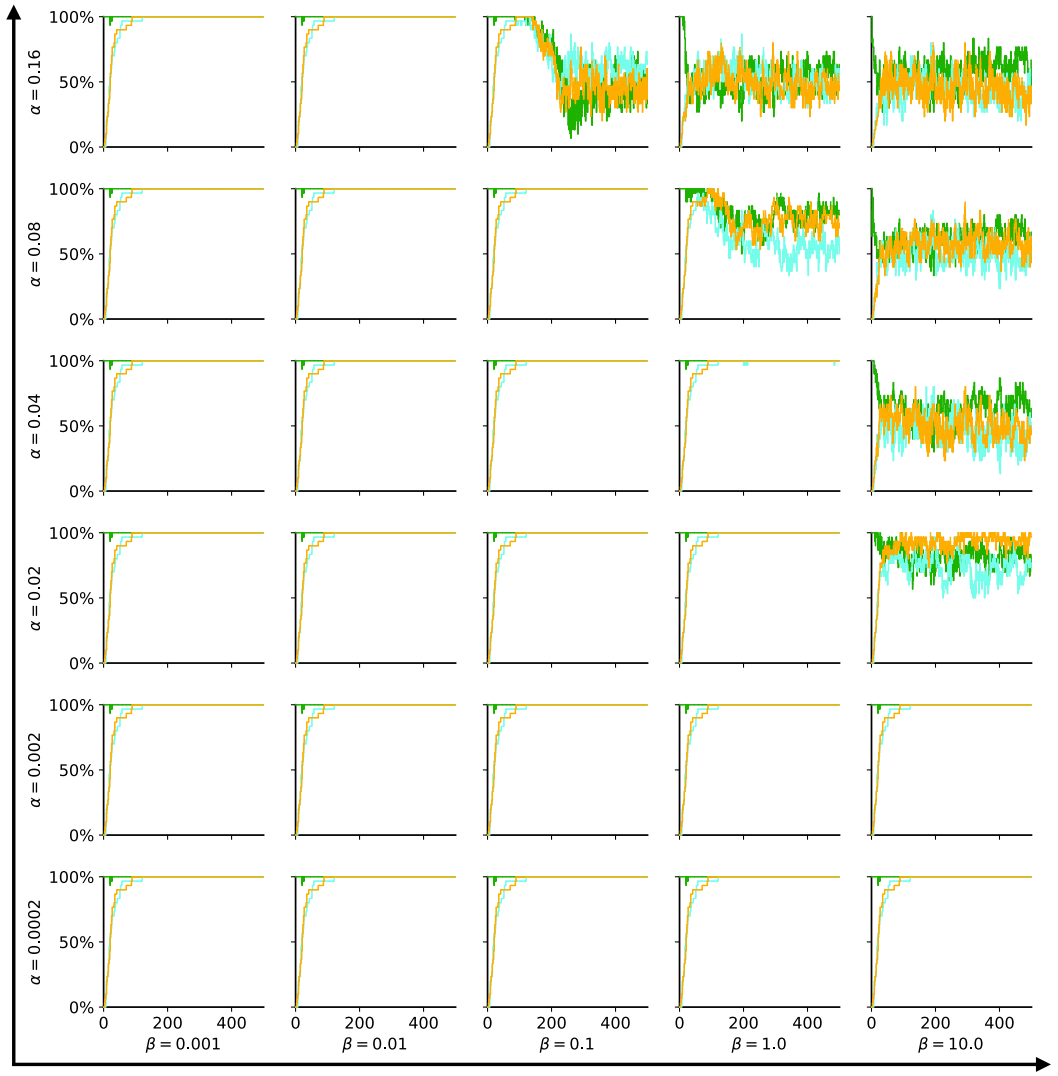

Figure 8.6: Dependence of map learning on the parameters α_M and β_M in Oja's rule. Each panel is for one combination of α_M and β_M and shows performance on the Gridworld task (Figs 8.2, 8.3-i). The fraction of successful navigations is plotted vs the number of steps in the exploratory random walk, averaged over 30 different walks. The 3 curves show navigation to the 3 goals, color coded as in Figure 8.3-i.

Algorithm 1 Online Learning via Oja's Rule

j : pre-synaptic neuron
i : post-synaptic neuron
$w = 0.3$ (fractional activity at neighbor nodes)
$s_g = 1$ (except dual-target tasks)
$\alpha_{\mathrm{M}} = 0.05$
$\beta_{\mathrm{M}} = 0.02$
$\alpha_{\mathrm{G}} = 0.5 \cdot \alpha_{\mathrm{M}}$
$\beta_{\mathrm{G}} = 0.03$

$\mathbf{M} \leftarrow 0$
$\mathbf{G} \leftarrow 0$

for *step t in node visit sequence* **do**
  Compute Neural Activity
  $\boldsymbol{u}_{node(t)} \leftarrow 1$
  for *each neighboring node i* **do**
    $\boldsymbol{u}_{node(i)} \leftarrow w$
  end for
  $\boldsymbol{u}_{node(others)} \leftarrow 0$
  $\boldsymbol{v} = \boldsymbol{u} + \mathbf{M}\boldsymbol{v} = (\mathbf{1} - \mathbf{M})^{-1}\boldsymbol{u}$
  $\boldsymbol{g} = \mathbf{G}\boldsymbol{v} + \boldsymbol{s}_{node(t)}$

  Synaptic Learning
  $M_{ij} \leftarrow M_{ij} + \beta_{\mathrm{M}}(\alpha_{\mathrm{M}} v_i v_j - M_{ij} v_i^2)$
  $G_{ij} \leftarrow G_{ij} + \beta_{\mathrm{G}}(\alpha_{\mathrm{G}} g_i v_j - G_{ij} g_i^2)$

end for

Dynamics of learning

Figure 8.7 illustrates the state of the synaptic networks over the course of online learning, as observed during a random walk on the binary maze graph (Figure 8.3A-ii). The norm of the map matrix $\|\mathbf{M}\|$ increases continuously through steady small updates $\|\mathrm{d}\mathbf{M}\|$. By comparison the goal matrix $\|\mathbf{G}\|$ increases in noticeable steps of $\|\mathrm{d}\mathbf{G}\|$ every time the agent visits a goal location. With sufficiently low α and β, the network learns stably and gradually approaches a steady state. However, as demonstrated in the text, even the first visit to a goal location already produces a goal signal that allows a reliable return to that location.

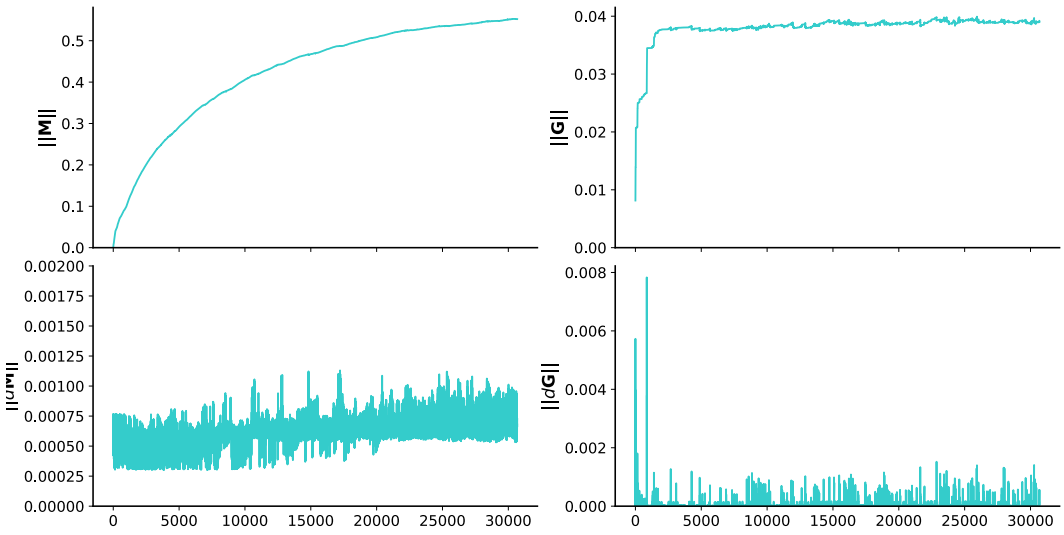

Figure 8.7: Dynamics of online learning. Evolution of the map matrix (‖**M**‖ and ‖d**M**‖) and the goal matrix (‖**G**‖ and ‖d**G**‖) during exploration of the binary maze graph of Figure 8.3A-ii. See text for details.

Robustness to noise

We tested how robust the map learning is to noise. Figure 8.8 illustrates the results using the Gridworld task (Figure 8.3-i). At each step of the simulation we perturbed each neuron's signal with multiplicative noise, by adding a Gaussian noise variable to the logarithm. Performance of learning and navigation was robust for signal-to-noise ratios of 2 or higher.

Data and code availability

Data and code to reproduce the reported results are available at `https://github.com/tonyzhang25/Zhang-2021-Endotaxis`.

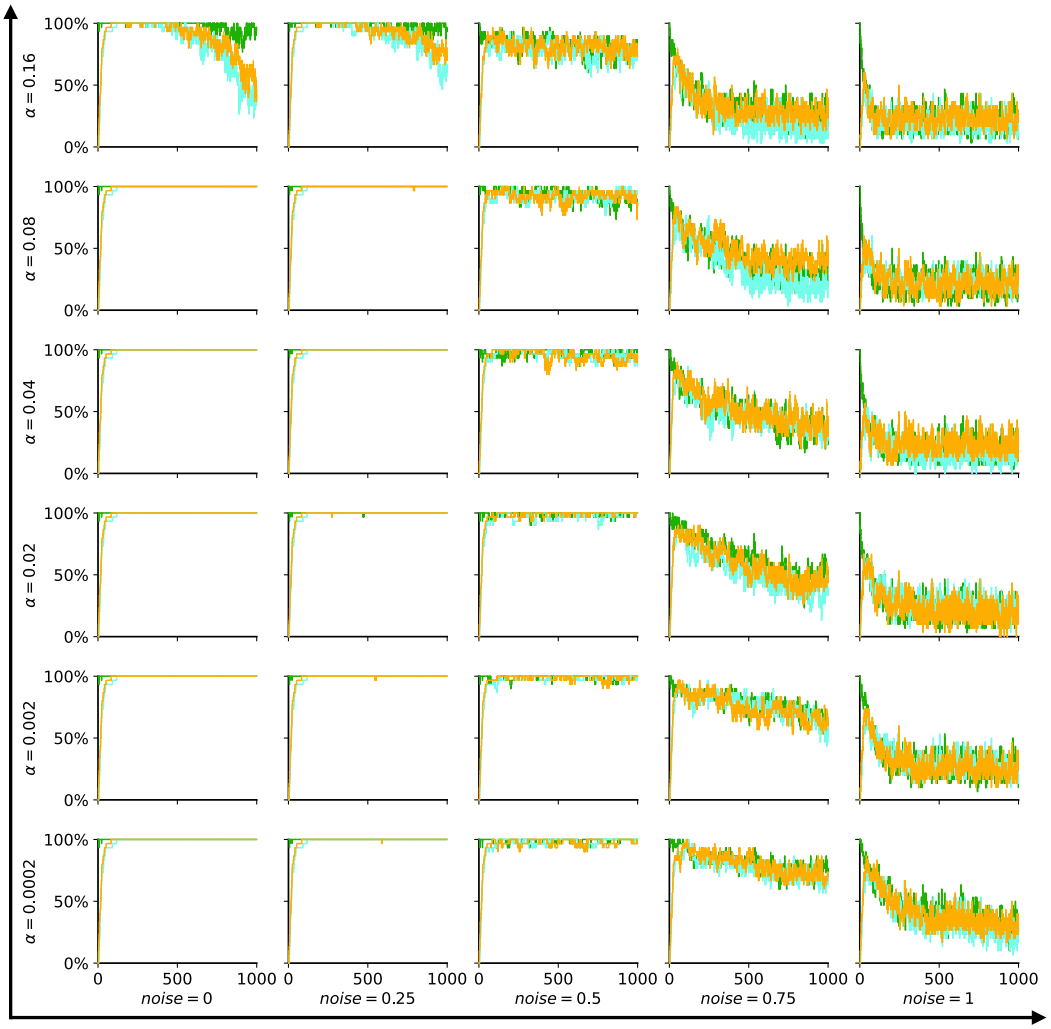

Figure 8.8: Learning tolerates perturbation by neural noise. Each panel shows navigation performance on the Gridworld task (Figs 8.2, 8.3-i), plotted as in Figure 8.6. Each neuron's activity was perturbed by multiplicative noise proportional to the unit's activity. The panels differ by the combination of α_M (rows) and noise level (columns). The noise level as a fraction of the unit's firing rate is listed below each column.

References

[1] L. F. Abbott and W. G. Regehr. "Synaptic Computation." In: *Nature* 431 (Oct. 2004), pp. 796–803. DOI: 10.1038/nature03010.

[2] Francisco Aboitiz and Juan F. Montiel. "Olfaction, Navigation, and the Origin of Isocortex". English. In: *Frontiers in Neuroscience* 9 (2015). ISSN: 1662-453X. DOI: 10.3389/fnins.2015.00402.

[3] Charlotte B. Alme, Chenglin Miao, Karel Jezek, Alessandro Treves, Edvard I. Moser, and May-Britt Moser. "Place Cells in the Hippocampus: Eleven Maps for Eleven Rooms". en. In: *Proceedings of the National Academy of Sciences* 111.52 (Dec. 2014), pp. 18428–18435. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1421056111.

[4] Alice Alvernhe, Etienne Save, and Bruno Poucet. "Local Remapping of Place Cell Firing in the Tolman Detour Task". eng. In: *The European Journal of Neuroscience* 33.9 (May 2011), pp. 1696–1705. ISSN: 1460-9568. DOI: 10.1111/j.1460-9568.2011.07653.x.

[5] Y. Aso, D. Sitaraman, T. Ichinose, K. R. Kaun, K. Vogt, G. Belliart-Guerin, P. Y. Placais, A. A. Robie, N. Yamagata, C. Schnaitmann, W. J. Rowell, R. M. Johnston, T. T. Ngo, N. Chen, W. Korff, M. N. Nitabach, U. Heberlein, T. Preat, K. M. Branson, H. Tanimoto, and G. M. Rubin. "Mushroom Body Output Neurons Encode Valence and Guide Memory-Based Action Selection in Drosophila." In: *Elife* 3 (2014), e04580. DOI: 10.7554/eLife.04580.

[6] Keeley L. Baker, Michael Dickinson, Teresa M. Findley, David H. Gire, Matthieu Louis, Marie P. Suver, Justus V. Verhagen, Katherine I. Nagel, and Matthew C. Smear. "Algorithms for Olfactory Search across Species". In: *The Journal of Neuroscience* 38.44 (Oct. 2018), pp. 9383–9389. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.1668-18.2018.

[7] Curtis C. Bell, Victor Han, and Nathaniel B. Sawtell. "Cerebellum-like Structures and Their Implications for Cerebellar Function". eng. In: *Annual Review of Neuroscience* 31 (2008), pp. 1–24. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.30.051606.094225.

[8] H. C. Berg. "A Physicist Looks at Bacterial Chemotaxis." In: *Cold Spring Harb Symp Quant Biol* 53 Pt 1 (1988), pp. 1–9.

[9] Katie C. Bittner, Aaron D. Milstein, Christine Grienberger, Sandro Romani, and Jeffrey C. Magee. "Behavioral time scale synaptic plasticity underlies CA1 place fields". en. In: *Science* 357.6355 (Sept. 2017). Publisher: American Association for the Advancement of Science Section: Report, pp. 1033–1036. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aan3846. URL: https://science.sciencemag.org/content/357/6355/1033 (visited on 08/15/2020).

[10] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. *JAX: composable transformations of Python+NumPy programs*. Version 0.2.5. 2018. URL: http://github.com/google/jax.

[11] Cornelia Buehlmann, Beata Wozniak, Roman Goulard, Barbara Webb, Paul Graham, and Jeremy E. Niven. "Mushroom Bodies Are Required for Learned Visual Navigation, but Not for Innate Visual Behavior, in Ants". eng. In: *Current biology: CB* 30.17 (Sept. 2020), 3438–3443.e2. ISSN: 1879-0445. DOI: 10.1016/j.cub.2020.07.013.

[12] Thomas S. Collett and Matthew Collett. "Memory Use in Insect Visual Navigation". en. In: *Nature Reviews Neuroscience* 3.7 (July 2002), pp. 542–552. ISSN: 1471-0048. DOI: 10.1038/nrn872.

[13] Peter Dayan. "Improving Generalization for Temporal Difference Learning: The Successor Representation". In: *Neural Computation* 5.4 (July 1993), pp. 613–624. ISSN: 0899-7667. DOI: 10.1162/neco.1993.5.4.613.

[14] Peter Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Computational Neuroscience. Cambridge, Mass.: MIT Press, 2001.

[15] J. Epsztein, M. Brecht, and A. K. Lee. "Intracellular Determinants of Hippocampal CA1 Place and Silent Cell Activity in a Novel Environment." In: *Neuron* 70 (Apr. 2011), pp. 109–20. DOI: 10.1016/j.neuron.2011.03.006.

[16] S. M. Farris. "Are Mushroom Bodies Cerebellum-like Structures?" In: *Arthropod Struct Dev* 40 (July 2011), pp. 368–79. DOI: 10.1016/j.asd.2011.02.004.

[17] Robert W. Floyd. "Algorithm 97: Shortest Path". In: *Communications of the ACM* 5.6 (June 1962), p. 345. ISSN: 0001-0782. DOI: 10.1145/367766.368168.

[18] Loren M. Frank, Garrett B. Stanley, and Emery N. Brown. "Hippocampal Plasticity across Multiple Days of Exposure to Novel Environments". en. In: *Journal of Neuroscience* 24.35 (Sept. 2004), pp. 7681–7689. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1958-04.2004.

[19] Mathieu Galtier, Olivier Faugeras, and Paul Bressloff. "Hebbian Learning of Recurrent Connections: A Geometrical Perspective". In: *Neural computation* 24 (May 2012), pp. 2346–83. DOI: 10.1162/NECO_a_00322.

[20] Jesse P. Geerts, Fabian Chersi, Kimberly L. Stachenfeld, and Neil Burgess. "A General Model of Hippocampal and Dorsal Striatal Learning and Decision Making". en. In: *Proceedings of the National Academy of Sciences* 117.49 (Dec. 2020), pp. 31427–31437. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2007981117.

[21] David Ha and Jürgen Schmidhuber. “Recurrent World Models Facilitate Policy Evolution”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

[22] Martin Heisenberg. “Mushroom Body Memoir: From Maps to Models”. en. In: *Nature Reviews Neuroscience* 4.4 (Apr. 2003), pp. 266–275. ISSN: 1471-0048. DOI: 10.1038/nrn1074.

[23] Stig A. Hollup, Sturla Molden, James G. Donnett, May-Britt Moser, and Edvard I. Moser. “Accumulation of Hippocampal Place Fields at the Goal Location in an Annular Watermaze Task”. en. In: *Journal of Neuroscience* 21.5 (Mar. 2001), pp. 1635–1644. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.21-05-01635.2001.

[24] Lucia F. Jacobs. “From Chemotaxis to the Cognitive Map: The Function of Olfaction”. en. In: *Proceedings of the National Academy of Sciences* 109.Supplement 1 (June 2012), pp. 10693–10700. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1201880109.

[25] Kirsten Brun Kjelstrup, Trygve Solstad, Vegard Heimly Brun, Torkel Hafting, Stefan Leutgeb, Menno P. Witter, Edvard I. Moser, and May-Britt Moser. “Finite Scale of Spatial Representation in the Hippocampus”. eng. In: *Science (New York, N.Y.)* 321.5885 (July 2008), pp. 140–143. ISSN: 1095-9203. DOI: 10.1126/science.1157086.

[26] Markus Knaden and Paul Graham. “The Sensory Ecology of Ant Navigation: From Natural Environments to Neural Mechanisms”. In: *Annual Review of Entomology, Vol 61*. Ed. by M. R. Berenbaum. Vol. 61. 2016, pp. 63–76. ISBN: 978-0-8243-0161-3. DOI: 10.1146/annurev-ento-010715-023703.

[27] K. S. Lashley. “Visual Discrimination of Size and Form in the Albino Rat”. In: *Journal of Animal Behavior* 2.5 (1912), pp. 310–331. ISSN: 0095-9928(Print). DOI: 10.1037/h0071033.

[28] Spike W. S. Lee and Norbert Schwarz. “Bidirectionality, mediation, and moderation of metaphorical effects: the embodiment of social suspicion and fishy smells”. eng. In: *Journal of Personality and Social Psychology* 103.5 (Nov. 2012), pp. 737–749. ISSN: 1939-1315. DOI: 10.1037/a0029708.

[29] Feng Li, Jack W Lindsey, Elizabeth C Marin, Nils Otto, Marisa Dreher, Georgia Dempsey, Ildiko Stark, Alexander S Bates, Markus William Pleijzier, Philipp Schlegel, Aljoscha Nern, Shin-ya Takemura, Nils Eckstein, Tansy Yang, Audrey Francis, Amalia Braun, Ruchi Parekh, Marta Costa, Louis K Scheffer, Yoshinori Aso, Gregory SXE Jefferis, Larry F Abbott, Ashok Litwin-Kumar, Scott Waddell, and Gerald M Rubin. “The Connectome of the Adult Drosophila Mushroom Body Provides Insights into Function”. In: *eLife* 9 (Dec.

2020). Ed. by Leslie C Griffith, Eve Marder, Leslie C Griffith, Jason Pipkin, and Chris Q Doe, e62576. ISSN: 2050-084X. DOI: 10.7554/eLife.62576.

[30] Jeffrey C. Magee and Christine Grienberger. "Synaptic Plasticity Forms and Functions". In: *Annual Review of Neuroscience* 43.1 (2020). _eprint: https://doi.org/10.1146/annurev-neuro-090919-022842, pp. 95–117. DOI: 10.1146/annurev-neuro-090919-022842. URL: https://doi.org/10.1146/annurev-neuro-090919-022842 (visited on 06/23/2021).

[31] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. "Model-based reinforcement learning: A survey". In: *arXiv preprint arXiv:2006.16712* (2020).

[32] R. G. M. Morris, P. Garrud, J. N. P. Rawlins, and J. O'Keefe. "Place Navigation Impaired in Rats with Hippocampal Lesions". en. In: *Nature* 297.5868 (June 1982), pp. 681–683. ISSN: 1476-4687. DOI: 10.1038/297681a0.

[33] May-Britt Moser, David C. Rowland, and Edvard I. Moser. "Place Cells, Grid Cells, and Memory". In: *Cold Spring Harbor Perspectives in Biology* 7.2 (Feb. 2015), a021808. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a021808.

[34] R. U. Muller, M. Stead, and J. Pach. "The Hippocampus as a Cognitive Graph". eng. In: *The Journal of General Physiology* 107.6 (June 1996), pp. 663–694. ISSN: 0022-1295. DOI: 10.1085/jgp.107.6.663.

[35] Robert U. Muller, John L. Kubie, and Russ Saypoff. "The hippocampus as a cognitive graph (abridged version)". en. In: *Hippocampus* 1.3 (1991). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hipo.450010306, pp. 243–246. ISSN: 1098-1063. DOI: 10.1002/hipo.450010306. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.450010306 (visited on 07/02/2021).

[36] Martin Müller and Rüdiger Wehner. "Path Integration in Desert Ants, Cataglyphis Fortis". en. In: *Proceedings of the National Academy of Sciences* 85.14 (July 1988), pp. 5287–5290. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.85.14.5287.

[37] Plato. "The Apology". In: (ca 400 BCE).

[38] A. D. Redish and D. S. Touretzky. "The Role of the Hippocampus in Solving the Morris Water Maze". eng. In: *Neural Computation* 10.1 (Jan. 1998), pp. 73–111. ISSN: 0899-7667. DOI: 10.1162/089976698300017908.

[39] A. David Redish. "Vicarious Trial and Error". In: *Nature reviews. Neuroscience* 17.3 (Mar. 2016), pp. 147–159. ISSN: 1471-003X. DOI: 10.1038/nrn.2015.30.

[40] Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. "Mice in a Labyrinth Exhibit Rapid Learning, Sudden Insight, and Efficient Exploration". In: *eLife* 10 (July 2021). Ed. by Mackenzie W Mathis, e66175. ISSN: 2050-084X. DOI: 10.7554/eLife.66175.

[41] Diogo Santos-Pata and Paul F. M. J. Verschure. “Human Vicarious Trial and Error Is Predictive of Spatial Navigation Performance”. In: *Frontiers in Behavioral Neuroscience* 12 (Oct. 2018), p. 237. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2018.00237.

[42] Marielena Sosa and Lisa M. Giocomo. “Navigating for Reward”. en. In: *Nature Reviews Neuroscience* (July 2021), pp. 1–16. ISSN: 1471-0048. DOI: 10.1038/s41583-021-00479-z.

[43] Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. “The Hippocampus as a Predictive Map”. eng. In: *Nature Neuroscience* 20.11 (Nov. 2017), pp. 1643–1653. ISSN: 1546-1726. DOI: 10.1038/nn.4650.

[44] Kathrin Steck, Bill S. Hansson, and Markus Knaden. “Smells like Home: Desert Ants, Cataglyphis Fortis, Use Olfactory Landmarks to Pinpoint the Nest”. In: *Frontiers in Zoology* 6.1 (Feb. 2009), p. 5. ISSN: 1742-9994. DOI: 10.1186/1742-9994-6-5.

[45] Xuelong Sun, Shigang Yue, and Michael Mangan. “A Decentralised Neural Model Explaining Optimal Integration of Navigational Strategies in Insects”. In: *eLife* 9 (June 2020). Ed. by Mani Ramaswami, Michael B Eisen, and Stanley Heinze, e54026. ISSN: 2050-084X. DOI: 10.7554/eLife.54026.

[46] Richard S Sutton. “Integrated architectures for learning, planning, and reacting based on approximating dynamic programming”. In: *Machine learning proceedings 1990*. Elsevier, 1990, pp. 216–224.

[47] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. en. MIT Press, Nov. 2018. ISBN: 978-0-262-03924-6.

[48] Michael Tarsitano. “Route Selection by a Jumping Spider (Portia Labiata) during the Locomotory Phase of a Detour”. In: *Animal Behaviour* 72.6 (Dec. 2006), pp. 1437–1442. ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2006.05.007.

[49] D. Thistlethwaite. “A Critical Review of Latent Learning and Related Experiments”. In: *Psychological Bulletin* 48.2 (1951), pp. 97–129. ISSN: 0033-2909. DOI: 10.1037/h0055171.

[50] E. C. Tolman. “Cognitive Maps in Rats and Men”. In: *Psychological Review* 55.4 (1948). WOS:A1948UY69500001, pp. 189–208. ISSN: 0033-295X. DOI: 10.1037/h0061626.

[51] Barbara Webb and Antoine Wystrach. “Neural Mechanisms of Insect Navigation”. en. In: *Current Opinion in Insect Science*. Pests and Resistance * Behavioural Ecology 15 (June 2016), pp. 27–39. ISSN: 2214-5745. DOI: 10.1016/j.cois.2016.02.011.

[52] M. A. Wilson and B. L. McNaughton. “Dynamics of the hippocampal ensemble code for space”. eng. In: *Science (New York, N.Y.)* 261.5124 (Aug. 1993), pp. 1055–1058. ISSN: 0036-8075. DOI: 10.1126/science.8351520.

[53] D. M. Wolpert, R. C. Miall, and M. Kawato. “Internal Models in the Cerebellum”. In: *Trends in Cognitive Sciences* 2.9 (Sept. 1998), pp. 338–347. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(98)01221-2.

[54] Uri Zwick. “Exact and approximate distances in graphs — A survey”. In: *Algorithms — ESA 2001*. Ed. by Friedhelm Meyer auf der Heide. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 33–48. ISBN: 978-3-540-44676-7.

Part V

Conclusion

Chapter 9

SUMMARY

This thesis presented new methods and findings in studying biological intelligence in learning and decision-making, including a new automated setup for behavioral data collection, class of tasks for studying sequential decision-making, and a theoretical neural circuit model that could produce key behavioral phenomena in mapping and navigation.

In the first work, we developed and validated an automated tool for training and collecting behavior data for groups of animals, examined problems with 2-AFC tasks commonly used in neuroscience, and developed an explainable learned model for decoding the policies driving suboptimal behavior. In the second work, we proposed a new navigation-based sequence learning task that is more suitable for understanding fast and complex learning and decision-making. We find that with this task, animals are able to reach goals that require the integration of many actions very quickly, including homing and reward-seeking. And finally, in the third work, we proposed a concrete neural circuit that can solve sequence learning in navigation via simultaneous mapping and goal learning one-shot.

The general approach used in this thesis can be used to study a wide range of intelligent, learned behavior, as long as there exists hardware sensors and computational tools to quantify behavioral states. Overall, this thesis highlighted two important criteria for studying behavior: automating behavioral setups for a large quantity of data collection, and ethologically relevant task design in order to evoke natural behavior that may involve more relevant neural pathways for neuroscience. The results highlighted in the behavioral work show that animals are capable of fast learning when placed in ethological tasks, even in complex tasks that involve learning many decisions. Finally, as shown in the theory work, even a mechanistic neural circuit could be useful for thinking about behavior, as these models could inform the types of learned behavior that is feasible.

Chapter 10

DISCUSSION

In Chapter 7, although the behavioral data presented were all from mice navigating a maze, the underlying neural mechanisms of sequence learning that makes the mapping and navigation possible may well generalize across species and possibly even tasks. It could be argued that there exists a set of tasks of equal complexity measured in terms of roughly the number of decisions learned, in the space of all possible tasks, that each organism has evolved to excel at solving due to relevance for survival. For instance, while mice may be able to reliably navigate in complex underground burrows, a human might instead rely on more complex visual cues both locally and in the far distance for navigation [3, 7].

In the theory work presented in Chapter 8, because point cells heavily rely on sensory stimuli for stability, the sensory modality should be tailored to the subject based on the primary modality of the species. For humans, visual cues in a richly rendered three-dimensional environment may be necessary, while perhaps narrow tunnels with tactile cues would work well for mice. Generally, there is little reason to believe that the various types of sequence learning could not leverage the same type of neural mechanism, even if the actual circuits responsible for each kind of sequence learning are distributed across brain regions. It could even be argued that the mechanisms underlying the learning of correlated structures in time can be generalized across species for general-purpose reasoning and cognition.

Endotaxis also has no specialized properties that limit it to any particular organism. The basic building blocks of the model are truly universal across all nervous systems: excitatory synapses and neurons. When learning to navigate, learning the adjacency graph is equivalent across species as long as there exist point cells that are localized to specific regions of physical or abstract space, and are co-active within the window for synaptic potentiation. Given the right form of ethological task, across the many motile organisms that could learn, there may already exist a large set of sparsely coded cells similar to point cells that respond to specific locations of an environment. These low-level point cells may be the result of complex multisensory integration that allows the animal to attach localized place fields to location with unique properties, or just tiled regularly in an open environment.

What is not addressed in this thesis is how these point cells might come about—it is assumed that this is the process of some sensorimotor information integration over time, and their presence is supported by experimental evidence [1, 4]. In an actual network of neurons, there may be well be drifts in the place fields of each point cell over time due to accumulated errors in sensing. Drifts in the place fields of point cells could potentially affect the ability of the goal cells to generate signals that are flawless for direct-route navigation, depending on the amount of error and the complexity of the environment. Nonetheless, the ability of the model to work on all graphs in simulation flawlessly gives us confidence that with more noisy signals coming from the input, the system could still operate in a way that is not unlike actual mice or humans, which we do not expect to learn maps one-shot without errors. By modeling these errors in point cell place fields and comparing against actual neural and behavioral data, perhaps one could establish some notion of the fault tolerance of such a system, and thereby establish some bounds on the types of learning possible.

Chapter 11

FUTURE DIRECTIONS

There is a range of follow-up directions one could pursue from the works presented in this dissertation, in terms of both behavior and theory.

On the behavioral front, a clear extension would be to expand on the interplay between tasks and models, and use a model-based approach to iteratively optimize and inform the development of new tasks that can tease apart the different mechanisms of behavior. With a flexible task design that allows for easy reconfiguration, one can imagine using an adaptive approach to studying behavior, where the task can be iteratively reconfigured to maximize a particular objective of interest, such as increasing the distance in some behavioral feature space between two proposed competing models of behavior.

On the theory front, the neural theory work presented in Chapter 8 could also be extended in many other directions. The most obvious would be to look for the proposed neural signals in the brain. However, there are also other extensions possible on the modeling side: for instance, the framework could be extended to allows for general relational reasoning. One can imagine the brain may use a similar approach to learn distances between features of more abstract concepts, and learn a relational network that allows for general cognition and deduce a causal graph of the world that can be used for any downstream task.

Endotaxis could also be applied on the algorithmic front by implementing it in silicon. One advantage of the framework discussed is its potential in solving graph search problems with better time complexities than existing graph algorithms, should it be implemented in a neural or neuromorphic circuit. One could imagine using a neuromorphic chip that can implement the RNN described in the paper, and using it to instantly solve APSP. Solving APSP can have implications for various applications, such as path planning in robotics [2].

BIBLIOGRAPHY

[1] William N. Butler, Kiah Hardcastle, and Lisa M. Giocomo. “Remembered Reward Locations Restructure Entorhinal Spatial Maps”. en. In: *Science* 363.6434 (Mar. 2019), pp. 1447–1452. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aav5297.

[2] Voemir Kunchev, Lakhmi Jain, Vladimir Ivancevic, and Anthony Finn. “Path planning and obstacle avoidance for autonomous mobile robots: A review”. In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2006, pp. 537–544.

[3] Eleanor A Maguire, Neil Burgess, and John O’Keefe. “Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates”. In: *Current opinion in neurobiology* 9.2 (1999), pp. 171–177.

[4] May-Britt Moser, David C Rowland, and Edvard I Moser. “Place cells, grid cells, and memory”. In: *Cold Spring Harbor perspectives in biology* 7.2 (2015), a021808.

[5] Mu Qiao, Tony Zhang, Cristina Segalin, Sarah Sam, Pietro Perona, and Markus Meister. “Mouse Academy: high-throughput automated training and trial-by-trial behavioral analysis during learning”. In: *bioRxiv* (2018), p. 467878. URL: https://www.biorxiv.org/content/10.1101/467878v2.

[6] Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. “Mice in a Labyrinth Exhibit Rapid Learning, Sudden Insight, and Efficient Exploration”. In: *eLife* 10 (July 2021). Ed. by Mackenzie W Mathis, e66175. ISSN: 2050-084X. DOI: 10.7554/eLife.66175.

[7] Jesse N. Weber, Brant K. Peterson, and Hopi E. Hoekstra. “Discrete Genetic Modules Are Responsible for Complex Burrow Evolution in Peromyscus Mice”. en. In: *Nature* 493.7432 (Jan. 2013), pp. 402–405. ISSN: 1476-4687. DOI: 10.1038/nature11816.

[8] Tony Zhang, Matthew Rosenberg, Pietro Perona, and Markus Meister. “Endotaxis: A Universal Algorithm for Mapping, Goal-Learning, and Navigation”. In: *bioRxiv* (2021). URL: https://www.biorxiv.org/content/10.1101/2021.09.24.461751v1.