Modeling hippocampal replay in spatial navigation with the theory of reinforcement learning : a neuroscientific and robotic approach

Elisa Massi

SORBONNE UNIVERSITÉ

DOCTORAL THESIS

Modeling hippocampal replay in spatial navigation with the theory of reinforcement learning: a neuroscientific and robotic approach

---

Author:
Elisa MASSI

Supervisor:
Dr. Benoît GIRARD
Co-supervisor:
Dr. Mehdi KHAMASSI

President of the jury:
Dr. Laure Rondi-Reig – Sorbonne Université, CNRS
Reviewers:
Prof. Lola Cañamero – CY Cergy Paris Université, ENSEA, CNRS
Prof. Sen Cheng – Ruhr-Universität Bochum
Examiner:
Dr. Nicolas Cuperlier – CY Cergy Paris Université, ENSEA, CNRS

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Institut des Systèmes Intelligents et de Robotique (ISIR), UMR 7222, CNRS, Sorbonne Université, Paris, France
École Doctorale ED3C Cerveau, Comportement, Cognition

SORBONNE UNIVERSITÉ

Abstract

École Doctorale ED3C Cerveau, Comportement, Cognition

Doctor of Philosophy

Modeling hippocampal replay in spatial navigation with the theory of reinforcement learning: a neuroscientific and robotic approach

by Elisa MASSI

The experience gained by interacting with the surrounding world is the primary way animals and humans learn. The mammalian brain can re-elaborate past experiences and contextually organize them through neural circuitries which involve the hippocampus. Hippocampal reactivations of place cells seem to exploit experience to infer the outcome of new situations, as it has been studied in rodent spatial navigation experiments.

Recently, the Reinforcement Learning (RL) theory has been proved to be very efficient in modeling goal-directed navigation and the contributions of different types of hippocampal replay. However, how it can account for the richness of exploratory behaviors is still a matter of debate. Thus, our first contribution has been designing and validating a data-driven exploration model for rodents.

Our first research interest was identifying common behavioral characteristics in rodent free exploration and modeling them as a valued-based decision-making model. Starting from observations and data analyses performed on a new rodent dataset, we propose a parametrized general decision-making model, where decisions are based on the perceived safety of a location and the biomechanical cost and persistence of the animal's exploratory dynamic. Eventually, we validate the adoption of the same model on two new rodent datasets, freely exploring different mazes for different periods. We discuss future model improvements to better adapt it to a broader range of situations.

Free exploration represents a particular case of exploratory behavior when no external conditioning emotionally affects the animal. When animals experience positive or negative external stimuli, their exploratory behavior changes. Research studies have shown that hippocampal reactivations represent emotional-related locations more frequently. However, very few studies directly address this phenomenon following aversive and appetitive stimuli and comparing the mechanisms involved. Our second contribution concerns the extension of the free exploration model to describe these mechanisms. Starting from existing RL models of hippocampal replay, we extended our free exploration model with a component accounting for learned and replayed stimulus valence. Our results can qualitatively reproduce the correlation between the estimated amount of hippocampal replay during sleep and the differential occupancy of the shock zone in post- and pre-conditioning found in the experimental data by our collaborators. Moreover, they raise new interesting experimental predictions concerning the increasing relevance of sleep replay in proper learning the post-conditioning behavior of the animal in negative conditioning, compared to their relevance in the positive conditioning case.

Learning goal-directed behaviors is also crucial in designing adaptive artificial agents and robots. Autonomous robots usually have limited knowledge of the stochastic nature of the real world surrounding them, and one of the most powerful sets of online learning algorithms, RL, is often neither responsive enough nor time-efficient for the constraints imposed by real robotic applications. Even before the first neurophysiological studies on hippocampal reactivations, machine learning research proposed mechanisms for experience replay in RL algorithms to enhance the speed of learning and the adaptability of the already existing algorithms.

The last scientific contribution of this thesis concerns a prospective analysis of the possible benefits and disadvantages of different state-of-the-art hippocampal replay-inspired RL algorithms in neurorobotics. Since the impact of different types of replay in neurorobotics scenarios has only recently started to be investigated, we test a model combining different RL replay strategies and test their interaction in different goal-oriented robotic navigation tasks, going from a pure theoretical simulation to a complete robotic experiment.

SORBONNE UNIVERSITÉ

Résumé

École Doctorale ED3C Cerveau, Comportement, Cognition

Docteur

Modélisation des réactivations hippocampiques en navigation spatiale avec la théorie de l'apprentissage par renforcement : une approche neuroscientifique et robotique

par Elisa MASSI

L'expérience obtenue par les interactions avec le monde qui nous entoure est le principal moyen par lequel les animaux et les humains apprennent. Comme cela a été étudié dans la navigation spatiale chez les rongeurs, le cerveau des mammifères peut réélaborer l'expérience passée de manière contextuelle, grâce aux réactivations des cellules de lieu dans l'hippocampe.

Récemment, la théorie de l'apprentissage par renforcement (AR) s'est avérée très efficace dans la modélisation de la navigation dirigée vers un but et des différents types de réactivations de l'hippocampe.

Dans cette perspective, notre première contribution scientifique concerne la conception et la validation d'un modèle d'exploration spatiale inspiré par des données comportementales de rongeurs. Nous avons identifié des caractéristiques comportementales communes chez les rongeurs dans un contexte d'exploration libre, et les avons modélisées sous la forme d'un modèle de prise de décision. En exploitant ces données comportementales, nous avons proposé un modèle décisionnel d'exploration général, où la prise de décision repose sur la sécurité perçue d'un lieu dans le labyrinthe et sur le coût et la persistance biomécanique de la dynamique exploratoire de l'animal. Enfin, nous avons validé l'adoption du modèle proposé sur deux nouveaux groupes de données comportementales de rongeurs et nous avons discuté les possibles améliorations futures du modèle pour mieux l'adapter à un plus large éventail de situations.

L'exploration libre correspond à un cas particulier du comportement exploratoire, où aucune condition externe n'affecte émotionnellement l'animal. Des recherches antérieures montrent que les réactivations hippocampiques représentent plus fréquemment des lieux à haut contenu émotionnel, mais il existe peu d'études traitant directement de l'apparition de réactivations hippocampiques à la suite d'un évènement aversif ou appétitif et qui comparent les mécanismes impliqués. Pour répondre à ce manque, notre deuxième contribution concerne l'extension du modèle d'exploration libre à la modélisation de ces mécanismes. En s'inspirant des modèles d'AR existants sur les réactivations hippocampiques, nous étendons notre modèle d'exploration libre avec un composant qui décrit la valence apprise et rejouée du conditionnement positif ou négatif. En exploitant des nouveaux données expérimentales, nous avons reproduit qualitativement la corrélation entre la quantité estimée de réactivations hippocampiques pendant le sommeil et l'occupation différentielle de la zone de conditionnement aversif, après et avant le conditionnement. De plus, elles soulèvent de nouvelles prédictions intéressantes, qui concernent la plus grande

importance des réactivations pendant le sommeil pour reproduire au mieux le comportement des souris après un conditionnement négatif, plutôt qu'après un conditionnement positif.

L'apprentissage des comportements dirigés vers un but est également essentiel dans la conception d'agents artificiels et de robots adaptatifs. En dépit de l'efficacité des méthodes d'AR pour l'apprentissage en ligne, celles-ci ne sont en général pas assez réactives pour répondre aux contraintes de la robotique réelle. De plus, avant même les premières études neurophysiologiques sur les réactivations hippocampiques, la recherche en apprentissage automatique proposait de rejouer l'expérience passée dans l'AR pour améliorer la vitesse d'apprentissage et l'adaptabilité des algorithmes.

La dernière contribution scientifique de cette thèse concerne une analyse prospective des avantages et des inconvénients en neurorobotique de différents algorithmes d'AR inspirés par les réactivations hippocampiques. Nous avons proposé un modèle qui combine différentes stratégies de réactivations d'AR et avons testé leurs interactions dans différentes tâches de navigation robotique dirigée vers un but, en partant de simulations purement théoriques jusqu'à arriver à des expériences robotiques complètes.

Acknowledgements

This thesis work has been financed and made possible by the CNRS 80 | Prime RHiPAR project.

First of all, I would like to thank the reviewers and the members of my Ph.D. jury, Prof. Lola Cañamero, Prof. Sen Cheng, Dr. Laure Rondi-Reig, and Dr. Nicolas Cuperlier, for taking the time to evaluate my manuscript and my work.

My most significant acknowledgment goes to my supervisors, Benoît and Mehdi, for having supervised and guided me all these three years. Most of the things I have learned through this Ph.D. come from your precise and technical feedback. Also, your expertise and positive attitude have always let me leave our meetings and discussions with new ideas and possible solutions that I was struggling to find myself.

I sincerely thank all of our collaborators because, without their precious work, this thesis would have lacked many important contributions. I want to thank Dr. Karim Benchenane for the behavioral data he provided us at the beginning and the end of the thesis, which constitute the primary dataset in our neuroscientific contribution. I also want to thank significantly our collaborators from KU Leuven, Prof. Sebastian Haesler and Eléonore Schiltz, and Prof. Michaël Zugaro, Raphaël Brito, and Linda Kokou, at College de France, who makes these exchanges and the following discussions extremely easy. Thank you for always being available to answer my questions and excited to discuss.

All the bachelor and master students who, through their internship, contributed to some of the presented results, were also an important part of my Ph.D. journey. I really enjoy supervising them, but in particular, working and discussing with them: it was very stimulating to work together, and I learned a lot about making my work more precise while discussing with you. Thank you very much to Juliane, Julien, Esther, Artem, Lakshwin, Fousseyni, Lydia, Léo, Laurine, and Ilke.

I want to thank all the people working at the ISIR, particularly the AMAC team. Even though the Ph.D. is meant to be a one-person adventure, it was a pleasure to share the most critical and joyful moments with the other Ph.D. students of the lab and not only: Nicolas, Alex, Ahmed, Elias, Maud, Jeanne, Johann, Giuseppe, Rémi, Jérémy, Yoones, Paul, Matthieu, Olivier, Charly, Lise, Leïla, Elias, Elias, Augustin, Rick, Nicolas, Hippolyte, Alessia, Astrid, Quentin, Douina, Emily, François, and the others. Despite the pandemic that forces most of us to spend a lot of time smart-working, I was really enjoying my time in the lab with you. From the moment I arrived when I did understand none of your "blagues", my will to be able to interact with you appropriately was a powerful drive to get a decent level of French, also. Thank you for often taking the time to help me when I needed it, professionally and personally. Together with the professors, all the researchers and students here are brilliant and passionate about their work, and it constantly motivated me to give the best I could.

Also, the friends I have here in Paris have been part of my Ph.D. life for these three years, and aside from having fun together, walking in the city, breathing art and culture in all its imaginable forms, they have always been very supportive of my work and happy to participate at the first occasion. Thanks to Stefano, Silvia, Michela, Natalia, Marianna, Andrea, Andrea, Selene, Sara, Amanda, Daniele, Natalia, Carlos, Julia, Sera, Rebecca, and the others. Thanks to my friends from outside Paris: even though we do not live close by, you are the first ones to have the energy and motivation to keep in touch with me and organize weekends, holidays, and meetings: in particular Martina, Marta, Francesco, and all the Bionics, Jenny, Chiara,

Giulia, Cesare, Andrea, Francesca, and all the others from Russi. Thank you all for sharing so much with me.

Thanks to Lorenzo, among all things, for always believing in me and encouraging me. I am glad you have jumped into this Parisian life “avec moi”. Your dedication and method in studying what you love have inspired me. Finally, thank you for the millions of small things you have helped me with, particularly for being a genius and having studied RL to understand my thesis. Thank you for accepting my fragilities, sharing yours, and always giving your best to stay by my side.

In the end, thanks to my family, in particular my sister, my parents, and my grandparents, who have supported me, not just concretely, but completely, through my educational journey and had never let me down for my reckless decisions. You taught me to be independent and prioritize my curiosity, even though that almost always meant being far from you. You taught me how to study when I was a kid and never missed the opportunity to listen to my small projects or presentations, even today. Thanks for your honesty, understanding, protection, and all the possibilities, freedom, and love you gave me.

Contents

List of Figures

List of Tables

Acronyms

MDP	**M**arkov **D**ecision **P**rocess
AI	**A**rtificial **I**ntelligence
DM	**D**ecision-**M**aker
RL	**R**einforcement **L**earning
MB	**M**odel-**B**ased
MF	**M**odel-**F**ree
NN	**N**eural **N**etwork
IGS	**I**nternally **G**enerated **S**equence
SWR	**S**harp **W**ave **R**ipple
MC	**M**eta-**C**ontroller
PFC	**P**re**F**rontal **C**ortex
RBPF	**R**ao-**B**lackwellized **P**article **F**ilter
SLAM	**S**imultaneously **L**ocalization **A**nd **M**apping
NSGA	**N**on-**D**ominated **S**orting in **G**enetic **A**lgorithms
ReAL	**R**einforcement **A**active **L**earning
CMA-ES	**C**ovariance **M**atrix **A**daptation - **E**volution **S**trategy
A-O	**A**ction-**O**utcome
DLS	**D**orso**L**ateral **S**triatum
VS	**V**entral **S**triatum
PAG	**P**eri**A**queductal **G**ray
OFC	**O**rbito**F**rontal **C**ortex
DMS	**D**orso**M**edial **S**triatum
BOLD	**B**lood-**O**xygen-**L**evel-**D**ependent
STDP	**S**pike-**T**iming **D**epended **P**lasticity
MEC	**M**edial **E**ntorhinal **C**ortex
ADN	**A**ntero**D**orsal thalamic **N**uclei
TD	**T**emporal-**D**ifference
RPE	**R**eward **P**rediction **E**rror
VPI	**V**eward **P**rediction **E**rror
ROS	**R**obot **O**perating **S**ystem
VTE	**V**icarious **T**rial and **E**rror
MO-EA	**M**ulti-**O**bjective **E**volutionary-**A**lgorithm
LIDAR	**LI**ght **D**etection **A**nd **R**anging
PC	**P**lace **C**ell
GC	**G**rid **C**ell
IMU	**I**nertial **M**easurement **U**nit
TaVAR	**T**able lumineuse pour **V**ulgariser l'**A**pprentissage par **R**enforcement
MemR	**Mem**ory **R**eactivations
SimR	**Sim**ulation **R**eactivations
EC	**E**ntropy and **C**ost
ITI	**I**nter **T**rial **I**nterval
NREM	**N**on-**R**apid **E**ye **M**ovement

To all the experiences and people who inspired me,
to my friends,
to Lorenzo,
and, most of all, to my family

Chapter 1

Introduction

> *In the practical use of our intellect, forgetting is as important as remembering.*
>
> William James, *Psychology: Briefer Course*, 1984

Animals and humans continuously learn through their interaction with others and the surrounding environment. They are also very efficient when learning from their past experience, particularly from unexpected mistakes. This learning efficiency is desirable and often required to survive in harsh conditions and emerge in situations where resources and possibilities are limited. To cope with this natural need, evolution has provided many creatures with neural mechanisms that can exploit what they have already experienced to possibly learn faster from their past.

When animals encounter new environments and situations, many factors can influence their behavior. In the absence of significant exogenous stimuli, this exploratory behavior is mainly driven by internal factors such as their current level of motivation, anxiety, or curiosity. Thus, this animal behavior is usually referred to as a *free exploration* of the environment. Clearly, these factors may change when rewarding or adverse stimuli are introduced in the scene, and the animal modifies its decision-making mechanisms accordingly, typically to maximize the reward and to minimize adverse situations. To computationally understand these mechanisms, behavioral neuroscientists are interested in modeling how mammals make decisions while exploring new environments and studying if the behavioral patterns at the base of these decisions are consistent across different environments, timescales, and species.

Then, how do these behavioral attitudes change when unexpected variables are introduced in the environment? To survive, animals need to learn a proper behavior through very few interactions with a positive event. In case of an adverse event (*e.g.,* injury, a threat by a predator), this is even more important in order to preserve the agent's physical integrity and increase its chances of survival. More than one century ago, with his famous experiment, Pavlov remarked on a solid ability for instrumental conditioning and learning in mammals. From that point, research efforts have been focused on investigating which parts of the mammalian brain cooperate to generate such instrumental and associative learning mechanisms and which neural strategies made them possible.

One of the most popular theories came from Tolman (1948) who suggested that, among other brain structures, the hippocampus was the one specialized in encoding the existing relationships between elements of the outside world and the agent experiencing them and eventually building *cognitive maps*. How animals and humans

organize, elaborate and abstract from the noisy and redundant information sensed in the world is essential to efficient learning.

From the seminal works of Scoville and Milner (1957) and Pavlides and Winson (1989), the role of the hippocampus as a center for cognitive mapping has been deeply studied in neuroscience, in particular through neurophysiological studies in rodent spatial tasks. This research shed light on the fact that some patterns of sequential activations of hippocampal neurons, observed during task performance, are later replayed during either sleep or periods of quiet wakefulness of the animal. This activity was called *hippocampal reactivations* and recognized as a powerful mechanism used, in particular by place cells, to recall, organize and consolidate past experiences and infer future ones.

From a machine learning point of view, it has also been known that performing off-line replay of learning algorithms enables accelerated learning, following a small number of real-world interactions with rewarding or punishing events in the environments (*e.g.,* Lin (1992)). Since the first works in the field of neurorobotics (interdisciplinary field between neuroscience and robotics), such as the one from Miyamoto et al. (1988), computationally understanding and reproducing exploratory and learning mechanisms from humans and animals has been vital in designing agents and robots that could adapt to the noisy, dynamic and rich interaction with the outside world. Nevertheless, adding replay strategies to learning algorithms poses important challenges for robotics, especially concerning what is called *the reality gap*: the challenge of transferring a robotic behavior from simulated environments to real ones. In fact, a robot's replay or mental simulations can produce very different results than performing these actions in the real world. So, how can we design efficient replay algorithms that minimize the reality gap in robots? And potentially, does the biological brain face the same problem? If yes, which neural mechanisms have been selected through evolution to solve this problem? These questions suggest that designing computational strategies that can easily transfer to real-world scenarios is one of the main challenges of bridging artificial intelligence, neuroscience, and robotics.

The great interest in implementing computational strategies inspired by hippocampal reactivations in robotics lies in tasks in which past experience and acquired knowledge should be re-evaluated and re-elaborated to perform better during future decision-making steps. This is the case of reinforcement learning paradigms, where at the beginning, when no prior knowledge is usually available, the best strategy is to interact within an environment by trial and error, and only when the level of experience increases the agent can exploit its previous knowledge to reach a sequence of actions that is closer to the optimal behavior for that situation. In mammals, this knowledge consolidation is not only due to the animal performing the same actions in the same situations: memories, in particular, targeted experience recalls, are fundamental for efficient learning from a small batch of episodes.

Recalling past experience through hippocampal reactivations has been observed to have many functions, such as memory consolidation and spatial learning (Girardeau et al., 2009; De Lavilléon et al., 2015; Foster, 2017; Ólafsdóttir, Bush, and Barry, 2018), but how to develop efficient computational strategies to replicate these functionalities is still a matter of debate, in particular when the agent that needs these strategies is a mobile robot, with low computational power and battery life. While in simulation, strategies that accumulate reward faster are usually unreservedly chosen, real robotics imposes new constraints which push the investigation towards the inspiration from neural strategies that can better juggle the compromise between learning accuracy and computational cost. The proper identification of what should

be kept in memory for recall and what should not is fundamental to efficiently orchestrate past experience when time and resources are limited. Finally, what is also designating robotics as a field worthy of digging into for behavioral neuroscientists, is that it represents more than the implementation of autonomous machines with cognitive capabilities. By emboding, testing and, validating computational theories for adaptive behavior and intelligence in interaction with the real world, robotics constitutes a compelling technology also for behavioral neuroscience.

1.1 Scientific questions

This thesis focuses on studying different aspects around the topic of hippocampal reactivations and their role in spatial navigation tasks, both in computational neuroscience and robotics. In particular, our research addresses three main scientific questions:

- **Is it possible to interpret rodent decision-making during free exploration in a novel environment through a combination of value functions?** We want to examine the main characteristics that can influence rodent free exploration. Then, we investigate if a decision-making mechanism that assigns a value to these characteristics is able to produce a simulated behavior closer to one of the real animals than the one produced by a random decision-maker.
- **How does opposite valence conditioning, *i.e.*, reward and punishment, influence hippocampal reactivations?** Here, we are interested in optimizing our proposed reinforcement learning-based spatial exploration model on mice behavioral data to qualitatively predict the replay activity during opposite valence conditioning.
- **Which replay mechanism based on reinforcement learning would be better suited for dynamic goal-directed robotic navigation tasks?** We want to investigate which navigation mechanism based on reactivations is more appropriate while passing from theoretic simulations to real robotic experiments and trying to predict which computational strategies better explain goal-directed spatial navigation in different conditions.

This manuscript is the scientific report of our research process to answer the above questions. From our current knowledge of different fields, such as behavioral neuroscience, computational modeling, and neurorobotics, we propose an interdisciplinary approach to address these scientific questions transversely. All these questions together aim at a deeper understanding of mammals' behavior and learning in spatial navigation through the lense of reinforcement learning (RL) and the cooperation between computational neuroscience and robotics. The common value-based RL framework allows, on one side, for an easy transfer of the model designed for describing rodents' free exploration and learning (first and second scientific questions) to artificial agents and robots in similar spatial tasks (third question). On the other side, it facilitates new neuroscientific hypotheses about the role of hippocampal replay in spatial learning, not only by the analyzing the animals' behavior, but also by analysing the best strategies that contribute to learning efficiency and adaptability in robotic goal-directed tasks.

1.2 Thesis development

After this introduction section, the manuscript presents a first macro-section about the state-of-the-art and related works to contextualize and insert the scientific contributions of this thesis in the literature (Chapter 2). This background review is divided into three sections: Neuroscience (Sect. 2.1), Computational modeling (Sect. 2.2), and Neurorobotics (Sect. 2.3).

We are going to tackle different scientific questions in this thesis (Sect. 1.1), covering from behavioral neuroscience (Chapter 3) to robotics (Chapter 4). We pass from one domain to the next one by consistently relying on the common modeling sequential framework of discrete Markov Decision Processes (MDPs), Value-based models, and discrete tabular Reinforcement Learning (RL) (Sect. 2.2.1).

The sections that follow gather the scientific outputs of the thesis (Chapters 3-4). Firstly, we present the design of a new value-based decision-making model for rodent free exploration and its evaluation against three different datasets (Sect. 3.1.2). Secondly, we show the adaptation of the same model to evaluate and predict the contribution of replay-like mechanisms in opposite valence spatial learning (Sect. 3.2). Also in this case, we based our predictions on behavioral data. Thirdly, analyses of the impact of different reinforcement learning replay-inspired mechanisms in neurorobotics are presented (Sect. 4.1). Here, we present the results published in Massi et al. (2022) and the design and development of an immersive robotic demonstration on the role of reinforcement learning-based reactivations in robotic navigation (Sect. 4.2).

Finally, we discuss the contribution of the presented results (Chapter 5) by describing the possible limitations and the future perspectives derived from this thesis (Sect. 5.2).

1.3 Collaborators and contributors

The results presented in this thesis would not be possible without the exchanges and discussions we had with our collaborators and the work conducted by bachelor and master students supervised by us during part of the period when this thesis was conducted. The author intends to acknowledge the contributions of our collaborators and students concerning the different sections of this thesis.

Regarding the free exploration model in rodents, the first design was driven by behavioral data (Sect. 3.1.1) we had through the collaboration with the research team of Karim Benchenane at ESPCI Paris and PSL University, in the context of the CNRS RHiPAR project which funded this thesis. Moreover, preliminary work and results on the design and validation of this proposed model have been started by Eléonore Schiltz, and Artem Dobrosmyslov did some preliminary analyses during their master internships in the lab. Then, Eléonore Schiltz and her current Ph.D. supervisor Sebastian Haesler, from KU Leuven, exchanged with us some of their behavioral data to enrich the validation process of our free exploration model and to generalize our hypothesis on new data (Sect. 3.1.1). Finally, during the last year of the thesis, more behavioral data (Sect. 3.1.1) were also accessible to us for validating the model thanks to the team of Michaël Zugaro at College de France, and in particular to Raphaël Brito and Linda Kokou, who are the Ph.D. students who directly set up the experiments and recorded these data.

The results obtained in Sect. 3.2 were then possible thanks to new data received from Karim Benchenane's team during the last months of the thesis.

Finally, Jeanne Barthélémy, Juliane Mailly, Esther Poniatowski, and Julien Canitrot contributed to implementing part of the code for Massi et al. (2022) and did some preliminary analyses on the results. Next, Mehdi Khamassi worked on the simulations, results analysis, and writing of Sect. 4.1.2. Then, Lakshwin Shreesha and Fousseyni Sangaré contributed to the execution of some robotic navigation experiments for their master internships, and Lydia Gaillot, Léo Laval, Laurine Le Peutit, and Ilke Tuzun built the immersive demonstration table and part of the visualization program for their university project (Sect. 4.2).

Chapter 2

Background and related works

The research results presented in this thesis cover and merge knowledge from different domains of science. In this section, the literature concerning the topic of rodent navigation, the neural circuitry involved in instrumental behavior, spatial representation, and learning, is analyzed from the three main perspectives that are then tackled in the scientific contributions of this work: neuroscience (Sect. 2.1), computational modeling (Sect. 2.2) and neurorobotics (Sect. 2.3).

2.1 Neuroscience

While talking about spatial navigation and learning, different levels of study and abstraction can be employed.

Firstly, it is important to consider the previous researches that have questioned the relevance of particular behavioral traits in animal spatial navigation. In Sect. 2.1.1, we are going to analyze the main insights about behavioral patterns in animal spatial exploration. Indeed spatial exploration has been extensively studied in rodents, in the last 50 years.

Secondly, it is relevant to look into the decision-making processes that generate learning and adaptive instrumental behavior. Sect. 2.1.2 summarizes the studies on this subject, in particular regarding the dichotomy between two distinct systems for guiding action selection; a *habitual* and a *goal-oriented* one.

Thirdly, we are going to describe the current understanding of the brain circuits and mechanisms which process spatial information. Sect. 2.1.3 reviews the state-of-the-art studies about the role of the hippocampal-entorhinal circuitry in self-localization and orientation in space.

Finally, understanding the neural mechanisms that make spatial adaptation and learning possible is key to transferring these principles to a higher behavioral level. In Sect. 2.1.4, we provide a summary of the principal brain structures that are involved in spatial learning, as well as the mechanisms that underlie spatial memory.

2.1.1 Spatial exploration in rodents

Exploration has always played a key role in human and animal survival. Being able to safely navigate unfamiliar places and to optimize energy to collect food discriminates which species are more likely to sustain and then prevail over others.

With the use of the term *exploration*, we are particularly referring to a spatial navigation behavior where the main trigger is novelty (Belzung, 1999). Thus, we will typically use this term when describing an animal or a human that experiences an environment for the very first time and whose behavior is strongly guided by novelty and surprise (Berlyne, 1950).

The interest in understanding the leading motivations which guide navigation in a novel environment has developed decades ago when novelty already was thought to be a salient factor for exploration. During the last 20 years, some researchers have focused their work on similar questions, in particular by studying spatial exploration in rodents, like mice and rats.

Rodent spatial behavior has been studied in different types of environments, such as open fields (Walsh and Cummins, 1976), mazes with corridors, such as Y-mazes (Peyrache et al., 2009), and hole-board apparatus (G. R. Brown and Nemes, 2008). Here, we focused our attention on the so-called *free exploration*, by discussing the exploratory dynamics in contexts where the animals are navigating mazes without any constraints or exogenous stimuli from the experimenter.

A crucial distinction that can be observed when going through past studies about free exploration concerns the presence or not of a departure home cage and whether the animal is given the possibility to come back there or not. When rodents cannot return to a familiar safe location, they are much more active and this level of hustle subsequently decreases within experimental sessions (Welker, 1957). On the contrary, if a very familiar location, such as a home cage, is present, rodents are less active and tend to come back there very often during the beginning of their exploration, because they perceive it as the safest accessible place.

Extensive work in modeling rodent exploration in novel open-field environments has been carried out by the research group of Golani at Tel Aviv University. Since the early nineties, their research has progressed and enriched the knowledge in the field, by disclosing the existence of some common patterns in a free exploratory behavior that seems often haphazard.

The first relevant feature of free exploratory behavior in rodents is that places that are considered safe, such as home-cages for instance, stand out from the rest of the environment as attraction points in exploration. Further, Golani, Benjamini, and Eilam (1993) identified the fact that when one or more home cages are present in the maze, the attraction to their location increases faster when the number of the animal's stops in its "excursion" increase. Usually the animal's stops during excursions does not exceed an intrinsic upper bound, which interestingly does not scale with the size of the explored area. Thus, rats' exploratory behavior can be split into *excursions* (paths covered by the animal between two visits of its home base places, Golani, Benjamini, and Eilam (1993)) that often present a low-velocity profile and an intermittent progression while the animal is moving away from the home bases and instead high-velocity profiles while it is going back to them (Tchernichovski and Golani, 1995).

A second important characteristic of free exploratory behavior in rodents is the fact that if the animal is modeled as a moving point in space, then the dynamics of this point should be described as an alternation between moving and stopping bouts (Tchernichovski and Golani, 1995). As long as the exploration of a novel environment proceeds, rats unravel new reference places (home bases) that gradually become more connected in time (Tchernichovski, Benjamini, and Golani, 1996).

Finally, another relevant aspect is that the dynamics of free exploration in rodents develop in a way that the animals feel safer and safer and free to navigate far away from the maze walls and home bases and cover complex trajectories and movements. As shown in Fig. 2.1, the dynamics of the behavior of their twelve studied mice can be decomposed into twelve *landmark motions* that describe a gradual passage of the animal into a broader and more unconstrained navigation of the space, in an about 3-hour period. The animal starts the exploration by periodically peeping its nose beyond the home cage (1) and then by completely exiting the safe cage and

immediately going back (2). Then it starts to move in a circle close to the departure spots (3) and still comes back inside the cage (4). The first roundtrip excursions are very short strictly close to the walls, and in the same direction out of the cage (5) and become progressively longer (6). Finally, the mice begin doing roundtrips in the opposite direction out of the home cage (7) and finally, a full circle, rigorously close to the wall can be accomplished (8). Later, the trajectories become more complex and they start with a brief excursion towards the center of the maze (9) and go for some border-related shuttles (10). This gradual process of interaction with novelty and new information to process in the environment ends when the mice's trajectories finally reach the center of the open field maze (11) and then they try to jump out of the maze (12).

The ordered coverage of these landmark motions is consistent for 5 over 12 mice studied in the research, during an unfettering exploration. The remaining six mice, whose exploration was not completely consistent with the proposed 12 *landmark motions* would need from 1 to 3 swaps of adjacent motions to replicate the exact sequence (Fonio, Benjamini, and Golani, 2009).

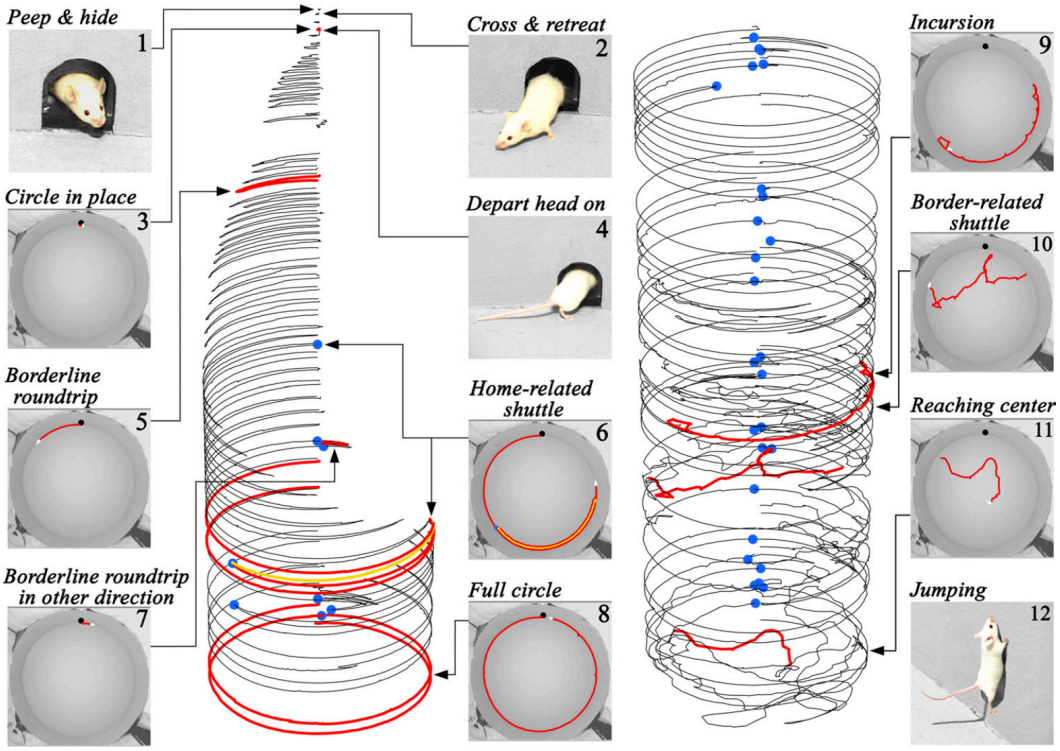


FIGURE 2.1: Illustration of the 12 *landmark motions* identified in the work from Fonio, Benjamini, and Golani (2009). The 2D evolution of the navigation of the animal is shown with a black line. The red demarcations better highlight some of the landmark motions, while the yellow ones indicated Home-directed-shuttle. Finally, blue dots refer to turns when the animal stopped before going back to the home cage. Figure reprinted from Fonio, Benjamini, and Golani (2009).

All these observations contribute to the current state-of-the-art knowledge on free spatial exploration in rodents, and the majority of them have reported results from experiments where rodents had explored circular open-fields environment.

These studies have been able to show that, without any particular constraint, rodents tend to have some common persistent behavioral features in free exploration (Drai et al., 2001) and that they deal with novelty, safety spots, and dynamics of space coverage. In this thesis, we suggest that similar common patterns can be found even when rodents explore different types of mazes and for various temporal durations, proposing a new framework for the description and modeling of diversified free explorations.

2.1.2 Instrumental behaviour in the brain

Rodent navigation does not concern just free exploration, but also moments where the animal is goal oriented (*e.g.,* navigating towards a desired food location). These goal-oriented behaviours are usually referred to as *Instrumental behaviours*. They consist in a series of actions that are undertaken to reach a desired goal, either to get a reward or to avoid a punishment (Fragaszy and Liu, 2012). That animals can perform instrumental behaviour was demonstrated by Hammond (1980). They showed that animals were receptive about the causal relationship between their action and the subsequently obtained reward. Dickinson (1985) was then the first one to hypothesize the existence of two general behavioural modes for animal instrumental behaviour. The first mode is close to a direct and reactive relationship stimulus-response and can be described as *habit*. The second one, instead, is *goal-oriented* and assumes a more complex knowledge on the relation between actions and following consequences. The existence of this second goal-directed model has been observed first in rats, where their response to a lever-press task is modulated by reward alteration (Adams, 1982).

After several behavioural and neural studies supported the presence of this duality between an habitual and a goal-oriented model for instrumental choices (among them, Dickinson and B. Balleine (1994), Lieberman et al. (2002), and Dickinson and B. Balleine (2002)). Daw, Niv, and Dayan (2005) were the first to investigate which kind of computational mechanism should then be in place to arbitrate their coexistence. They proposed that the arbitration between such a duality of action control systems can be based on the reliability of their predictions.

Thus, neurophysiological lesion studies from fifty years ago have already identified that, basal ganglia and, in particular the striatum, are crucial areas for instrumental conditioning (Konorski, 1967; B. W. Balleine, Delgado, and Hikosaka, 2007), and thus for everything that has to deal with the control of voluntary behavior (Yin and Knowlton, 2006). In rodents, on the one hand, the dorsolateral striatum (DLS) has been proven to be necessary for habit formation (Yin, Knowlton, and B. W. Balleine, 2004) and its inactivation enhances sensitivity to changes in the Action-Outcome (A-O) contingency in rats (Yin, Knowlton, and B. W. Balleine, 2006). On the other hand, the dorsomedial striatum (DMS) has been identified as supporting center for goal-directed behavior (Yin, Knowlton, and B. W. Balleine, 2005) hypothesize that this functional difference comes from the diversity in types of plasticity present in these areas (Partridge, Tang, and Lovinger, 2000) which lead to a difference in computational learning rules. Yin et al. (2009) showed also that the activity of these two areas (*i.e.,* DMS and DLS) change through the task learning phase: at the beginning, the DMS is more active than the DLS, suggesting that goal-directed behaviour is preferred in this phase, while towards the end, the DLS, prevails, representing a higher contribution from habitual behavior in this last phase. Finally, the prefrontal cortex (PFC) represents the brain areas devoted to high cognitive functions which can orchestrate different type of behaviour and associate them to particular contexts

and goals (Koechlin and Summerfield, 2007; O'Doherty et al., 2021). The PFC has the role of switching among models or task-sets, depending on the external context, while the characteristics of these models are modulated and learned by the basal ganglia (Daw et al., 2011b; Gläscher et al., 2010). Either the DLS (habituation system) or the DMS (goal-oriented system) could be selected based on their reliability on their predictions, making the inferior lateral prefrontal and frontopolar cortex inhibiting the habitual system when the goal-oriented one is preferred in rodents (S. W. Lee, Shimojo, and O'Doherty, 2014). In their review, Dolan and Dayan (2013) underline the fact that even if four generations of research works have been published on the cognitive dichotomy of reflective versus reflexive decision making, it is not yet clear at which level model-based and model-free control are intertwined, in particular in humans. On one hand, Blood-oxygen-level-dependent (BOLD) signal in the human orbitofrontal cortex was showed to be modulated by outcome devaluation, arguing a role of the area in goal-directed decision-making (Valentin, Dickinson, and O'Doherty, 2007). On the other hand, overtrained human subjects showed increased BOLD signals in the right posterior putamen/globus pallidum compared to earlier sessions of the training. This denotes that that these areas are habituation-related and this analysis is consistent to the results previously found in rodents (Tricomi, B. W. Balleine, and O'Doherty, 2009).

The study of habitual versus goal-oriented behaviour has also been studied for specific tasks in rodents. Khamassi and Humphries (2012) underlined the fact that the interaction between these two decision-making systems is also very relevant in spatial navigation. In this view, an habitual behaviour would correspond to a *cue-driven* and *map-free* exploration, while a goal-directed behaviour would correspond to a *place-driven* and *map-based* strategy. The model they propose that suggest specific computational functions to the neural substrate of basal ganglia, hippocampus, amygdala and other areas in the cortex and in the pedunculopontine nucleus will be illustrated in Sect. 2.2.2.

They also suggested that ventral striatum is responsible for building the model of a rewarding interaction with the environment (Joel, Niv, and Ruppin, 2002; Yin, Ostlund, and B. W. Balleine, 2008) and to learn important stimulus-response associations that constitute a model between actions and reward (Lansink et al., 2009).

More recently, Daw et al. (2011b) found that human ventral striatal BOLD signal seems to responds either to model-free and model-based predictions, suggesting that the computation alternating these two learning systems could be more combined than expected.

These results concerning the coexistence of two behavioral systems, *i.e.,* habit-oriented and goal-oriented, shared important insights on animal and human decision-making and learning processed. Looking at the computational principles behind them, we could be able to better explain how animals explore and in particular adapt to different situations, and in the case of this thesis, to different spatial cues and environments.

2.1.3 Spatial mapping in the brain

During the last decades, all the literature related to spatial navigation has been progressively involving the concept of spatial memory and the role of the hippocampus in navigation.

Spatial memory is defined as the ability of an organism to know or to have a representation of where it is in a particular environment and thus to be capable of effectively navigating in there (C. Barnes, 1988). One of the first pieces of evidence

that spatial memory exists arrived when Morris (1981) demonstrated that rats can rapidly learn to locate an object that they cannot sense as long as it remains in a fixed spatial location. The research implied the use of a circular pool, around 10 times the size of the animal, filled with opaque water. The rats were given the possibility to escape water by climbing over a platform and they were able to rapidly localize it either if it was above or below the water's surface. The same thing happened also if the rats were introduced to a different area of the pool, showing a generalization capability related to allocentric spatial memory. The only case when the localization of the platform was slower than usual were the ones where the underwater platform was moved to a new location. In this latter case, 5 over 6 animals showed a strong searching strategy around the previous location of the platform.

The fact that rodents can efficiently navigate in a known environment without the need for sensory clues, such as visual markers, familiar odors, and so on, is possible thanks to neural hippocampal formations that serve as cognitive maps (O'Keefe and Dostrovsky, 1971; O'Keefe and Conway, 1978).

Different types of cells in the hippocampal-entorhinal circuit play a special role in declarative memory formation and in encoding relevant spatial information in both rodents and humans (M.-B. Moser, Rowland, and E. I. Moser, 2015; Alkon et al., 1991):

- **Place cells**: they respond specifically to the current location of the animal *i.e., place field*, (O'Keefe, 1976) and their combination of activity is unique for a specific environment (O'Keefe and Conway, 1978). Interestingly, the spatial organization of the place fields and the anatomical location of the corresponding place cells do not seem to have any topographical relationship (O'Keefe et al., 1998). Fig. 2.2 shows the strong spatial selectivity of a group of pyramidal cells recorded simultaneously in the hippocampus of a rat. The represented 35 place fields seem to collectively cover almost the whole available space (O'Keefe et al., 1998).

- **Grid cells**: they were identified in the medial entorhinal cortex (MEC, in Fig. 2.3-A) and express geometrical and directional information about the animal location (Hafting et al., 2005). Their activity significantly increases whenever the animal is placed in one of the vertices of a regular grid of equilateral triangles superimposed on the environment (Fig. 2.3 C, top). These *grid fields* increase in size from the dorsal to the ventral entorhinal cortex and their activity persists even in the absence of stable external markers.

- **Head direction cells**: their firing rate increases selectively when the rodent's head is facing a certain preferred direction which varies among cells and is independent of the animal's location and behavior (Taube, 2007). They were first discovered by Ranck Jr (1984) to be present in rat's presubiculum (Taube, Muller, and Ranck, 1990a; Taube, Muller, and Ranck, 1990b), and then their presence has been remarked also in the anterodorsal thalamic nuclei (ADN) (Taube, 1995), in the lateral mammillary nuclei (LMN) (Stackman and Taube, 1998) and in the retrosplenial cortex (Chen et al., 1994). Eventually, neurons showing this specific behavior have been found in deeper layers of the medial entorhinal cortex (Sargolini et al., 2006).

- **Border cells**: they also are in the entorhinal cortex and their activity is related to the proximity of the animal to the borders of the environment and they are orientation-specific (Solstad et al., 2008). They are thought to play an important

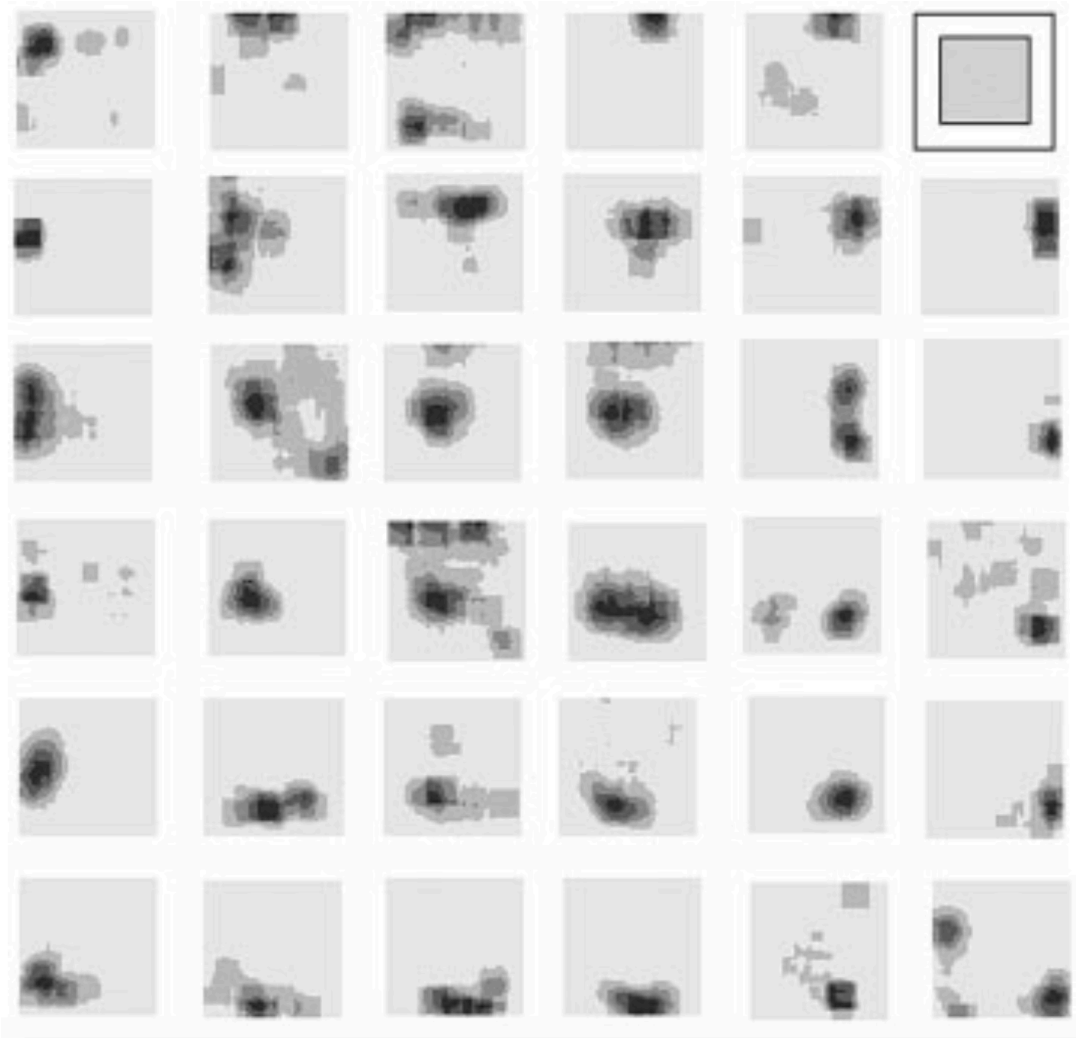


FIGURE 2.2: Location-related activity of 35 simultaneously recorded hippocampal place cells O'Keefe et al. (1998). The activity of the cells is ordered to spatially represent the animal's location in the 40cm x 40cm open platform, while it is searching for grains of rice. The four gray scales represent the place fields firing rates (each gray shade depicts 20% of the peak firing rate). Figure reprinted from O'Keefe et al. (1998).

role in consolidating the structures of place and grid fields to an environment-related reference frame.

A description of a typical activation pattern of place cells and grid cells is shown in Fig. 2.3. Grid cells are located in layer III of the MEC and project their axons to the CA1 in the hippocampus, to place cells (Fig. 2.3 B). This suggests that grid cell neural input is crucial for initiating place cells' activity.

Thanks to all the studies that have been carried out in the last decades (Stensola and E. I. Moser, 2016), we can affirm today the importance of the hippocampal-entorhinal circuitry (Fig. 2.3 A) in memory and space-related information processing, even in episodic memory *mental travels* (Buzsáki and E. I. Moser, 2013). Nowadays, the computational mechanisms which regulate these phenomena have reached a deep level of understanding and consensus in the community.

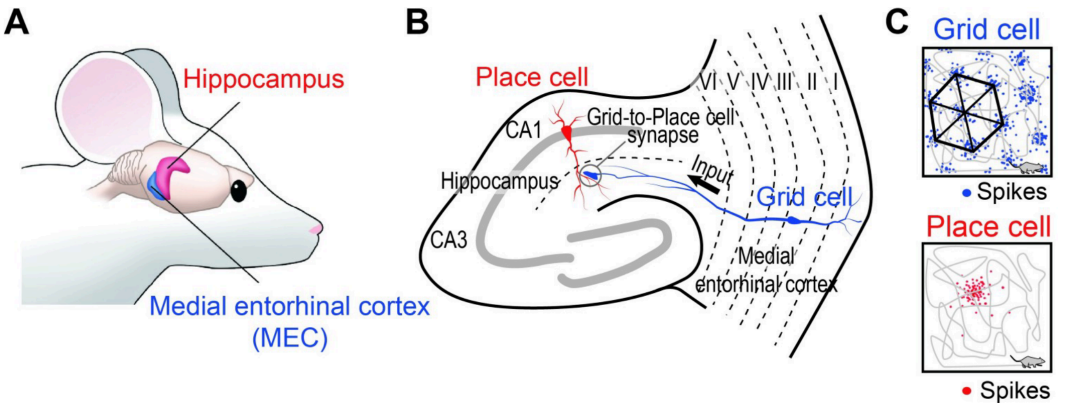


FIGURE 2.3: Hippocampal-entorhinal circuitry: grid and place cells. A) Anatomical representation of the hippocampus and medial entorhinal cortex (MEC) in rodents. B) Anatomical representation of the Grid-to-Place cell synaptic communication circuitry. C) Examples of a spiking pattern for a grid cell and a place cell where the rodent occupies a particular position in space. Figure reprinted from Park et al. (2019).

2.1.4 The role of hippocampal replay in spatial navigation and memory

A consistent understanding of the brain mechanisms underlying the representation of spatial information and self-location in space is not sufficient to explain how this information is stored and then retrieved when needed. From the first pieces of evidence that there exist cells in the hippocampus that spike according to the animal's spatial location, *i.e.*, place cells (O'Keefe and Conway, 1978; E. I. Moser, Kropff, M.-B. Moser, et al., 2008), and that lesions to the hippocampus strongly impair the formation of new spatial memory (Scoville and Milner, 1957; Morris et al., 1982), many more research studies have been conducted in this direction. The current consensus is that the hippocampus displays a particular activity pattern called *sharp wave ripples (SWR)*, at a frequency of 150-200 Hz, which is temporally compressed with respect to the timescale of the neural activity happening during the real spatial experience, and thus enhances spike-timing depended plasticity (STDP) (Dan and Poo, 2004). Hippocampal sharp wave ripples encode temporally structured spatial patterns and drive the initial storage and the later retrieval of relevant spatial experience (Pfeiffer, 2020). A very effective and simple example of the neural activity happening during hippocampal replay is shown in Fig. 2.4. On the top of the figure (Fig. 2.4 A) and highlighted in the analysis below (Fig. 2.4 B), we can see that the selective place cells that were active during the time-lapse when the animal covered the corridor, are reactivated, in a compressed timescale and in reverse order, when the animal reached the end of the corridor and stops there (the animal trajectory is the light blue line in Fig. 2.4 A bottom). This example is coherent with what has been found recently by Diba and Buzsáki (2007); studying replays in linear elevated tracks, they found that backward replay is usually elicited at the end of a one-way run while replaying forward sequences was more common before starting a new run. The bi-directionality of this phenomenon can indicate also a duality in the nature of hippocampal reactivations; backward sequences are more devoted to consolidating relevant and recent past experience and forward ones to deliberate and plan the next trajectories (Pfeiffer and Foster, 2013; Johnson and Redish, 2007). Different types of replay can be also

triggered by task demands; Ólafsdóttir, Carpenter, and Barry (2017) showed that the content of task-focused replay is more related to planning and subsequent spatial decisions when the animal is active on the task, while it is more synchronized with grid cells activity during rest periods, and thus oriented towards memory consolidations.

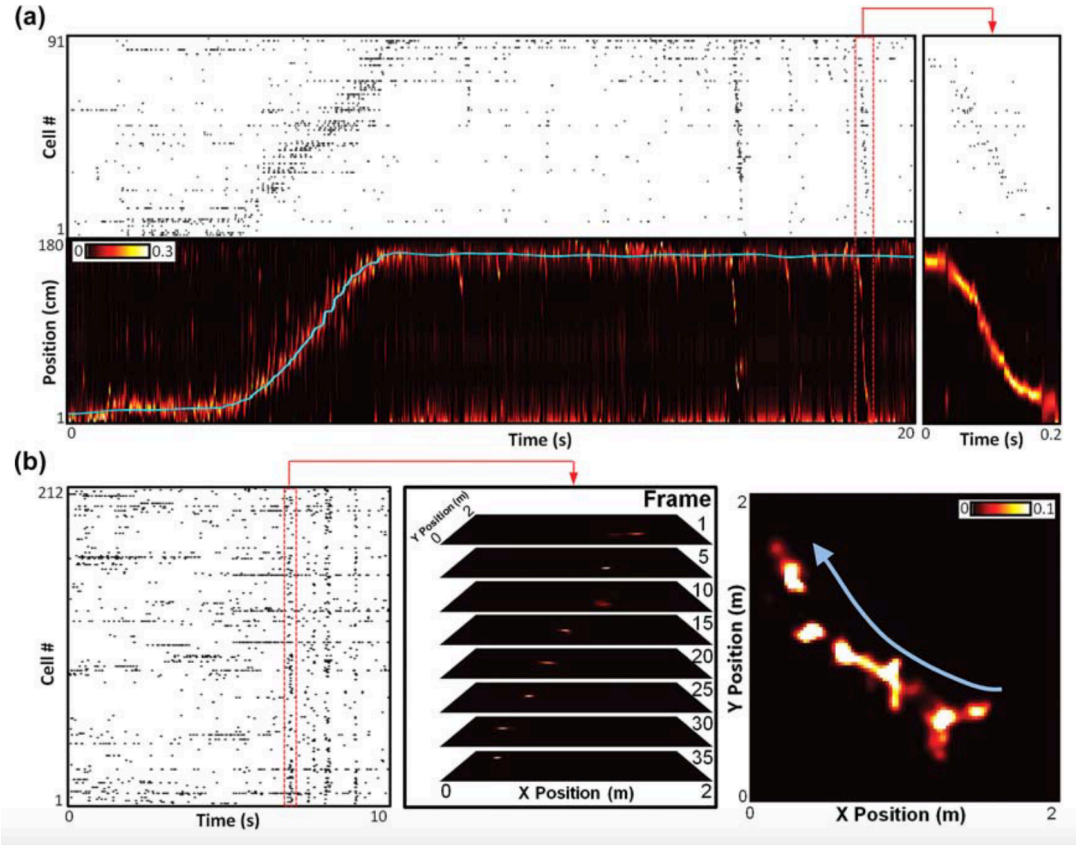


FIGURE 2.4: Example of recorded reverse hippocampal replay in rats. A, top) The raster plot representing the activity of 91 simultaneously recorded place cells in the animal CA1 hippocampus. The cells are ordered based on the location of the place field peak. A, bottom) the heat map represents the rat's estimated position on the linear track, based on Bayesian decoding of the spiking activity above. In this time interval, the animal is covering the linear track from one end to the other (the position of the animal is shown here with a light blue line). On the right, we can see expanded windows on the reverse replay activity, once the animal reaches the end of the track. B) Example of the decoding of an open-field replay. From the left to the right, we can look at the raster plot on the activity of 212 simultaneously recorded place cells, then the Bayesian estimation of the place fields for a selected ripple in the raster plot, and finally the reconstructed replay trajectory from the temporal frames to the maze. Figure reprinted from Pfeiffer (2020).

Further studies have brought to light even more diversity in hippocampal reactivation; such ripples can happen at relevant moments during the active exploration (Foster and M. A. Wilson, 2006), but also during inactive time intervals and sleep (M. A. Wilson and McNaughton, 1994; A. K. Lee and M. A. Wilson, 2002). This distinction between *awake* and *sleep* replay has been associated with a two-step memory

consolidation process that enforces versatile memory during day time and stabilizes long-term memory during night (Buzsáki, 1989). Also, Maingret et al. (2016) has recently validated the hypothesis that sleep hippocampal sharp wave-ripples, coordinated with prefrontal cortical delta waves and spindles, causally contribute to memory consolidation processes in rats.

To consolidate the hypothesis that hippocampal replays are not exclusively happening for long-term memory consolidation, (Gupta et al., 2010) demonstrated, in a double T-maze task, that they are not just representing past experience, but they can also encode possible environment-related contents that have never been experienced before by the animal, generating the so-called *imaginary* replay.

In addition, hippocampal SWRs contribute to spatial learning. Their suppression during sleep phases, which happen just after the task, leads to a performance impairment during spatial reinforcement learning tasks (Girardeau et al., 2009). De Lavilléon et al. (2015) managed also to create fictitious memory for a location preference by positively stimulating the animal during sleep, simultaneously with the SWR appearance for the place cell representing that location. An overall view of the acquired knowledge and current questions on the role of hippocampal replay is provided thanks to the works by Foster (2017) and Ólafsdóttir, Bush, and Barry (2018).

Interestingly, memory consolidation processes happening through replay, are often triggered by novelty (Cheng and Frank, 2008), recency (Foster and M. A. Wilson, 2006) or saliency (Singer and Frank, 2009), suggesting that some distinctive experiences could be replayed more or at the cost of others. On another interesting note, hippocampal reactivation can already appear after a single episode and reinforce with experience, by slowing down the reactivation of the same trajectory for encoding a greater level of spatial details (Berners-Lee et al., 2022).

Although replay phenomena happen in other parts of the brain, namely in the motor cortex (Dave and Margoliash, 2000), in the prefrontal cortex (Euston, Tatsuno, and McNaughton, 2007), and in the visual cortex (Ji and M. A. Wilson, 2007), the ones occurring in the hippocampus can thus be described as one of the primary neural mechanisms that allow adaptive spatial behavior and spatial learning. Yet, the idea that the hippocampus is a pivotal center for the creation, consolidation, and retrieval of cognitive maps, knowledge abstraction, and inference (Tolman, 1948) keeps getting stronger evidence recently, extending from rodents to human models (Behrens et al., 2018). Stella et al. (2019) found that the length and the timescale of replayed trajectories approximate the Brownian motion of particles, showing that this reactivation was not encoding the exact recent experience of the animal, but random trajectories, from the map, recently created on the environment by the animal. That could extend the importance of the replay phenomenon to contextual memory, generalization, and learning, beyond the spatial context and across multiple domains.

2.2 Computational modeling

Building computational models is what allows humans to get closer to a full understanding of many natural processes and animal and human behaviors, and then to be able to emulate these processes as well. A computational model describes a particular phenomenon in one or more equations, depending on its complexity. One of the main advantages of computational modeling consists in the description and explanation of a large amount of data by a precise and synthetic mathematical formulation. Another important benefit of having models for describing a certain natural

phenomenon resides in its potential for prediction; in fact, by looking at the results of the dynamics from the model, researchers can predict what would be observed in a real experiment. For example, lesion studies can be simulated in brain models by removing the connections among the areas of interest (Dollé et al., 2018) or changing environmental conditions in a robotic experiment can be simulated by changing the dynamic contact friction in physical robotic simulation environment (Massi et al., 2019). If the reader is interested in knowing more about the bases of computational modeling, an exhaustive introduction, with open-source code is proposed in French by A. Collins and Khamassi (2021).

This section summarizes the state-of-the-art about computational modeling in the subjects already covered in the previous section (Sect. 2.1). This literature represents the research niche to which this thesis contributes the most.

The computational framework which is the basis of the main contribution of this thesis is Reinforcement Learning (RL). RL represents a great modeling framework for explaining learning processes where a repetitive action or exposition to a positive or negative stimulus leads to the development of attraction or aversion towards it, and thus, RL can be considered a valid strategy to model instrumental learning (Sect. 2.1.2). Thus, the first section (Sect. 2.2.1) will cover the bases of the RL theory (R. Sutton and A. Barto, 1998), from the definition of Markov decision processes (MDPs) to the description of the main RL algorithms adopted in this thesis.

A step further is then taken to explain what is currently known to associate instrumental behavior with RL. Sect. 2.2.2 will go through the major computational principles that are associated with habitual and goal-directed instrumental behaviors, described in Sect. 2.1.2, and the way their coexistence and interplay are then modeled in the literature.

Once defining this theoretical framework for spatial decision-making and learning, we will show how these principles have been applied in the recent literature to model rodent free navigation. Thus, the analysis of the existing behavioral models for spatial navigation in rodents will be reviewed in Sect. 2.2.3.

Following the short survey on the modeling of free exploration, we then examine how rodent spatial learning can be explained and modeled by using reinforcement learning and most importantly the hippocampal replay-inspired strategies (Sect. 2.2.4). Here, we report the late efforts to associate different replay-inspired RL mechanisms to the underlying biological phenomena (awake or asleep hippocampal reactivations, reverse, forward, unordered or imaginary sequences).

Ultimately, to explain all the computational instruments used in this thesis, an introduction to the bases and main working principles of evolutionary computation is provided in Sect. 2.2.5. Here more details are given on the methodologies applied in the scientific contributions of the presented research, meaning single- and multi-objective optimization strategies for model parameters estimation.

2.2.1 Reinforcement learning and value-based decision-making models

Modeling behavior can be achieved at many levels and, in this thesis, we model it as a series of decision-making steps in a discrete and sequential environment. For this purpose, we introduce finite Markov decision processes (MDPs), which are mathematical models for sequential decision-making, where an agent, its state, and actions can be described in their search for the maximization of future reward (R. S. Sutton and A. G. Barto, 2018). In the recent literature, the definition of a Finite Markov Decision Process is what has been most commonly used to model agents' behavior,

thanks to its precise formalization and enormous capability to be applied in very different problems of learning through interaction (Bellman, 1957).

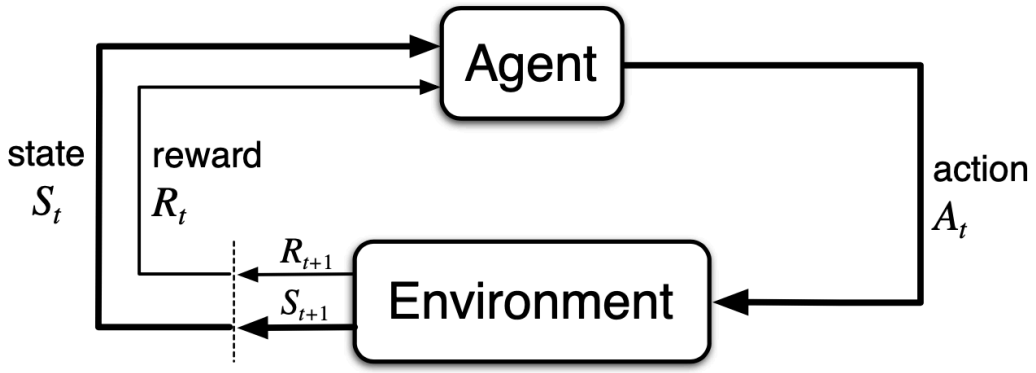


FIGURE 2.5: The agent interacting with the environment in a Markov decision process. Figure reprinted from R. S. Sutton and A. G. Barto (2018).

In the framework of discrete MDPs, well illustrated in Fig. 2.5, when we consider a single time step, the decision-maker is the *agent* and it can interact with the environment through a selected action A_t. Following this action, the agent will then access the new conditions of the environment, state S_{t+1} and the instantaneous reward R_{t+1}. Then, the decision-making process goes on to the next time steps, allowing the learning agent to keep enriching the experience of its interaction with the environment and the accumulation of a reward signal. This formalization gets its name from the *Markov property* due to the assumption that the probability of reaching possible states S_t and rewards R_s depends only on the previous state S_{t-1} and action A_{t-1} and the interaction agent-environment can be fully described by its *state-transition probabilities*:

$$p(s'|s,a) \doteq Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r|s, a) \tag{2.1}$$

where $\mathcal{R}$ is the set of possible rewards. This implies that the previous state S_{t-1} includes all the information of the past agent-environment interaction that is useful for the future choice of A_t.

Almost all problems of goal-directed learning can be expressed as an MDP where the continuous agent-environment interaction makes training and learning possible. In MDPs, learning rules can be derived just from the combination of the three signals, represented in Fig. 2.5: The choices made by the agent (*actions*) A, the context where these choices are taken (*states*) S, and the agent's goal (*rewards*) R.

The branch of Machine Learning which approaches this kind of goal-directed learning from interaction in MDPs is called *Reinforcement Learning (RL)*. Reinforcement learning takes inspiration from the way humans and animals learn how to achieve goals via their interaction with the environment. No precise indications or guidance is given to infants when they learn how to properly grip an unknown object or when a mouse learns how to enter the proper maze corridor with food, but their interactions with the world are the same as an RL agent. They are learning an optimal behavior through an Action-Outcome instrumental conditioning protocol

(Sect. 2.1.2), where they gradually discover that a certain rewarding situation is associated with a specific action, and they want to optimize their strategy to maximize their reward.

These main concepts are *trial-and-error search* and *delayed reward*. On the one hand, trial-and-error refers to the absence of complete knowledge of the agent-environment interaction or on the task that makes the learner usually experiment with many sub-optimal sets of actions before getting to the optimal one, called *optimal policy*. In RL, the term *policy* indicates the agent's behavior. The policy represents the mapping of the perceived state and chosen action, and it is usually improved via the experience (for example encountering unexpected rewards). In most cases we are interested in maximizing the sum of future rewards, expressed as the discounted return the notion of expected return, that is the sum of $G_t = \sum_{t=0}^{t=+\infty} \gamma^t R_{t+1}$. This definition of G_t is a deliberate choice and it would be possible to define other return criteria and this would lead to different algorithms and associated policies (for example Jarboui and Akakzia (2022)).

One of the best strategies to improve the policy with respect to the maximization of G_t is through the use of the *Bellman equation* which describes the relationship between the value of a state and the values of its successor states under an optimal policy. The use of this equation makes it easier to derive learning algorithms. In fact, the unique solution for the Bellman equation for a specific policy π is the value function for that policy v_π, as expressed in Eq. 2.2.

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum rp(s', r|s, a)[r + \gamma\, \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \quad \text{for all } s \in \mathcal{S} \end{aligned} \tag{2.2}$$

This equation states that the value of each state s in the set of all possible states $\mathcal{S}$ is the discounted value (with a discount factor γ) of the expected next state s', plus the expected reward r. The value function v_π for a given policy π is what is most important when a learning agent needs to take a decision. Proposing methods for efficiently estimating the value function is one of the most important contributions of RL; values functions are built on top of the environment and the reward signal, and through the experience, they can map and propagate the rewarding information and decisions in the state-action space to improve the policy. Thus, the value of a state represents its long-term desirability, that is the total amount of reward that the learner can expect to achieve in the future, from that particular state.

In the scope of this thesis, we will focus specifically on *RL tabular solution methods*, which are the simplest form of RL algorithms. These algorithms are called tabular since their state and action spaces are discrete and not too large and can be represented in tables. The main RL algorithms we are going to adopt in the scientific contributions of this thesis are *Temporal-difference* methods, *Q-learning*, and *Value iteration*.

Firstly, temporal-difference (TD) methods update the estimate of the value function of a certain policy v_π by *bootstrapping* on other estimates. Compared to other RL methods, such as the Monte Carlo ones (R. S. Sutton and A. G. Barto, 2018), TD methods can also learn from one action. They combine the sample update strategy

of Monte Carlo methods without waiting until the end of a run and use it to bootstrap the value of the current state considering only the sample successor state and not the complete distribution of all possible successors as in dynamic programming strategies (R. S. Sutton and A. G. Barto, 2018). The equation for updating the value function in case of one-step TD learning (TD(0)) is as follows, where R is the reward signal and α is the learning rate:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{2.3}$$

Secondly, q-learning is an off-policy temporal difference algorithm, that allows the agent to learn a direct approximation Q of the optimal action-value function q_*, independently of the agent's current policy (Watkins, 1989). Differently from TD(0), the value function is here called q-function and depends on both states and actions. Its action-state function is updated as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \tag{2.4}$$

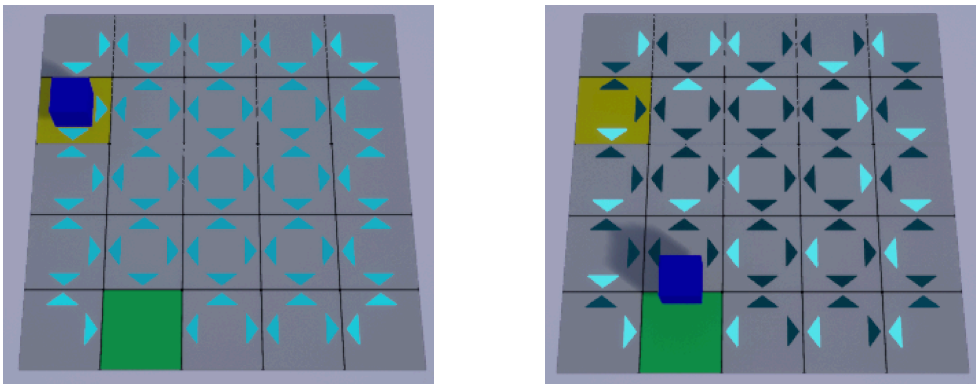
where α is the step size parameter or *learning rate* and γ is the discount factor, already presented in Eq. 2.2. Without any knowledge of the model of the environment, the Q-learning rule updates, along with the agent's experience, the state-action q-value $Q(S_t, A_t)$ based on the maximal q-value obtained from the arriving state $S(t+1)$.

As a relevant visual example of the q-learning at work, we can look at the grid-world in Fig. 2.6, where a blue cubic agent learns the q-table to efficiently navigate from the yellow to the green state. Firstly, Fig. 2.6a shows a visual representation of a q-table $Q(S_t, A_t)$ at its starting condition when all its values are zeros. Secondly, Fig. 2.6b exhibits the same scenario after a total number of 100 actions taken by the blue agent; Each state shows with a light blue arrow the preferential action to be taken by the agent to get to the reward state. Since the q-function has not converged yet to its optimal values q_*, the states which are further away from the rewarding state may suggest an action that is in the opposite direction of the green state, while the closer states express a better policy.

Eventually, value iteration instead fits in the type of RL algorithms, called *dynamic programming* (R. S. Sutton and A. G. Barto, 2018). This type of strategy requires complete knowledge of the environment, *i.e.* the transition and the reward functions, and usually have a high computational cost, due to the procedure of policy evaluation, which generally requires updating the value function for many sweeps through the state set. By having complete knowledge of the environment in an MDP, we can say that the algorithm has access to the $p(s', r|s, a)$ for all the (S, A, R) in the system. The advantage of using value iteration compared to other dynamic programming strategies lies in the fact that the computational cost is reduced by stopping the policy evaluation procedure after one sweep over all the states and still being able to converge to an optimal policy for a discounted and finite MDP. The update of the value function that is performed for each state during a value iteration sweep is:

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_k(s+1) | S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r|s, a)[r + \gamma v_k(s')] \end{aligned} \tag{2.5}$$

This equation above is obtained by converting the Bellman equation into an update rule where the termination criterion is the moment when the value function



(A) The q-table at t=0 (B) The q-table at t=100

FIGURE 2.6: Grid-world q-learning example: A blue cube agent exploring a 5x5 states world, which 4 possible actions to be taken in each state (north, south, east, west). The yellow square represents the initial state, where the agent returns after each trial, *i.e.* after having reached the rewarding green state. The colors of the 4 arrows per state indicate the q-values $Q(s,a)$ associated with that particular action a from that state s: Lighter arrows correspond to a higher $Q(s,a)$. The interface to generate this example is ReinforceMe! (https://loreucci.github.io/projects/reinforceme/).

does not change more than a small threshold ϵ after a sweep.

While the value or the action-value functions are being learned, the agent can exploit these values to make decisions for its next actions. For instance, in the computational models we described in Sect. 3.1-3.2, the simulated agent decides to go to a particular state s with a probability $P(s)$, computed as the Soft-max Boltzmann distribution applied onto the values $V(s)$ of its possible next states $\mathcal{S}$:

$$P(s) = \frac{e^{\beta V(s)}}{\sum_{i \in \mathcal{S}} e^{\beta V(i)}} \tag{2.6}$$

Here, in Eq. 2.6, β is called the *inverse temperature* and tunes the exploration/exploitation trade-off; lower βs favors exploration, while higher ones exploit more the information contained in $V(s)$. This decision-making strategy is one of the most used among the classical ones such as ϵ-greedy (R. S. Sutton and A. G. Barto, 2018). The use of the Soft-max Boltzmann distribution has been also previously adopted to model the brain mechanism of decision-making under exploration/exploitation trade-off (Daw et al., 2006; Khamassi et al., 2011).

2.2.2 Model-based and model-free reinforcement learning

Although all the RL methods presented in the previous section (Sect. 2.2.1) are based on the approximation of a value function which is based on an expectation of the future reward used to back up its values, they present also clear differences. RL techniques that can be used without a model of the environment, such as temporal difference learning, are called *model-free (MF)*. In contrast, RL strategies based on dynamic programming, such as value-iteration, require the knowledge of a model under the form of a transition and reward functions and are then called *model-based (MB)*. The model of the environment can be deterministic or stochastic and in this latter case, they are usually represented as *distribution models* where the agent knows

the transition and reward probabilities $p(s', r|s, a)$ of the current MDP. This knowledge can be used to simulate a sequence of actions and experiences, before taking the actual decision, in a process that is commonly addressed as *planning*.

Since (at least) two distinct behavioral strategies have been proved to exist in the brain for instrumental behavior (habitual vs goal-directed behavior, Sect. 2.1.2), different research groups have proposed models where these strategies respectively correspond to model-free and model-based RL (first among the others Daw, Niv, and Dayan (2005)), and have proposed mechanisms to describe how they could efficiently work together. Model-free RL strategies are indeed a proper model for habitual behaviors because they slowly converge to an optimal series of actions, by using a very low computational cost and also slowly reacting to external changes in their learning scenario. This scarce planning power and low behavioral reactivity are what mainly characterized habitual behavior in rodents and humans. On the other side, model-based RL algorithms are more efficient in adapting to context changes, but they require a higher computational cost to infer and plan for granting this fast reaction. This higher computational cost reflects the largest involvement of high cognitive brain areas required in goal-directed behavior (Sect. 2.1.2).

One of the first studies proposing a model for multiple learning strategies was Guazzelli et al. (1998) which applied it in a navigation task. They proposed a cooperation between two TD-learning algorithms: *Taxon-Affordances* which learned directly on the perceived stimuli, closer to a cue-guided behavior, and *World-Graph* which built an association between the perceived stimuli and the external environment, similar to a more structured map-based behavior. Then, decision-making is performed through a combination method based on the values associated with the next possible actions by each algorithm, which are then summed by a meta-controller.

An interesting theory for the combination of multiple learning systems in sequential procedure has been then proposed by Hikosaka et al. (1999). Their proposal suggested that the acquired information for learning a determined sequential behavior is distributed and elaborated in many parts of the brain, such as the prefrontal cortex (working memory), the hippocampus (declarative memory), and the basal ganglia (procedural memory). In the motor control case, which they analyzed, they suggested that two brain areas learn their sequential procedures at the same time, but for two different systems: the spatial coordinates and the motor coordinates. Finally, they proposed that it exists a neural loop circuitry between the basal ganglia and the cerebellum where the procedural learning control passes from the former to the latter once the inverse model between the goal spatial-space and the motor-space has been learned through the task.

Following these first proposals, another mechanism called *responsibility signal* has been developed to weigh the relative goodness of predictions between different MB-RL modules (Doya et al., 2002). Since each module is composed of a state prediction model and an RL controller, the responsibility signal is implemented as a gaussian softmax function of the errors in the outputs of the prediction models.

To organize the contributions of many behavioral navigation strategies, an interesting development arrived when, next to the more computationally expensive combination strategies (for example Guazzelli et al. (1998)), the first arbitration methods started to be considered. An important contribution in this direction came from Daw, Niv, and Dayan (2005), who, always using a computational framework, based on RL, proposed a Bayesian principle to choose which of the two modes (model-based or model-free) should take control over behavior based on its accuracy. Between a value-iteration (MB) system and a q-learning (MF) one, the chosen expert

would be the one with the lower variance of a posteriori distribution of expected value $\boldsymbol{Q}_{s,a(q)\equiv P(Q(s,a))=q)}$ [1].

The coexistence of two instrumental behavioral systems is thought to give a relevant contribution at different stages of learning (Hikosaka et al., 1999; Daw, Niv, and Dayan, 2005). On the one hand, MB systems are better in adaptation towards dynamical scenarios, but the computation to account for the elaboration of all the information to be that flexible is very expensive (Dromnelle et al., 2022). On the other hand, MF systems have a faster response, but they do not integrate enough knowledge of the environment to be adaptive. Keramati, Dezfouli, and Piray (2011), also using an RL computational framework, proposed that the arbitration mechanism between these two behavioral modes is based not only on the decision accuracy (as proposed by Daw, Niv, and Dayan (2005)) but also on the cost of deliberation. Looking at the *value of perfect information* $VPI(s,a)$ which is a measure of the decision accuracy from state s for each action a, for each expert, the arbitration mechanism would pick the expert with the highest value that it is not having a too high cost of deliberation. This means that $\bar{R}_\tau$, the amount of reward that could potentially be acquired in a deliberation time τ, should be lower than the actual reward that the deliberation time τ allows to get.

After further experimental evidence, also in human studies, supporting the existence of two unique learning modalities and the identification of the neural signatures for the reward prediction error (RPE) (D'Ardenne et al., 2008; Haruno and Kawato, 2006; O'Doherty et al., 2003) and the state prediction error (SPE) in humans (Gläscher et al., 2010), this instrumental behavioral duality has been extended as a leading mechanism in other functional scenarios. For instance, this model-based/model-free distinction turns out to be also relevant to describe mammal behavior during Pavlovian conditioning paradigms (Lesaint et al., 2014).

In light of the goal-directed and habitual behavioral strategies described above, many recent proposals for computational navigation models are based on the coordination of parallel learning systems. For instance, Dollé et al. (2010) proposed a new model for navigation based on the coordination of three different strategies by a meta-controller, called *Gating network*, based on TD-learning. Depending on the accuracy of the three strategies in predicting the future reward, the meta-controller can learn when it is better to follow a taxon expert (procedural, associative, and egocentric learning), a planning one (cognitive, planning, and allocentric learning), or pure random exploration. Their proposed novelty was in particular in the design of a meta-controller that could integrate the information (*e.g., common currency*, meaning the proposed goal-directed action) coming from different learning strategies and learn how to optimally choose among them. The proposed model has been later integrated with a more realistic hippocampal model (Ujfalussy et al., 2008) on the cognitive and planning part of the algorithm (Dollé et al., 2018). In this work, they were able to replicate some experimental results, among which Morris et al. (1982) and Devan and White (1999), showing that the gating network coordination strategy is indispensable for these replications and can coordinate the contributions of different learning strategies.

An innovative perspective came then from this hypothesis article, Khamassi and Humphries (2012), where they proposed that in spatial instrumental behavior, the main distinction between model-based (MB) and model-free (MF) learning depends on *how* the knowledge is used more that on *what* kind knowledge is used. As shown

[1] The MB variance of $\boldsymbol{Q}_{s,a(q)\equiv P(Q(s,a))=q)}$ has a fixed minimum value without strong theoretical justifications.

in Fig. 2.7, they propose a detailed functional model of the basal ganglia as a neural substrate for spatial learning.

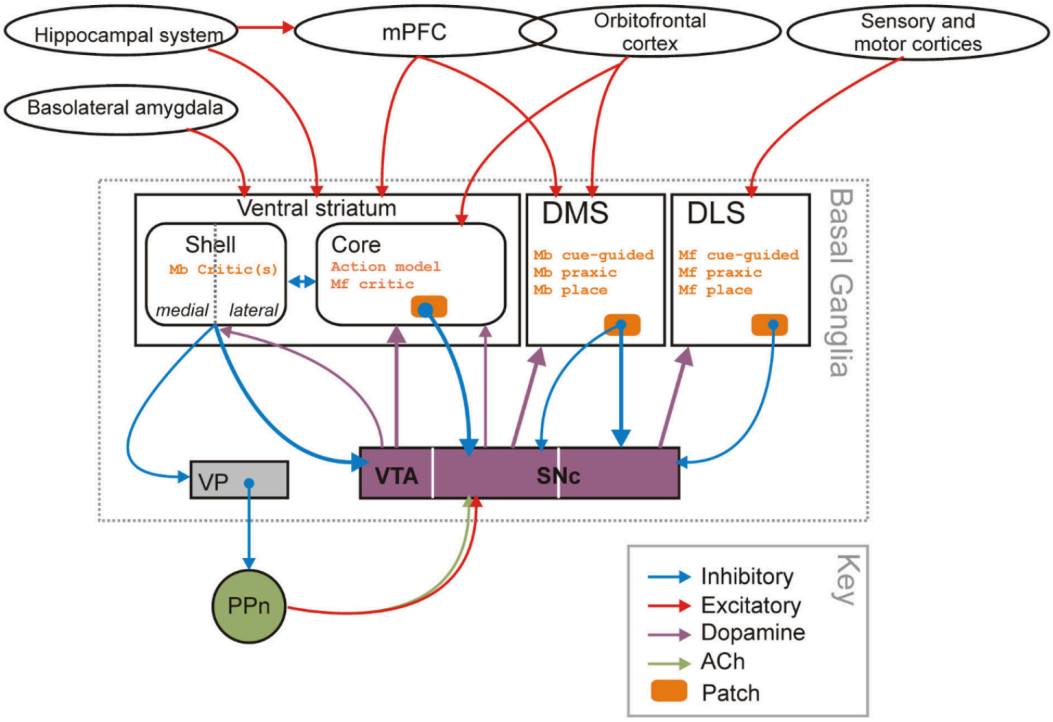


FIGURE 2.7: The proposed functional model for the integration of model-based (MB) and model-free (MF) navigation strategies in the brain. The basal ganglia are identified as the main center for spatial instrumental behavior thanks to their communication with the amygdala, hippocampus, medial prefrontal cortex (mPFC), orbitofrontal cortex, sensory and motor cortices, and pedunculopontine nucleus. In particular, the areas specifically related to the processing of spatial information are the hippocampus (Sect. 2.1.4), the sensory and motor cortices, and the mPFC, considered as a center for the elaboration of place representation (Hok et al., 2005) and deliberation on the output of the two strategies (Wunderlich, Dayan, and Dolan, 2012). Figure reprinted from Khamassi and Humphries (2012).

The dorsomedial striatum (DMS) works as an MB system and the ventral striatum is the corresponding part dedicated to building the model of the world. In fact, the striatum has been proven to be extremely relevant for value-based planning (Wunderlich, Dayan, and Dolan, 2012), and in this work, they proposed that its core, in particular, is dedicated to learning the transitions probability model of the environment. Finally, they argue that different parts of the ventral striatum could compute the RPE for the model-free system, in its core, and for the model-based one, in its shell, suggesting that the ventral striatum plays the role of the *critic* in a classical actor-critic RL algorithm (Konda and Tsitsiklis, 1999). On the other side, the DMS and the dorsolateral striatum (DLS) represent the *actors*, of the MB and MF systems, respectively.

More recently, other groups have proposed new theories and computational models on the coexistence of these two known strategies for instrumental behavior. Pezzulo, Rigoli, and Chersi (2013) proposed a *Mixed Instrumental Controller* that, through a costs and benefits analysis, decides whether to privilege a habitual behavior or to

re-evaluate the cached values for a specific state-action transition. Compared to the models proposed before (Daw, Niv, and Dayan, 2005; Keramati, Dezfouli, and Piray, 2011), this work introduces a more central control where the cost of model-based predictions is engaged just when needed, *i.e.*, in particular when the *Value of Information* surpasses the cost of the inference process. They defined the Value of Information (Voi) concerning the possible action $Act1$, over the other option $Act2$, as follows:

$$VoI_{Act1} = \frac{C_{Act1}}{|Q_{Act1} - Q_{Act2}| + \epsilon} \tag{2.7}$$

where C_{Act1} is the uncertainty linked to $Act1$, and Q_{Act1} and Q_{Act2} the values of the actions $Act1$ and $Act2$, respectively. ϵ will assure that the denominator is non-zero. The MB re-evaluation of the states-actions values Q is similar to a *trajectory sampling* episode (A. G. Barto, Bradtke, and Singh, 1995) where the length of the sampled trajectory is controlled by the VoI of the sequence of actions. A similar inference mechanism, with a controlled re-evaluation budget, called *Simulation Reactivations (SimR)*, is also adopted in our scientific contribution about the coordination between MB and MF replay strategies in goal-directed navigation (Sect. 4.1.4).

Inspired by the forward re-evaluation sweeps of the inference strategy proposed by Pezzulo, Rigoli, and Chersi (2013), Keramati et al. (2016) proposed a new computational strategy for the coordination of an MB and MF learning systems. Similarly to Doya et al. (2002), they suggested that the presence of these two learning systems could be seen not just as a dichotomy, where one of the two systems takes control over the other because more reliable at less expensive at that point, but as more like two strategies that always act together, modulating the weight of their contribution using their competence. In this way, many more learning strategies could exist in the animal and human brains based on the modulation of the contribution of the two experts (MB and MF). By testing their ideas on a theoretical framework based on RL, they propose a *plan-until-habit* solution, where the q-value $Q(s,a)$ of a given state and action couple (s,a) is computed by limited forward simulations to predict the future discounted rewards r until a predefined depth k when the habitual q-value for the consequences of the remaining future steps $Q^{habit}(s',a')$ is summed to the prediction.

$$Q(s,a) = r_0 + \gamma r_1 + \gamma^2 r_2 + .. + \gamma^k Q^{habit}(s',a') \tag{2.8}$$

In Sect. 4.1.4 of this thesis, we are going to explore deeper the role of the arbitration mechanism proposed by (Dromnelle, Renaudo, et al., 2020; Dromnelle, Girard, et al., 2020) in goal-oriented spatial navigation tasks and the possible advantages of adding MB and MF replay strategies, respectively integrated within the two systems.

In conclusion, the current consensus is that the animal and the human brain exercise instrumental conditioning and domain-general learning by using a framework that can be modeled as a mixture of experts in machine learning (O'Doherty et al., 2021) (usually MB and MF experts), and their prevalence can be regulated trial-by-trial (Otto et al., 2013). The arbitration among the experts is performed based on the uncertainty of their predictions and their cognitive cost and, thanks to the variety of their proposals, effective interaction, and adaption to different environments and situations are possible.

2.2.3 Models of spatial navigation in rodents

Despite the apparently arbitrary nature of free spatial exploration in rodents, Sect. 2.1.1 illustrates several regular patterns that portray a sort of systematic unconstrained navigation in novel environments. The same research group that conducted most of these behavioral studies, proposed also a first computational model for rodent free exploration behavior (Gordon, Fonio, and Ahissar, 2014a).

Firstly, they proposed a hierarchical curiosity model to mimic an autonomous active agent that optimizes the learning of its sensory-motor integration by recompensing curiosity, in the form of RL reward prediction error (Gordon and Ahissar, 2012). In their work, additional levels of sensory-motor mapping are built on top of each other as RL incremental natural actor-critic agents (Fig. 2.8a, Bhatnagar et al. (2007).

Assigning the square of the learning error e_2 to the reward function R would create a reinforcement active learning agent (ReAL) (Gordon and Ahissar, 2011) with a strong attitude towards exploration. The reward for this agent will then be:

$$R_{t+1} = e_t^2 = (\hat{o}_L(i_t) - o_t)^2 \tag{2.9}$$

where o is the current output for the input i and $\hat{o}_L$ is the estimated one that follows the input-output transformation L dedicated to learn the model of the interaction agent-environment. The parameters defining the transformation L are autonomously updated according to the reward prediction error e_2 by internal supervised learning. This implementation allows the gradual learning of knowledge and skills of increasing complexity concerning the interaction with the surrounding environment. They applied this strategy in particular to rodent whiskers system to learn the internal model of their motion and their object localization activity.

Secondly, the same principle was improved to model rodent whiskers' exploratory patterns in novel circumstances (Gordon, Fonio, and Ahissar, 2014b). In this case, the exploring agent uses the information gain as an intrinsic reward, instead of the reward prediction error as done in Gordon and Ahissar (2012). To measure the information gain, they used the Kullback-Leibler divergence of the new observed state:

$$D_{KL}(P_{t+1}(s'|o)||(P_t(s')) = \sum_{s'} P_{t+1}(s'|o) log\left(\frac{P_{t+1(s'|o)}}{P_t(s')}\right), \tag{2.10}$$

where $P(s')$ is the previous known probability to encounter the sensory state s' at time t, and $P(s'|o)$ is the new actual probability to encounter it at time $t+1$ given the observation o. This divergence measure indicates the amount of new information which is present in the observation o and this information gain is used then to guide the agent perception toward states s' with higher gains.

Here, they also introduced the concept of *novelty control* which allows an alternation between *exploratory* and *retreating* behavior by keeping constant the amount of novelty perceived by the agent in its whisker system.

Eventually, their research adapts this latter model to describe either the whisker system and the locomotion one to create a complete spatial exploration framework (Fig. 2.8b, Gordon, Fonio, and Ahissar (2014a)). As in their previous work, the perceiver is a Bayesian learner that aims to learn the sensory-motor forward model (Kawato, 1999), the critic tries to learn instead the expected future reward and the actor chooses the current action, given a stochastic policy (Fig. 2.8a). In the different exploration primitives, by having each actor start from a different random

distribution for exploring the agent-environment relationship, the main exploration primitives will be learned from the most to the least novel ones.

As already mentioned, the novelty of this work compared to Gordon, Fonio, and Ahissar (2014b) is also the introduction of four locomotion loops (Fig. 2.8b). These exploration primitives are fed by the two whisker loops which provide information about the presence of "contact points" to external objects in the environment. This allows the locomotion loops to learn, based on the morphology of the agent's surroundings, features of the environment, like the presence of *corners*, *walls*, and *open spaces*. Also, in this case, the novelty controller operates to make the agent retreat when the novelty level is too high or explore when it is too low (Fig. 2.8c). The probability for the agent to pass to a retreat mode from a particular exploration loop l is:

$$p^l_{retreat}(r_t) = \psi\Big((r_t - (\hat{j}^l_t + \tilde{r}^l))/\tilde{r}^l\Big), \tag{2.11}$$

where r_t is the current reward, $\hat{j}^l_t$ is the current loop l average reward, $\tilde{r}$ is the novelty-transition sensitivity and $\psi(x) = 1/(1+e^{-x})$ is a sigmoid function. So that if the reward r_t exceeds the current estimation of the average reward $\hat{j}^l_t$, there is a higher probability for a switch to retreat mode. On the contrary, if there is no more novelty in the current exploration loop l, the novelty controller passes to a higher exploration level loop with a probability:

$$p^l_{adv}(\tau^l) = \psi\Big(\tau^l - \hat{\tau}^l)/\tilde{\tau}^l)\Big), \tag{2.12}$$

where τ^l is the accumulated time of loop l, $\hat{\tau}^l$ is the advancement threshold and $\tilde{\tau}^l$ is the advancement sensitivity. Before any advancement in exploration loops, the agent returns to base to assure that all learned motor primitives have a common starting point.

To conclude, the innovative idea behind the design of this model is the identification of novelty as the primary driving force for free exploration and the fact that this novelty is gradually absorbed by the agent. This idea takes, however, inspiration from the assumption that curiosity and novelty propel and optimize exploration and can produce agents which can systematically and autonomously learn tasks of increasing complexity; a view that has been recently explored a lot in artificial intelligence and robotics (Oudeyer, Kaplan, and Hafner, 2007; Pape et al., 2012; Little and Sommer, 2013).

Another interesting approach for a bio-inspired adaptive agent for spatial navigation tasks came from (Cos et al., 2013). They propose an RL-motivated actor-critic agent whose exploration is guided by a *hedonic value (HV)* and constrained by physiological stability. The proposed HV is inspired by the role of phasic dopamine as an error signal to module the subjective behavioral assessment response to different stimuli and physiological states (Houk, Davis, and Beiser, 1995; Khamassi et al., 2005). By simultaneously considering all the agent's homeostatic variables (abstractions representing the dynamics of its internal resources), their proposal produces an autonomous agent which owns an internal modulation of the value relative to the environment and that can rapidly adapt to non-stationary environments.

In our scientific contribution (Sect. 3.1), we will propose a free exploration model that considers the information gain coming from the environment from a different point of view compared to the research presented so far. We will describe the decision-making process of the animal as a combination of many factors which model the importance of its different emotional components, such as anxiety, fatigue, and

activity as a reflection of its internal curiosity. Even though research on rodents has strong experimental protocols and a long history, research laboratories are not the real natural environments of such animals and that makes the job of understanding and modeling their spontaneous behavior very challenging. This situation could alter the anxiety levels, and thus the subsequent spatial exploration, as the preferences for certain corners of the maze, which are considered safer and more familiar. These inclinations could also vary in relationship to social interactions, individual status, and gender of the animal (Balcombe, 2006).

Besides, regarding the modeling strategy in rodent free exploration, a recent research also points out that by considering an exploratory session of a novel environment as a whole, some regularities emerge more than searching for consistent local patterns (Benjamini et al., 2011). An example of these regularities can be seen as the cumulative distance traveled and the percent time spent in the center of an open-field maze respectively as measurements for the levels of activity and anxiety for a specific animal.

In this thesis, we investigate whether it is possible to identify some regular patterns in the decision-making process of rodents exploring a novel environment and if it is then possible to encode the relevance of these behavioral features for each animal (Sect. 3.1). The model we propose considers both allocentric and egocentric spatial representations of the animal in the environment (Berthoz, 1991; Klatzky, 1998) and would constitute a framework for building decision-making, spatial memory, and learning.

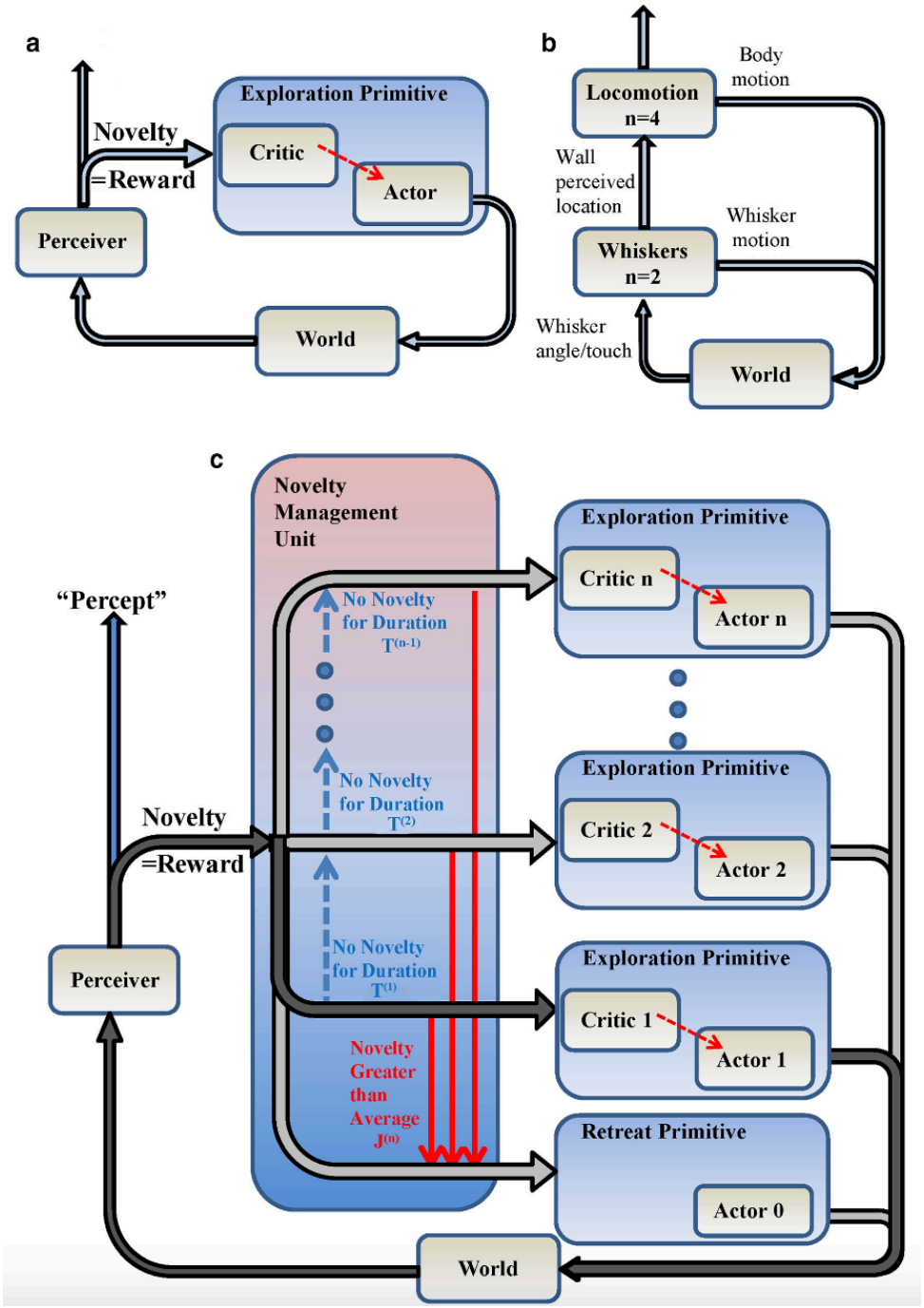


FIGURE 2.8: Scheme of the model for exploration and whiskers motion from Gordon, Fonio, and Ahissar (2014a). A) A single exploration primitive: The interaction agent-environment is dealt with an RL actor-critic agent, whose actions on the world depend on the amount of novelty (=reward) perceived by the critic. B) The system is composed of two exploration modules for the whisker control and four modules for the locomotion one. The locomotion system is of a higher rank compared to the whiskers' one and uses the sensory information provided by the latter to perform better localization. C) The novelty management unit as a mechanism to alternate between exploration and retreat; when the novelty for a specific exploratory module is higher than its average, the agent chooses the retreat primitive, otherwise the next exploratory module is activated.
Figure reprinted from Gordon, Fonio, and Ahissar (2014a).

2.2.4 Reinforcement learning-based replay mechanisms

Even before the numerous studies that have been conducted in neuroscience about hippocampal replay, researchers in RL were starting to design and test artificial memory mechanisms that improved the performance of their classical algorithms.

R. S. Sutton (1990) introduced the *Dyna-Q* algorithm where MF trial-and-error learning is integrated with MB online planning on a learned model of the world. Compared to the previously proposed *Dyna-PI* and considering the Q-function instead of a Value function for a specific policy π, Dyna-Q assures that q-values, and also the learning process will converge properly whatever policy is used. This strategy is also more robust to face dynamic experimental environments thanks to the fact that it learns a stochastic model of the world, differently from Dyna-PI, which learns a deterministic one. In this way, the system has a sort of intrinsic exploratory behavior, which can make it go for suboptimal actions at the beginning, but that better deals with stochastic and changing environments. Further, during the planning phase, the agent is allowed to experience actions that have not been tried before and so integrate this knowledge in the learning of the Q-function.

In the same years, (Lin, 1992) introduced a mechanism called *experience replay* that consisted in effectively reusing past experience. A past event was described in the form of quadruples (s, a, s', r) meaning that the agent had started from state s and having performed action a had ended up in state s' with a reward r. Sequences of events starting in the initial state and leading to the goal state, called *lessons*, are then re-experienced in the opposite direction (backward replay). This means that each event in a lesson is used to update the state-action values as if it were experienced again. In particular, they suggested that experience replay could be more efficient if the propagation of the past knowledge was a sequence of the above-explained transitions and if it was replayed in backward order. By adopting this strategy, classical MF-RL algorithms, such as Q-learning (Watkins, 1989), were demonstrated to converge faster.

Although the computational RL strategies explained above have been designed and conceived separately from hippocampal reactivations, their similarities and interplay have been discussed more and more recently.

As developed in Dyna-Q, Pezzulo, Kemere, and Van Der Meer (2017) proposes that both the experience-tied (MF) and the internally generative (MB) experience can be involved in the inference processes of rodents, and mammals in general. The way this experience is recalled in reasoning and imagination happens in the form of replay or *internally generated sequences (IGSs)*, mainly in the hippocampus, but also in many other areas (Sect. 2.1.4). They first suggest that *active inference* which appears during task-engagement and theta cycles, replays successively experienced elements more rapidly if the location is next to the current animal position. Also, they claim the importance of the different dynamics and timescales of respectively theta cycles and sharp wave ripples (SWR) (Sect. 2.1.4) since these differences could correspond to diversity in many aspects, such as communication to other areas of the brain, in the timing of occurrence (task-engagement, sleep, etc.) and functionality. The neural mechanism for task-engaged reasoning and mental time could be the same, with the difference that the clue coming from the actual action-perception cycle guides the mental experience triggered on the generative model in a strongly constrained way compared to detached imagination.

The most known and used heuristics that have been proposed to make the best use of simulation replay are *prioritized sweeping* and *trajectory sampling* (R. Sutton and A. Barto, 1998). Prioritized sweeping was first proposed as an efficient strategy to

enhance learning and increase computational efficiency and it was independently proposed and developed in the same year in two slightly different versions. Both methods compute the reward prediction error (RPE) of each state-action pair and they consider it as its priority value and prioritized replay is then performed from the most unexpected pair to the least unexpected one in the queue. The replay phase stops when the priority queue (P-queue) is empty, meaning that no relevant knowledge is left to be replayed. On the one hand, the algorithm proposed by Peng and Williams (1993) suggests prioritizing the order of the value function estimate updates in Dyna-Q (R. S. Sutton, 1990). They proposed to prioritize the largest recent updates and then their predecessors' states' priority is computed and they are evaluated next. Next, the algorithm prioritizes the updates, starting from the current state, whose estimate is relevant for future long-term reward. On the other hand, Moore and Atkeson (1993) proposed a similar prioritization technique which put forth the backups of the most recent surprising state and their predecessors, without re-compute the predecessors' priority when they are then considered in the priority queue. While prioritized sweeping prioritizes backups of predecessors of states that have recently changed their values, trajectory sampling instead consists of privileging backups on the on-policy distribution, thus on the successors of the current state (R. Sutton and A. Barto, 1998). Here, the replay phase starts by simulating the on-policy trajectory from the current state and then performing backups at each state-action pair. This backups strategy results in a computationally efficient way of updating state-action values by ignoring uninteresting parts of the task space, but it can hurt in the long-term since backing ups just the on-policy state-action couples would become rapidly irrelevant.

With the advancement of the research in this direction, new proposals for more comprehensive algorithms, aiming at generating and explaining more replay mechanisms and functionalities are arising. One of the first computational proposals that allow the spontaneous generation of both reverse and forward replay came by Aubin, Khamassi, and Girard (2018). Designing and testing their algorithm in a navigation task, they first propose a neural network (NN) version of the Dyna-Q algorithm (R. S. Sutton, 1990), which also uses prioritized sweeping (Peng and Williams, 1993; Moore and Atkeson, 1993). Then, their strategy includes also a new NN architecture, *GALMO*, dedicated to learning the world model. The GALMO NN is essential to the proposed navigation task (double T-maze, Gupta et al. (2010)) because multiple predecessors exist for certain states of the maze. Multiple predecessors require that the world model is learned offline by presenting the data in random order to disrupt their sequential correlation. They suggest that the above-described mechanism can be predictive of the fact that also rodents learn the world model offline by non-sequential hippocampal reactivations. Then, another interesting result found in this study comes indeed from the spontaneous generation of replay in the proposed Dyna-Q prioritized sweeping set-up; the majority of the generated reactivations are non-sequential, while the 15-20% of them are either backward or forward, predicting a strong relevance of unordered replay even if the priority sweeping strategy is usually encouraging sequential reactivations, driven by larger prediction errors.

Following this work and based on the ideas behind the previously proposed replay heuristics, prioritized sweeping (Peng and Williams, 1993; Moore and Atkeson, 1993) and trajectory sampling (R. Sutton and A. Barto, 1998), Mattar and Daw (2018) proposes a balancing mechanism between *need* and *gain* which can orchestrate forward and reverse reactivations. They formulate a *utility* measurement, based on need and gain, to prioritize the memory access during deliberation. The demand for replay a certain episode e_k and so to update its $Q(s_k, a_k)$ in an MDP is evaluated

based on this utility, also called *expected value of backup* $EVP(s_k, a_k)$:

$$EVP(s_k, a_k) = \mathbb{E}_{\pi_{new}}[\sum_{i=0}^{\infty} \gamma_i R_{t+1} | S_t = s_k] - \mathbb{E}_{\pi_{old}}[\sum_{i=0}^{\infty} \gamma_i R_{t+1} | S_t = s_k] \tag{2.13}$$

$EVP(s_k, a_k) = Gain(s_k, a_k) \times Need(s_k)$ since $Gain(s_k, a_k)$ is the expected improvement in return of state s_k and $Need(s_k)$ is the discounted number of times the agent is expected to visit the target state s_k. On the one hand, starting replay activity based on gain will generate backward sequences from reward locations, since gain is proportional to the surprise of an unexpected positive gain. On the other hand, if need drives the replay activity, a depth-first replay sequence generates from the current state. This happens in particular because the need contribution privileges the replay sequence to start from the most probable states where the agent will be next. However, the main limitation of this approach is the needed computation to assess the utility of each backup replay which includes the computation of the Bellman backup itself (Eq. 2.13).

Another computational model, capable of generating diverse replay mechanisms, such as forward and reverse replays, has been presented by Khamassi and Girard (2020) to suggest a computational principle to generate awake hippocampal reactivations. The strategy they proposed is constituted by a model-based bidirectional search, which is composed of a trajectory sampling phase (A. G. Barto, Bradtke, and Singh, 1995) and a prioritized sweeping one (Peng and Williams, 1993; Moore and Atkeson, 1993). This process is repeated until the two trajectories connect and the state-action values converge. Imaginary replay (Gupta et al., 2010) are also spontaneously generated in this model; The double t-maze task, simulated in their paper to test the bidirectional search algorithm, does not allow the agent to go back straight away after the decision points, while the simulated experience, either forward planned or prioritized backward replayed, can interestingly involve these forbidden paths. Interestingly, with this model, the agent is free to perform replay in any state of the environment and at any time, while Mattar and Daw (2018) forced it to perform replay only before the beginning of the trial and at the end, after getting a reward. In fact, the performance of forward replay before starting to move and backward replay after getting rewarded was thus an emergent property of the Khamassi and Girard (2020)'s model. Moreover, compared to Mattar and Daw (2018), the model minimizes computational costs by performing extensive replays only after surprising events (*i.e.*, first reward delivery and task changes).

In these same years, an important scientific contribution came by Cazé et al. (2018) who wrote a review study to associate several types of hippocampal reactivation mechanisms to RL algorithms (model-based, model-free or hybrid models such as *Dyna*, already mentioned above). A year after, Whelan, Vasilaki, and Prescott (2019) also reviewed the neuro-inspired models and artificial plasticity strategies which could explain different types of hippocampal replay. In Cazé et al. (2018), many replay-inspired RL algorithms are tested in a simulated spatial task, to accumulate as much reward as possible (where the reward is located in a specific position of the environment and then this position is changed in the middle of the experiment).

Tab. 2.1 shows also that depending on the strategy, past or hypothetical future experiences can be replayed in different orders, namely backward, forward, randomized the past order of the experience, or even imagining possible new experience configurations.

Algorithm	Step	Flavor	Awake (SWR)				Awake (VTE)	Asleep			
			Fwd	Bwd	Img	Uno	Fwd	Fwd	Bwd	Img	Uno
MF-RL	Value function learning	Vanilla (i.e., without replay)									
		Unordered experience replay				*					*
		Backward experience replay		*					*		
		Forward experience replay	*‡				*‡	*‡			
		Prioritized experience replay				*‡					*‡
		NN-based value function				*					*
	Inference	Vanilla									
MB-RL	World model learning	Vanilla									
		NN-based world model				*					*
	Inference	Vanilla (i.e., unordered)				*					*
		Prioritized sweeping		*		*				†	†
		Trajectory sampling	*	*	*	*	*	†	†	†	†
		Bidirectional search	*	*		*	*	†	†		†
Dyna-RL	Value function learning	Vanilla (i.e., unordered)				*					
		Prioritized sweeping		*		*			*		*
		Trajectory sampling	*	*	*	*	*	*	*	*	*
		Bidirectional search	*	*		*	*	*	*		*
		NN-based value function				*					
	World model learning	Vanilla									
		NN-based world model				*					*
	Inference	Vanilla									

TABLE 2.1: Possible RL algorithms explanation for some hippocampal reactivations, namely awake (sharp-wave ripples, SWR, and vicarious trial and error, VTE) and asleep ones. The observed reactivations can be in forward (Fwd), backward (Bwd), imaginary (Img) order, or unordered (Uno). NN, neural networks. * means that the considered algorithms can explain the observed type of replay. †refers to the fact that if the awake inference budget is limited, asleep reactivations of the same nature as the awake ones are also expected. ‡ corresponds to associations that have not been proposed in the literature before but tested in principle in Cazé et al. (2018). Table reprinted from Cazé et al. (2018).

They showed that different types of RL reactivations could describe either awake or asleep replay (Tab. 2.1), and some of them, like the MF Neural Network-based value function (Mnih et al., 2015), the MB bidirectional search (Khamassi and Girard, 2020) and many sub-types of Dyna-RL (R. S. Sutton, 1990), even both awake and asleep ones. Also in this case, some algorithms can replicate different categories of reactivations, like prioritized sweeping (Peng and Williams, 1993; Moore and Atkeson, 1993) and bidirectional search (Khamassi and Girard, 2020). Throughout this review, they showed that different RL replay algorithms can account for the same type of replay phenomenon (awake or asleep, backward, forward and so on), but the identification of this computational mechanism to a specific type of replay (different states and different oscillations) is of crucial importance to infer the possible content of hippocampal reactivations in the brain of navigating rodents. In fact, if the replay event is more likely to be modeled with an MF-RL algorithm, its content will more likely be referred to the past, while if the algorithm is an MB-RL one, its content will be referred to the future, due to the inference and planning nature of these latter algorithms. Generally, awake replay is more often connected to MB-RL algorithms which could work as trajectories sampler or as other planning mechanisms that can be modeled by bidirectional sampling, for example. MB-RL algorithms are also found more appropriate to model imaginary replay, since they would base their inferences on the model of the world that they have built and explore more, compared to forward MF-RL algorithms that would instead reactivate episodes which are closer to recent experience and for that not very exploratory. They further hypothesize that unordered replays can instead be more appropriate to describe the noisier dynamics of asleep reactivations and crucial to learning an internal model of the world when adopting neural networks-based strategies. Also, the Dyna-RL model-free replay on the inferred model of the world could be a good candidate for hippocampal reactivations during sleep in rodents, as well as MF-RL forward replay that updates action value for future use. Finally, their simulation results show that generally forward sequences are spontaneously replayed at decision points, while backward replay happens mostly around the reward spots, where the reward prediction errors (*i.e.*, surprise signals) are higher. Even though this has not been experimentally observed yet, the authors proposed that the vicarious trial and error (VTE) reactivations at decision points could not exclusively be forward, but also backward. In fact, many MB algorithms, listed in Tab. 2.1, can reproduce both backward and forward replays during inference; For instance prioritized sweeping (Peng and Williams, 1993; Moore and Atkeson, 1993), trajectory sampling (A. G. Barto, Bradtke, and Singh, 1995) and bidirectional search (Khamassi and Girard, 2020).

Very recently, Diekmann and Cheng (2022) proposed an RL-based replay method able to generate different types of replay mechanisms and observed statistics on hippocampal reactivations. They proposed a new strategy to prioritize memory access called *Spatial structure and Frequency-weighted Memory Access (SFMA)* where, given that an experience is defined as $e_t = (s_t, a_t, r_t, s_{t+1})$, each time one e_t is replayed, all the other stored experiences e get a priority value:

$$R(e|e_t) = C(e)D(e|e_t)[1 - I(e)] \tag{2.14}$$

where $C(e)$ represents the frequency of the experience and its reward-related value, $D(e|e_t)$ describes the spatial distance between the two experiences e and e_t accounting for structural environmental obstacle. Finally, the last component $1 - I(e)$ prevents that an experience which has just been recalled is replayed again and this

has the consequence of generating replay sequences. The SFMA can operate in two different modes: the default and the reverse mode, depending on which states are considered in the computation of the similarity measure. In the default mode, the similarity measure is computed between both current states of the two experiences, $D(s(t)_e|s(t)_{e_t})$, while in the reverse mode, the similarity is computed between the current state of the currently reactivated experience and the next state of the other one, $D(s(t+1)_e|s(t)_{e_t})$. Despite both modalities seem to reproduce diverse experimental observations on replay mechanisms, the default mode results very efficient in the generation of preplay (Ólafsdóttir et al., 2015) and shortcut replay (Gupta et al., 2010; Ólafsdóttir et al., 2015), while the reverse mode is better performing in spatial learning (Morris et al., 1982).

Being able to associate different phenomenological and functional aspects of neural phenomena, such as hippocampal replay, to computational algorithms which encode a precise scope, such as planning, retrieval of past experience, or memory prioritization, can be key in understanding which other parts of the brain could be involved and co-active during these episodes. Investigations in this direction, and towards more comprehensive computational models that solve tasks such as goal-directed navigation, affect the design of new experimental protocols and improve the comprehension of hippocampal reactivations. These works have recently pointed out that hippocampal replay could account for many computational aspects (e.g., context estimation, working memory, and spatial planning), depending on their timing, organization, and content (Pezzulo, Kemere, and Van Der Meer, 2017).

In Sect. 4.1 of this thesis, we will explore the advantages and disadvantages of some of MB- and MF-RL replay strategies explained in this current section, in dynamical goal-directed navigation tasks. To the best of our knowledge, the application of replay-inspired RL strategies in real robots is still in a preliminary phase. Our contribution will consist in identifying which strategies are the best to face the uncertainty caused by real robotic experiments. The introduction and combination of RL-based replay in goal-directed algorithms for navigation are crucial to saving real experimental time on the robot, but many challenges, regarding for example the computational costs of such strategies, have not been addressed yet.

2.2.5 Parameter estimation and evolutionary algorithms

A part of the research contribution of this thesis has been conducted by applying state-of-the-art evolutionary algorithms for optimization and model fitting. This section is going to introduce and explain the general bases of these methodologies.

To study and understand a particular phenomenon, data is usually collected. After this phase, one of the main interests is identifying and fitting this observed behavior into a model. This procedure is usually called *model fitting*. Following the model's design, the next step is often formulating a function to describe the best-fit criterion between the proposed formalism and the data. Once this function has been defined, the *parameter estimation* process starts with identifying the best parameters that bring the model to optimally fit the available data and hopefully generalize the physical phenomenon beyond that accessible sample.

In decision-making computational neuroscience, one of the most common modeled experimental protocols is the binary choice (left versus right choice) of participants (rats, monkeys, or humans) in tasks where they repeat around 100 or 200 repetitions. In these cases, the parameters of a proposed model are usually fitted by the Bayesian estimation of the maximum likelihood to observe a participant's sequence of choices (Daw et al., 2011a). Daunizeau, Adam, and Rigoux (2014) proposed the

variational Bayesian approach (VBA), a toolbox to perform robust model-based analysis of empirical data and parameters estimation based on Bayesian maximum likelihood. Once selecting a particular model against others, *model falsification* is also an essential practice in computational neuroscience (Palminteri, Wyart, and Koechlin, 2017). This means that for assessing the selection of one model over another, it is necessary to demonstrate that the other one cannot reproduce the behavioral trend of interest while the selected one can. This same falsification procedure must also be applied when selecting a set of parameters for the model. R. C. Wilson and A. G. Collins (2019) then published an introductory guide and tutorial to computational modeling good practices, and among them, relevant methodology and examples on model parameters fitting.

The classical strategies the already mentioned *Maximum-Likelihood* (Bard, 1974) or Bayesian estimators, such as *Log-likelihood* estimator (Carrera and Neuman, 1986) and *Weighted Least-Squares* estimator (Beck and Arnold, 1977). More recently, in particular, when the computational models have more than 4 or 5 parameters, *evolutionary strategies* or *genetic algorithms* (Yu and Gen, 2010) seem to be efficient and robust in solving these nonlinear optimization problems (Liénard, Guillot, and Girard, 2010). Darwin's theory inspires these algorithms on natural evolution, and they have been proven to be very robust in solving complex and nonlinear problems (Bäck and Schwefel, 1993). By emulating the biological process of natural evolution, a randomly initialized *population* of individuals (*i.e.* possible problem solutions, set of parameters) is gradually improved throughout *generations* by *recombination*, *mutation* and *selection*. The latter selection procedure is based on the definition of a *fitness function* which determines how good an individual is. In the vast majority of cases, the fitness of an individual is computed on its "phenotype," meaning on the "expression" (*i.e.* behavior) of its "genotype" (*i.e.* the parameters themselves to be optimized). To assess an individual's quality, the concept of *domination* is adopted. An individual or solution X_1 dominates the solution X_2 if $X_1 \succeq X_2$, meaning that $f(X_1) \geq f(X_2)$ if f has to be maximized, or $f(X_1) \leq f(X_2)$ if f needs to be minimized, with f being a single-objective optimized function $f : S \to \mathbb{R}$, where S and $\mathbb{R}$ are the search and the real space respectively (Hao et al., 2019).

The main reasons evolutionary algorithms are broadly applied nowadays to solve problems, such as complex parameter estimation, lies in their conceptual simplicity and strong robustness against nonlinear and chaotic practical problems, such as the biological evolution itself (Fogel, 1997). Classical search and optimization algorithms, as the ones cited above, often fail to find satisfactory solutions to real complex problems, but the application of strategies inspired by natural evolution that do not rely on any a priori knowledge or static conditions could often generate optimized results, simply with the definition of a fitness function. Moreover, evolutionary computation is an embarrassingly parallelizable process, given that the evaluation of the fitness of each individual is independent of the fitness of all the others. This allows rapid scheduling of this computation in highly distributed computer architectures. In Sect. 3.2.3, we use the Covariance Matrix Adaptation - Evolution Strategy (CMA-ES, Hansen (2006)), which is one of the most used state-of-the-art evolutionary techniques to solve nonlinear-non-convex-black-box optimization problems. Its main characteristic lies in the iterative estimation of the covariance matrix, which defines the contours of a second-order model of the objective function (that does not need to be accurately known). In more detail, possible solutions to the problem are sampled from a multivariate gaussian distribution, and after evaluation, they are sorted by their fitness value. Then the multivariate gaussian distribution parameters (*i.e.*, the mean vector and the covariance matrix) are updated based

on the ranking of the solutions' fitness values. Moreover, its other point of strength is that, in many of its implementations (for example Auger and Hansen (2005)), it does not need a fine parameters tuning because the hyperparameters of the strategies, for example the population size, are usually derived by the dimensionality of the problem, for example from the number of parameters to evolve (problem dimension). The important parameter to tune is the initial step size σ_0. It determines the spread of the covariance matrix, which defines the shape of the gaussian distribution ellipsoid where the algorithm looks for new solutions. The mean vector of this gaussian distribution represents the best temporary solutions (Hansen, 2016).

An exciting branch of genetic computations that will be used in this thesis concerns *multi-objecive optimization* (Murata, Ishibuchi, et al., 1995; Coello, Lamont, Van Veldhuizen, et al., 2007). As a result of the evidence that many real-world problems have many objectives to be simultaneously optimized, an urge to apply evolutionary strategies also to solve multi-objective problems arose. When many behavioral aspects must be optimized (for example, the participant's choice and their reaction times), the use of broader strategies that imply a multi-objective optimization is needed (Viejo et al., 2015). From Schaffer (1985), the main procedure adopted in these kinds of problems is the identification of Pareto optimal solutions. These solutions are a set of trade-off sub-optimal individuals that are improved throughout the evolutionary process and include the non-dominated individuals in the multi-objectives space. The evolution will lead to the recognition of an approximated *Pareto front* of the multi-objectives, which usually conflict with each other (Deb, 2011). Once a final version of the Pareto front has been assessed, the best individual can be identified according to the problem and preferences. The principal elements for guiding a good multi-objective optimization with genetic algorithms are a proper choice for mutation, cross-over, and selection operators and also a satisfactory definition of the objective functions (Abraham and L. Jain, 2005).

The main Multi-Objective Evolutionary-Algorithm (MO-EA) adopted in the scientific contribution of this thesis is a new version of the so-called Non-dominated Sorting in Genetic Algorithms (NSGA, Srinivas and Deb (1994)). Differently from scalarization approaches and the Vector Evaluated Genetic Algorithm (VEGA), proposed by Schaffer (1985), the non-dominated sorting, introduced by Goldberg (1989), proposed to find stable, uniform, and reproducible Pareto optimal solutions, exploring in a no-objective-biased and uniform way the solution space. Then, Deb et al. (2000) proposes a faster version of NSGA, called NSGA-II, where they introduce a more elitist approach and reduce the overall complexity of the algorithm of an order of magnitude. A glimpse of an evolution cycle of NSGA-II is shown in Fig. 2.9. Initially, a population P of size N is randomly generated. Then at each generation, t, this population P_t is paired to its offspring population Q_t of the same size N. So, the whole $2N$ individuals are then organized in a new population R_t by non-dominated sorting into several fronts $F_1, F_2, ...$ and then their *crowding distance* is computed. The sorting and selection are performed "by fronts", meaning that the individuals belonging to F_1 are the best in the objectives space, and none among them dominate the others. On the other side, the ones belonging to F_2 are all the individuals dominated only by those in F_1 but then dominating all the remaining ones. The crowding distance is used here as a metric to preserve population diversity because it measures, for each individual, the average distance of its two most neighboring solutions. Then, based on these sorting of the fronts, a new population P_{t+1} of size N is created by selecting all the individuals in the first fronts F_1, F_2. To preserve population diversity, the individuals with the larger crowding distance

in the last front F_3 are also picked. The remaining N individuals are rejected. Finally, through crossover, mutation and selection a new offspring population Q_{t+1} is created. Crossover and mutation operations are used to identify new individuals by combining or partially modifying the existing ones. In particular, crossover operations combine "genes" (=parameters) of two "parent" solution to create a new individual of the next generation, while mutation operations alter certain genes of an individual to generate a new one. Finally, P_{t+1} and Q_{t+1} will be then the $2N$ starting population for the next generation $t+1$.

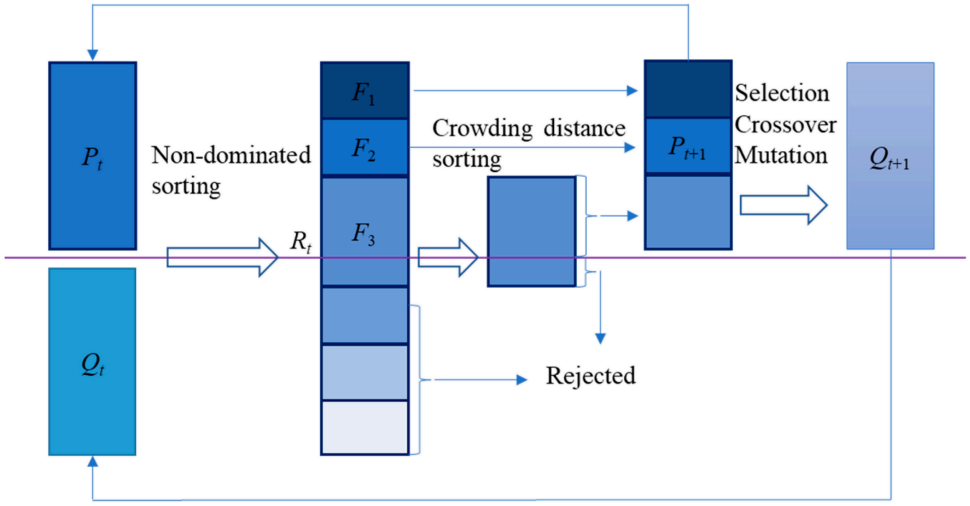


FIGURE 2.9: Scheme on the computations performed during a generation of NSGA-II. Figure reprinted from Jiang et al. (2021).

The last improvement to NSGA was made by Deb and H. Jain (2013) with the introduction of *NSGA-III* and of the concepts of *reference points*. Thanks to the establishment of some reference points in the solutions' space of the MO-EA, it is possible to better preserve solutions diversity and better guide the evolution and the search for the Pareto-optimal individuals. NSGA-II has already successfully been used to automatically optimize the parameters of different basal ganglia models for model selection (Wang et al., 2007; Liénard, Guillot, and Girard, 2010) and for the parameters estimation for a mean-field model of the basal ganglia against anatomical and electrophysiological data (Liénard and Girard, 2014). We will use the NSGA-III evolutionary strategy in Sect. 3.1.3-3.2.3 to estimate the best parametrization for our behavioral model based on the exploratory characteristics of different rodent datasets.

2.3 Neurorobotics

Recently, robotics and artificial intelligence have become very interconnected. Indeed robots have physical bodies which experience and act on the real world through their sensors and actuators. This embodiment is necessary to test the adaptive algorithms we design to model our learning behaviors. Otherwise, they are usually tested on simulated worlds, even if these can be very accurate.

Even if *neurorobotics* could classically refer to the integration of robotic systems with the human body for rehabilitation, recovery or augmentation purposes, the term is also concerning the domain that employs models of strategies used by the

brain to control a robotic device (Moxon, 2005). In this thesis, we will refer to neurorobotics from the latter point of view, in the specific domain of goal-oriented robotic navigation and spatial learning.

Mobile robotics has a long tradition and a wide range of daily applications, such as autonomous cleaner robots, rovers and mobile platforms for agriculture. In Sect. 2.3.1, we intend to present the main principles behind classic robotic space-mapping and self-location algorithms.

Then, the main neuro-controllers implemented in literature for bio-inspired goal-oriented robotic navigation are reviewed in Sect. 2.3.2. Here, our interest is in looking at the advantages that the inspiration from our nervous system, particularly from the role of place cells and grid cells in the hippocampus (Sect. 2.1.3), have brought to real robotic spatial learning.

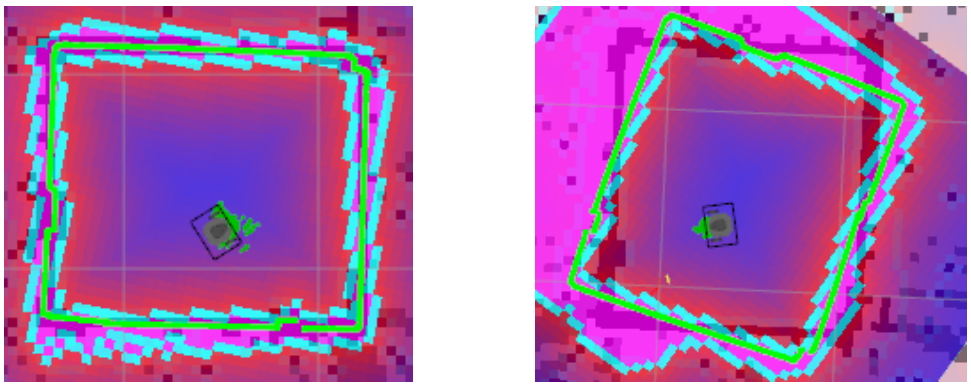
2.3.1 Robotic navigation and SLAM

Autonomous mobile robotics is strictly linked to the robot's ability to build maps of environments they have never explored before and to self-localise in these maps. This is a problem that is mutually recursive and indeed more complex than it may seem due to the inter-dependencies between building an accurate map and the proper localisation of the robot with respect to the same map. This problem is usually referred to as *Simultaneous Localization And Mapping (SLAM)* (Chatila and Laumond, 1985; Whyte, 2006; Aulinas et al., 2008). The mobile platform needs to be equipped with sensors for measuring its relative position with respect to external landmarks and with respect to the series of its previous positions. Such sensors can be cameras, proximity, light detection, and ranging (LIDAR) sensors for the estimation of the relative position of the robot with respect to the environment (in particular to specific landmarks) and motion sensors for the odometry estimation of the relative position of the robot with respect to its previous known location. Sensors commonly used for the odometry evaluation are Inertial Measurement Units (IMU), like accelerometers, gyroscopes and motor encoders.

The main techniques to simultaneously estimate the map and the robot position in that map are based on Bayes rule. These strategies are particularly efficient for the SLAM problem because of their ability to properly model probability distributions, uncertainty and noise. The main classical algorithms to tackle the SLAM problem are based on Kalman filters (KF, which are Bayesian filters with an assumption of normality of the data, which enables to use Gaussian distributions, Davison and D. W. Murray (2002)), Particle filters (PF, Montemerlo et al. (2002)) and Expectation Maximization (EM, Burgard et al. (1999)).

In the neurorobotic section of this thesis (Sect. 4.1), we will adopt the Rao-Blackwellized PF (RBPF) (Grisetti, Stachniss, and Burgard, 2007). This method creates grid maps thanks to laser range data. The algorithm computes a proposal distribution for the grid map based on the observation likelihood of recent sensors (in our case the LIDAR sensor), the odometry sensors and the scan-matching process. In Fig. 2.10, we show snapshots of the RBPF detecting the robot position with respect to the map. A different accuracy can be noticed in the case where the robot is following a straight line to accomplish a discrete navigation-step (Fig. 2.10a) and when it is rotating during its decision-making and inference phase (Fig. 2.10b, Sect. 4.2).

Rao-Blackwellized PF, very accurately compared to similar methods, defines a few particles (measurement estimations) and saves the computational time of creating unnecessary new ones. Rao-Blackwellized PF solves a full SLAM problem, that means that it can compute the joint posterior probability distribution $p(x_{1:t}, m|z_{1:t}, u_{1:t-1})$



(A) Forward motion SLAM (B) Rotation SLAM

FIGURE 2.10: RBPF in action on the Robot Operating System (ROS, Quigley et al. (2009)), recorded in the ISIR experimental setup. This screen capture shows how the real-time position of the Turtlebot3 burger robot from Robotis is estimated by the RBPF implementation on ROS, with the *Gmapping* package. The small green arrows around the robot represent the different particles' estimations of the current robot position. The green signal over-imposed on the arena's walls is the current information received by the LIDAR sensor. The pink area inside the two light blue lines is the subsequent estimate of the arena's borders over the previously memorized map (as a black trace in the background). A) During robotic forward motion, the algorithm estimates the walls better because they are tracked from the LIDAR from different map positions, but a poor estimation for the robot's current location. B) During the rotation phase instead, the robot position is better estimated than the contours of the map, which are temporally perceived as shifted compared to the real ones, due to the angular velocity of the robot.

over the whole trajectory $x_{1:t} = x_1, ..., x_t$ of the robot in the map m (Thrun, Burgard, and Fox, 2006). This posterior estimation is based on the observations from the current sensors $z_{1:t} = z_1, ..., z_t$ and the recent odometry measurements $u_{1:t-1} = u_1, ..., u_{t-1}$. The RBPF used the factorization $p(x_{1:t}, m|z_{1:t}, u_{1:t-1}) = p(m|x_{1:t}, z_{1:t}) \cdot p(x_{1:t}|z_{1:t}, u_{1:t-1})$ to first estimate the trajectory of the robot $p(x_{1:t}|z_{1:t}, u_{1:t-1})$ and then compute the estimate of the map $p(m|x_{1:t}, z_{1:t})$. The particle filter is used to estimate $p(x_{1:t}|z_{1:t}, u_{1:t-1})$; each particle represents the estimation of a possible trajectory $x_{1:t}$ and, for each of them, a map m is estimated. Once new sensory observations $z_{1:t} = z_1, ..., z_t$ and $u_{1:t-1} = u_1, ..., u_{t-1}$ are available, the set of particles and map couples are updated by using the new information and the weight of the particles: the so-called *importance weighting*. The novel proposal of the RBPF consists in taking advantage of the accuracy of laser range sensors (compared to visual sensors) to compute a Gaussian approximation of improved proposal for $\pi(x_{1:t})|z_{1:t}, u_{1:t-1}$ compared to the ones usually proposed by other PF-based methods. To sum up, RBPF proposes a dense grid map approach relying on a landmark-based SLAM.

In the following section, we will review the main computational proposals for transferring the principles behind self-location, mapping and spatial memory and learning from the mammals neural system to robotic navigation.

2.3.2 Neuro-inspired models for robotic navigation

We have already seen, in particular in Sect. 2.2.4, that the comprehension and the inspiration from neural mechanisms that regulate memory and learning in mammals can bring new insights to developing efficient strategies for artificial agents or robots which could be adaptable and time saving.

One of the first examples of parallelisms between rodent neural system for navigation and mobile robots' control architectures was presented by Touretzky, Wan, and Redish (1994). They proposed a robotic self-localization system inspired by the formation of place cells in the hippocampus (Sect. 2.1.3). By integrating the information regarding the distance and the egocentric orientation of the robot with respect to external landmarks with path integration (similarly to the principle with which the vestibular system helps mammals in the identification of a directional reference framework (McNaughton, Knierim, and M. A. Wilson, 1995), the proposed computational model can create and localize the system in fuzzy "external" states even though just partial information is available (either from visible landmarks or from self-perception sensors). This computational mechanism, together with the definition of "internal" states, where the robot position is identified compared to familiar *reference points*, allows for the allocentric localization of the robot, modeling the contribution of visual stimuli, head orientation and place cells in rodent navigation.

Following these first results, Arleo and Gerstner (2000) brought bio-inspiration even further, by modelling the interaction between CA3-CA1 hippocampal place cells for decoding spatial information, and the nucleus accumbens, which drives locomotion actions (M. A. Brown and Sharp, 1995). They proposed a closed-loop model that begins with the elaboration of visual inputs, in a model of the superficial entorhinal cortex, and that integrates the proprioceptive odometry information, processed in a model of the medial entorhinal cortex, on hippocampal place cells-like representation. Once a redundant spatial representation has been built online during exploration, population vector coding is used to extract the robot position in the following navigation phase. Then, reward based-learning is used to adjust the connection from the place cells to the action cells in the nucleus accumbens to generate goal-oriented actions. In this case, the chosen action is also extracted by

population coding. Finally, by using q-learning (Watkins, 1989) as a RL strategy to learn the weights between place and action cells, their robot can cope with changes in the reward location during the experiments.

To solve the problem of SLAM, whose main classical solutions have been briefly explained in the previous section, M. Milford and G. Wyeth (2010) implemented on a real robotic platform the biologically inspired solution *RatSLAM* (M. J. Milford and G. F. Wyeth, 2008). RatSLAM proposes to self-localize the robot in an extended and changing environments employing *local view cells*, *pose cells* and *experience maps*. Local view cells are an array of rate-coded units which encode the content in the field of view of the robot; usually one cell represents a unique scene, but cell can be also simultaneously active and different degree. The local view cells are then connected to the pose cells modelled by a 3D continuous attractor network (CAN). These networks have been previously used to model the mechanism that allow hippocampal place cells to encode spatial information in rodents (Samsonovich and McNaughton, 1997). Excitatory inputs strongly connect pose cells to the other neighbouring cells and other group of cells in different layers. These clusters of strongly connected cells are called experience nodes. In this way, the subsequent alternation among experience nodes creates an *experience map* and forms a spatial representation from visual external landmark, the robot can internally compute its allocentric position with respect to the external map. Moreover, thanks to the path integrator, which works on odometry, and so on idiothetic spatial representation, place cells can shift their activity to different experience nodes to stabilize the allothetic representation of the robot, without directly dealing with the sensors' uncertainty.

Inspired by the rat model proposed by Dollé et al. (2010) (Sect. 2.1.2), Caluwaerts et al. (2012) proposed then a bio-inspired algorithm, able to perform self-localization and navigation based on the alternation between different strategies and had tested it on a rat-like robot (Meyer et al., 2005). Using a RL framework, the robot can recognize a familiar context and switch towards the best behavioral strategy between pure exploration, a *taxon* and a *planning* response. These strategies took inspiration from studies on mammals (Trullier et al., 1997) and the meta-controller which arbitrates between them is inspired by the role of the prefrontal cortex (PFC) in rodents (Miller, J. D. Cohen, et al. (2001), Sect. 2.1.2). Following the model's extension they proposed in Caluwaerts et al. (2012), the robot could then switches among an exploratory, a planning (model-based) and, a taxon (model-free) strategy. In this work, this model's extension can detect all the environmental landmarks and autonomously learns the relevant ones, instead that associating its movements' direction just to proximal intra-maze landmarks (as in Caluwaerts et al. (2012)). The possibility to switch between many navigation strategies constitutes a simple way to generate adaptive behaviors, particularly by modulating the exploration/exploitation trade-off. Learning the association between gating patterns and sub-parts of the task through a context-switching detector, is either a proposed computational mechanism for the role of rodents and primates PFC in context evaluation and retrieval of past related information, but also a valuable contribution in the search for new mechanisms to produce adaptive goal-directed navigation in mobile robots.

Given the growing interest in hippocampal cells and their ability to robustly and flexibly interpret spatial information (Sect.2.1.3), bio-inspired models mimicking their roles have become interesting for roboticists. Jauffret, Cuperlier, and Gaussier (2015) propose a model of the communication between place cells (PC) and grid cells (GC), through a compression mechanism, that encodes the visual sensory information in the PC, by condensing it in the GC. This compression mechanism, based on

neural field coding (Wittmann and Schwegler, 1995), results in a satisfying path integration strategy for their robuLAB 10 robot from Robosoft Inc. They also argue that this mechanism can computationally explain the connectivity from the cortex to the entorhinal cortex and play the role of compressing and compensating sensory information (Gaussier et al., 2007). As in Arleo and Gerstner (2000), redundancy results as one of the main features of robust spatial localization, with different levels of GC which project on fewer emerging PC in the hippocampus, allowing this deep brain structure to easier detect the transitions in the cortical activity (Gaussier et al., 2002).

The coexistence of multiple instrumental conditioning strategies for goal-directed behavior and, in particular for navigation (Sect. 2.1.2), is of particular interest also for adaptive mobile robotics (Caluwaerts et al., 2012). Maffei et al. (2015) presents a comprehensive framework of alternation among navigation strategies where relevant policies are extracted by memory consultation. This model is based on the Distributed Adaptive Control (DAC) cognitive architecture which proposed that goal-directed behavior is produced not just by an unique computation, but by multiple learning, memory and planning systems at the same time (Verschure, Pennartz, and Pezzulo, 2014)). Particularly interesting in the context of hippocampal replay (Sect. 2.1.4), they manage to partially model the emergence of reactivations showing that forward-shifted spatial representations happen largely at decision points to predict the consequences of the next actions.

Another recent work on implementing bio-inspired replay on a real robotic platform has been presented in Whelan, Prescott, and Vasilaki (2020) and Whelan et al. (2022). They implement a model of a network of CA3 place cells which reproduce reverse sequences of recent experience. They design a biophysical network model where controlled replay can be generated thanks to the intrinsic plasticity (Pang and Fairhall, 2019) of the recent active cells. Fig. 2.11 shows an example of network rates (on the top) and intrinsic plasticity (on the bottom) where the robot explores the 10x10 states environment for the first time, from the left bottom corner (Fig. 2.11a and zone 1) to the top centre (Fig. 2.11b and c and final zone 14 in Fig. 2.11a). At the reward state (Fig. 2.11c and Fig. 2.11a zone 14) the reverse sequence is then triggered over its recent experience action-steps.

The main brain functions related to the phenomenon of hippocampal replay seems to be linked to memory and learning (Sect. 2.1.4). Therefore, modeling hippocampal reactivations is important in understanding how the brain stores experience to model the outcome of sequences of actions or simply off-line updates the knowledge on the state-action relationship to a reward. This comprehension, as a consequence, would advance the development of artificial agents and robots which show spontaneous memory recalling and learning for better tackling tasks for which they do not have any prior knowledge.

Many bio-inspired models of goal-directed navigation have been tested on real robotics platforms, and some of them were able to generate phenomena that recall hippocampal replay (Maffei et al., 2015; Whelan, Prescott, and Vasilaki, 2020). What is missing in the literature and that constitutes our contribution in the field is indeed the evaluation of different replay strategies, in the framework of RL, as a mean of enhancing adaptive mobile robotics and also better understand the implication of the different strategies (namely MB and MF) at behavioral level. Our research focuses will be also in unveiling how the contributions of the tested replay strategies persist or change when switching from a completely theoretical simulation to a real robotic one 4.1.

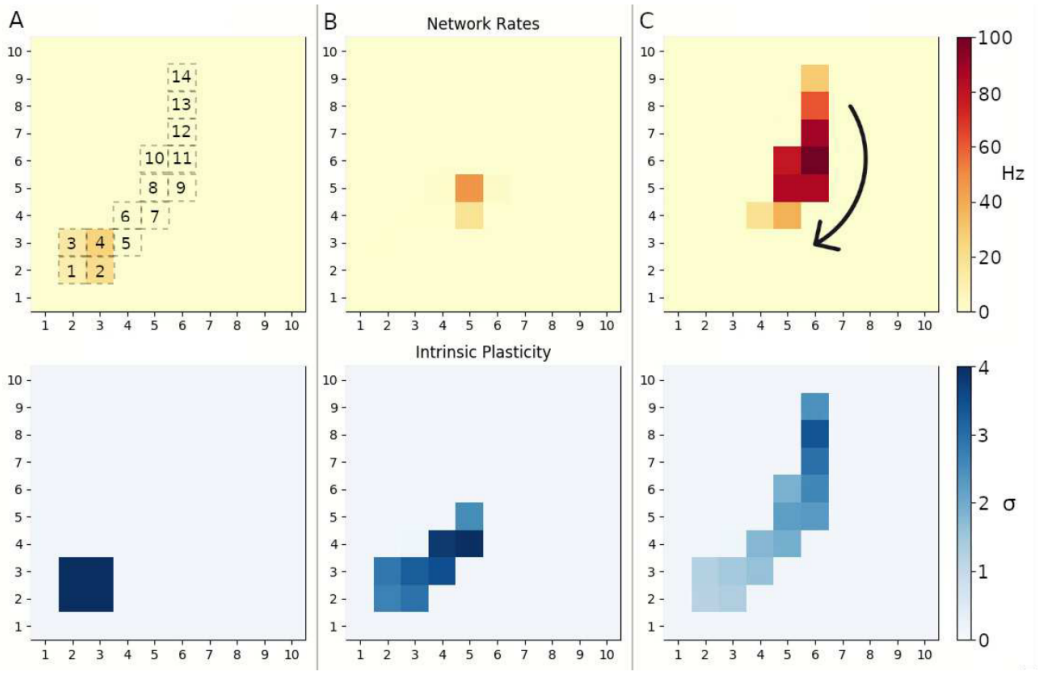


FIGURE 2.11: Network activity and intrinsic plasticity of the CA3 model running on the robot. A) Starting phase: from area 1 to area 14 (reward). At the beginning just a few areas close to the starting position are active and have a strong intrinsic plasticity. B) Exploration phase: the plasticity is more diffused along the past active cells. C) Reward phase and reverse replay: thanks to the reverse reactivation, the network activity propagates backward from the reward state (following the arrow) according to intrinsic plasticity. Figure reprinted from Whelan, Prescott, and Vasilaki (2020).

Chapter 3

Scientific contributions in neuroscience

As seen in the previous section, hippocampal reactivations nowadays are very interdisciplinary. Passing from neurophysiological recordings to behavioral studies and finally to computational models, recent research has drastically improved our knowledge on this topic (Sect. 2.1.4, Cazé et al. (2018), Foster (2017), Ólafsdóttir, Bush, and Barry (2018), and Whelan, Vasilaki, and Prescott (2019)). Nevertheless, we are far from a complete understanding of the generation and implications of such reactivations.

In Sect. 3.1, we describe newly identified common patterns in rodent free exploration, by modeling under the same constraints three rodent datasets. Our purpose is to unveil the existence of common free navigation patterns, which can explain rodent behavioral approaches into new scenarios or mazes across different timescales.

Sect. 3.2 will then take a step further, by merging the assessed model of Sect. 3.1 and employing it to model spatial learning tasks. In that case, we investigate, with our proposed computational model, how simple forms of RL-based reactivations could give insights on the emergence of divergent trends concerning the need for off-task (asleep) hippocampal reactivations based on the valence of the stimuli *i.e.*, positive or negative.

3.1 A data-driven computational model for free exploration in rodents

Exploration is an essential component in human and animal life. In nature, animals continually face new contexts, situations, and environments. In these new conditions, they need to look for resources and food and, simultaneously, avoid dangers and predators to survive. The interest in understanding exploratory behavior, and in particular exploratory decision-making, in animals is evident both from a neuroscientific and a robotic point of view. Taking inspiration from animal exploration to recreate a bio-inspired exploration can be useful for mobile robots to generate safer exploration paths and more robust localization maps.

In this section of our scientific contribution, we are going to investigate and discuss the following scientific questions:

- Which factors impact rodent decisions the most in the spatial exploration of new environments?
- Is it possible to model navigation in a novel environment as a value-based decision-making system inspired by rodent behavior?

- Is it possible to extend such a model to other rodents, different mazes, and longer exploratory sessions?

We will examine the above questions starting on observations and analyses we performed on rodent behavioral data, collected by our collaborators' research teams, during their neurophysiological experiments. The new computational model we propose tries to unveil common decision-making patterns among different rodents. To do so, we specifically fit the relevant behavioral components to each rodent subject's exploratory behavior. We suggest that common behavioral trends emerge if a consistent modelization of the environment and possible actions as a Markov Decision Process (MDP) is performed across different experiments, with different animals, and with different timescales.

As described in Sect. 2.2.3, the existence of common 'degrees of motion' in rodents has been studied and remarked on in the last 30 years. From our perspective, what is missing and can be useful to bridge the current knowledge on the subject to be easily deployed on reinforcement learning artificial agents or robots, is a common computational model which can account for MDP's decision-making, inspired by rodent free exploration.

In the following sections, the rodent exploration behavioral datasets that inspired the model's design and then were used to evaluate its generalization capabilities are presented (Sect. 3.1.1). Then we are going through the formalization of the computational model, in the details of its three behavioral components (Sect. 3.1.2). Finally, the model optimization process performed by evolutionary algorithms is explained in Sect. 3.1.3, where we also presented our results, and a discussion is proposed in Sect. 3.1.4.

3.1.1 Behavioral data

The design of this model is evaluated on a great amount of behavioural data that contribute to its novelty and reliability. The data concern mice and rats and range from novel exploration in a u-maze to open square maze and, to an original urban grid maze. The data was made available to us thanks to three important collaborations during this thesis (Sect. 1.3). The following section will explain in detail these three datasets and our data analysis.

U-maze

The first dataset contains trajectories of eight C57BL6jRj mice exploring for the first time and continuously for 15 minutes a u-shape maze of 1m x 1m size (Fig. 3.1, Bryzgalov (2021)). First, the mice position was tracked by an overhead thermal camera and then the center of mass of the hottest recorded point was registered as the mouse current position. The data were recorded at 15 Hz and Karim Benchenane and Dmtri Bryzgalov recorded them.

This first exploratory phase we are analyzing is considered a *habituation* phase for the animals to get familiar with the environment before starting the desired experiment. The research question behind these experiments regards studying the difference between the hippocampal replay activity in case of aversive or rewarding stimuli (Bryzgalov, 2021). For this question, the u-shape of the maze is key to better observe the presence of conditioned behaviours when the stimulus is given just on one of the two corridors' ends. In Fig. 3.2, the trajectories of the eight animals, extracted by Karim Benchenane and Dmtri Bryzgalov are showed.

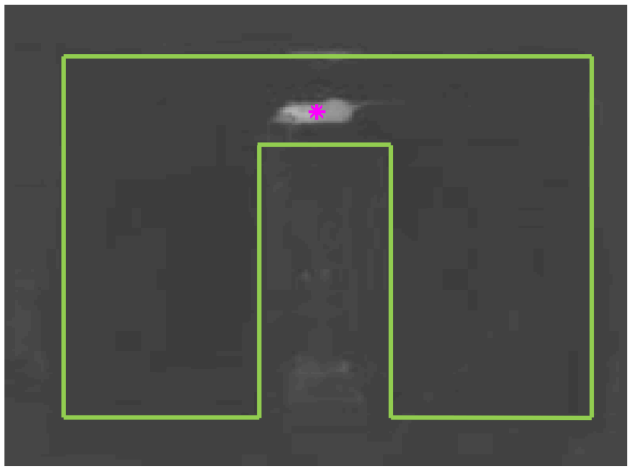


FIGURE 3.1: A caption from the recorded videos. From this and other similar videos, the trajectories of the body center (magenta star) of the 8 mice have been extracted. In green, the borders of the u-maze are highlighted. Figure reprinted from Bryzgalov (2021).

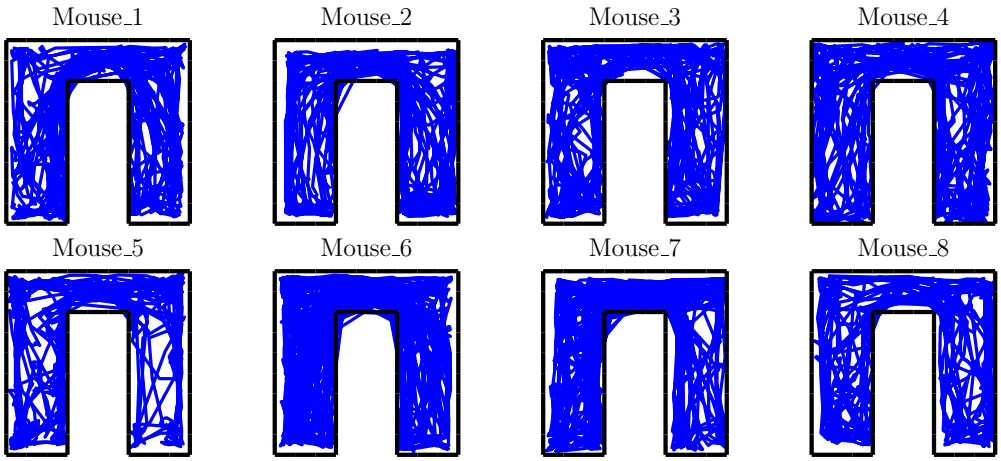
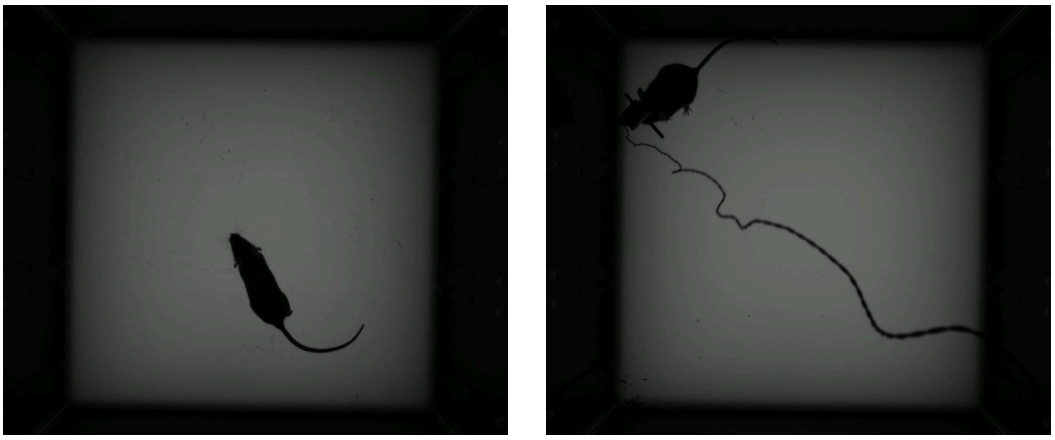


FIGURE 3.2: Trajectories followed by the eight mice during the habituation phase in the u-maze.

Square open maze

The second dataset also concerns mice behavioural data for habituation purposes. The concerned animals are female C57Bl/6J mice, between 8 and 24 weeks old. In that case, the trajectories contain the behaviour of these mice exploring for the first time a square open maze of size 30cm x 30cm, with 30cm high walls (Fig. 3.3). An overhead infra red camera records the data with a sampling time of 53ms and the trajectories of the body center of the mice has been extracted by our collaborators, using a DeepLabCut (Mathis et al., 2018) network, retrained for their experimental set-up. This data has been recorded and the trajectories extracted with DeepLabCut by Sebastian Haesler and Eléonore Schiltz.

The interesting aspect of this data is that these habituation phases are longer then the previous u-maze dataset (Sect. 3.1.1), they are 35 and 33 minutes long, respectively for the no-implanted and the implanted mice.



(A) No-implanted mouse (B) Implanted mouse

FIGURE 3.3: Captures from the habituation phases of the two mice exploring the square open-maze.

Even if the ratio between the maze and the animal size is three times smaller than for the previous maze (Sect. 3.1.1), the animals' behaviour could still show interesting pattern, in particular thanks to the long duration of the habituation phases. Fig. 3.4 shows the trajectories for the two mice of this dataset, individually. In particular, in this case, we have also a different condition between the two mice, since the second mouse was doing the habituation phase with the recording electrodes already implanted (Fig. 3.3b), while the first one was not (Fig. 3.3a). This aspect, together to the fact that our collaborators observed that the first mouse (Fig. 3.3a) was also more stressed and active than the other, result in a more uniform occupancy of the maze (Fig. 3.4, Mouse_1).

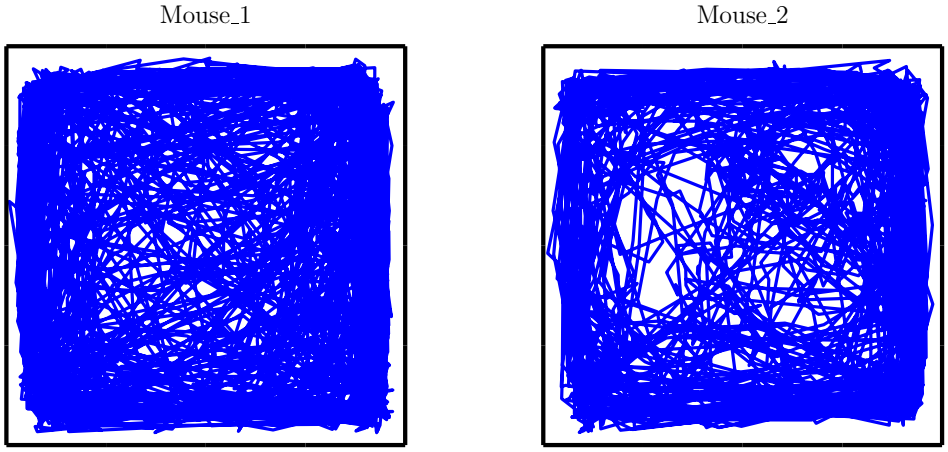


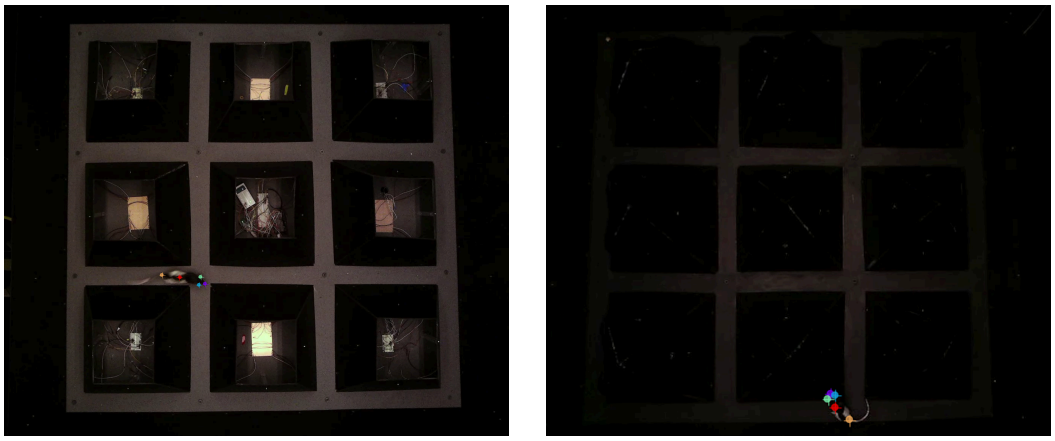
FIGURE 3.4: Trajectories followed by the two mice of the square open maze experiment.

Our collaborators used this open-maze set-up for investigating novelty versus familiar neural network encoding of the stimulus during an odor task (Sect. 1.3).

Grid maze

Finally, the last dataset collects the habituation behavior of 21 male Long-Evans rats (from Charles River), recorded over 3 years. Given that rats are bigger than mice, the maze's size is also bigger, around 2,1 x 2,1 m. The duration of these habituation sessions varies from 5 to 11 minutes. The data was recorded by Michaël Zugaro, Raphaël Brito, and Linda Kokou.

Fig. 3.5 shows the morphology of the custom-built maze, constituted by a 4 by 4 corridors grid, inspired by urban architecture. Through the months, some aspects of the maze have been adapted to the need of the researchers conducting these experiments; you can see it from the different appearance of the captures in Fig. 3.5a and Fig. 3.5b (Sect. 1.3).



(A) First experiments. (B) More recent experiments.

FIGURE 3.5: Captures from the videos from where the trajectories of the 21 rats have been computed for the grid maze. The colored dots indicates how the deep neural network identifies the body landmarks of the rat over the labelled one (colored crosses) that were assigned for the training of the network. The color are purple, green, blue, red and yellow, respectively for nose, left ear, right ear, body center and tail start.

This unique morphology for the maze has been created to have the possibility to deliver rewarding food in different crosses, and for temporally penalizing the animals' choice for passing through certain corridors (with an aversive sound). This study is also addressed to study the emergence of hippocampal replay in this complex and changing environment.

As you can notice, the color and quality of the videos are quite different from Fig. 3.5a to Fig. 3.5b. This is because in the first experiments, the central islands of the maze are opened, exposing the electronics that controls the reward and sound delivery. In the second version instead, our collaborators built pyramidal covers to protect the electronics and a sliding door to impede the animal from re-entering the starting home-nest (in the bottom left corner of the maze).

In Fig. 3.6 the trajectories we have extracted from the videos with DeepLabCut (Mathis et al., 2018) are shown. After a brief pre-processing on the videos for adjusting their size and orientation, we have extracted the trajectories by using a 50-layers deep residual network (ResNet) (He et al., 2016; Insafutdinov et al., 2016) implemented on DeepLabCut (Mathis et al., 2018), running on a graphics processing unit (GPU). This algorithm was able to properly generalize and identify the rats' nose,

ears, body center and tail start (Fig. 3.5) for all the 21 available videos, starting from a sample of around 100 video frames, labelled by hand.

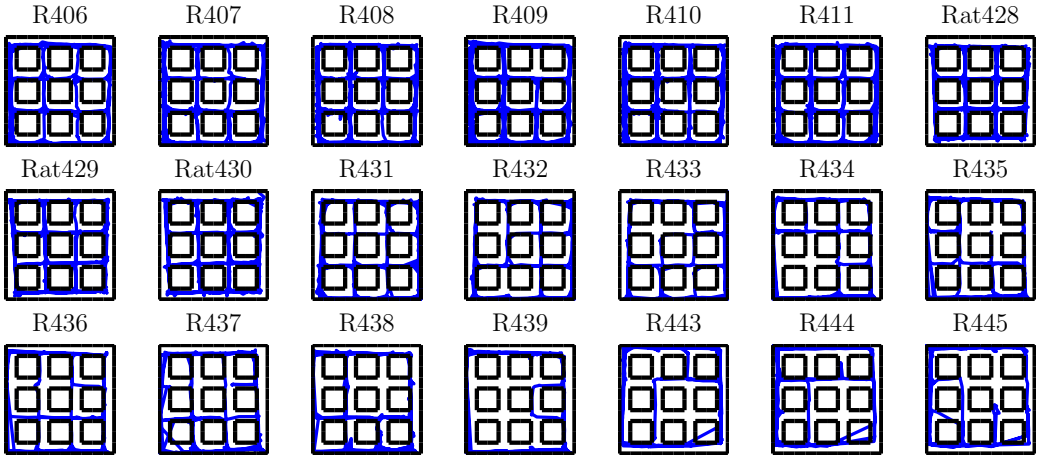


FIGURE 3.6: Trajectories extracted from the videos for all the 21 rats freely navigating the grid maze.

3.1.2 Free exploration computational model

Our aim in designing our computational model is to understand which factors influence rodent decision-making while exploring an environment for the first time. We have seen that this question has also been addressed by studying rodent behavior on three hours long sessions on a circular open arena by Fonio, Benjamini, and Golani (2009), and that the identification of exploratory behavioural patterns is possible (Drai et al., 2001). However, it is unclear if behavioural tendencies could be identified even when a home-cage is not present in the maze and if these trends could be consistent for more types of maze and exploratory sessions of different durations. A very recent study has also observed that mice's spontaneous spatial exploration is modulated by dorsolateral striatum (DLS) dopamine fluctuations (Markowitz et al., 2023). Their result implies that the same neural circuits and reinforce computational mechanisms at the base of goal-directed behavior could also explain rodent free exploration.

Compared to the previous literature models, our proposal's novelty consists of modelling, in a data-driven approach, the rodent as a decision-making agent, exploring an environment modeled as a Markov Decision Process (MDP), where each possible next discrete state has a defined value for the agent. In order to formally describe the behaviour of the mouse in its environment, we will use an atypical MDP in which we consider both continuous and discrete states, but only discrete time. The agent (mouse) has a continuous state CS (Eq. 3.1) and a finite set of continuous actions A (Eq. 3.2): the state is a pair with spatial coordinates and an absolute rotation, while the actions are a set of relative displacements.

$$CS = \mathbb{R}^2_+ \times [0, 2\pi] \tag{3.1}$$

$$A = \{(\Delta x_1, \Delta y_1), \ldots, (\Delta x_{N_A}, \Delta y_{N_A})\} \tag{3.2}$$

At each iteration step t the mouse has thus a continuous state $cs_t \in CS$ and chooses to perform a certain action $a_t \in A$ among the possible N_A actions. Moreover, the fact that the actions are relative implies that the agent, from the same starting positions and choosing the same action, can arrive in two different ending positions, depending on its orientation, as described by the deterministic transition function T:

$$\begin{array}{rccccc} T: & CS & \times & A & \to & CS \\ & (x,y,\theta) & \times & (\Delta x, \Delta y) & \mapsto & \left(x', y', \arctan\left(\frac{\Delta y}{\Delta x}\right)\right) \end{array} \tag{3.3}$$

$$\text{where } \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} + \begin{bmatrix} x \\ y \end{bmatrix}$$

The movement of the agent is actually constrained by the environment (maze) in which it acts. The maze has a discrete description: we divide the space in squared tiles of fixed size ts and consider all continuous positions in the same tile as the same discrete space using the relationship $s_t = discrete(cs_t)$ as described below.

$$S = \{1, \ldots, M\}^2 \tag{3.4}$$

$$\begin{array}{rccc} discrete: & CS & \to & S \\ & (x,y,\theta) & \mapsto & \left(\left\lfloor \frac{x}{ts} \right\rfloor, \left\lfloor \frac{y}{ts} \right\rfloor\right) \end{array} \tag{3.5}$$

In the description of S, for the sake of simplifying the notation, we consider only the case of squared mazes of size M (that is the case for all the datsets we are going to analyze in this thesis) and that the continuous position only lies within the extent of the environment. As the maze can contain tiles which the mouse cannot visit due to walls or holes, we also need to define a function that tells us which tiles are visitable, which depends on a set of visitable states V that is different for each environment:

$$V \subseteq S \tag{3.6}$$

$$\begin{array}{rccc} visitable: & S & \to & \{0,1\} \\ & s & \mapsto & \begin{cases} 1 & \text{if } s \in V \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{3.7}$$

By using the discretization and the *visitable* function we can restrict the possible actions for each state to the subset of actions $A(cs)$ that have a visitable endpoint inside the maze:

$$A(cs) = \left\{a \in A \mid s' = discrete(T(cs,a)) \in S \wedge visitable(s') = 1\right\} \tag{3.8}$$

The subset $A(cs)$ is the set of actions that we will consider during the MDP simulation.

In real experiments, the conditioning signal is given when the mouse is in a certain position, but only at specific points in time. This can be formalized by defining a set of reward values RS that depends on these variables (discrete position $s = (x,y)$ and discrete time t) and the corresponding reward function R:

$$RS \subset S \times \mathbb{N} \times \{-1, 1\} \tag{3.9}$$

$$RS = \{(x_0, y_0, t_0, r_0), \ldots, (x_{N_R}, y_{N_R}, t_{N_R}, r_{N_R})\}$$

$$\begin{aligned} R: \; S \times \mathbb{N} &\rightarrow \{-1, 0, 1\} \\ (x, y)_t &\mapsto \begin{cases} r & \text{if } (x, y, t, r) \in RS \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.10}$$

where N_R is the number of rewards for a specific session. As for the set of possible actions $A(cs)$ (Eq. 3.8), the set of rewards RS depends on the particular experiment. In the rest of the manuscript we will use either R or P interchangeably when referring to reward or punishment signals

To summarize: at each discrete time t, the agent has both a continuous state cs_t and a discrete one $s_t = discrete(cs_t)$; it will then transition to a new continuous state cs_{t+1} using an action $a \in A(cs_t)$; to choose such an action, the decision making process will be based on the value assigned to the possible next discrete states $s'_t \in \{s \in S | \exists a \in A(cs_t).s = discrete(T(cs_t, a))\}$.

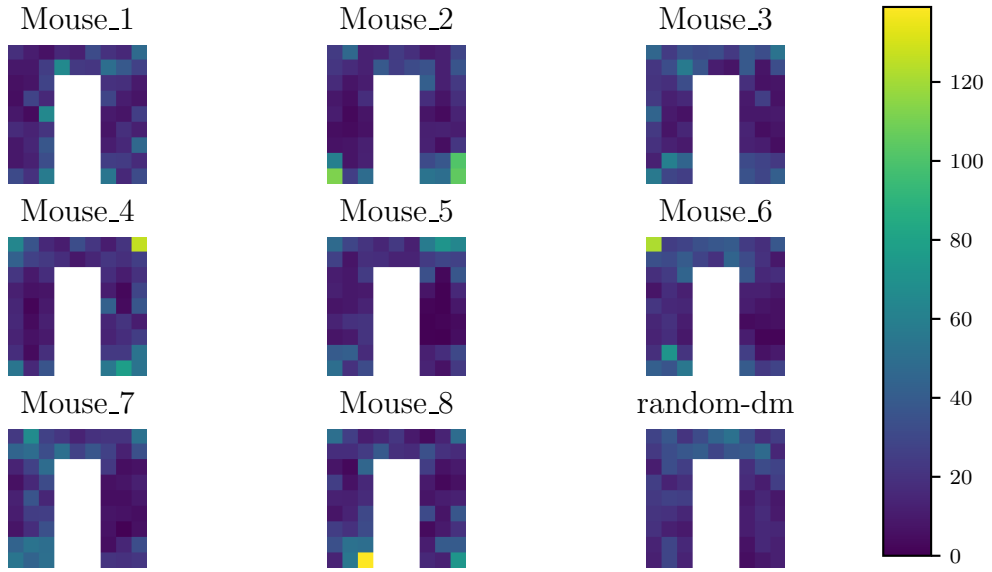
All of our available datasets are organised as x and y coordinates (in m) of the Center of Mass (CoM) of the animal recorded for each time sample (in ms). To define states and actions for the definition of the MDPs, we looked directly at the behaviour of the rodents.

On the first hand, the discretization of the position in states was done based on the ratio between the animal size and the maze size, but also targeting a clear definition of areas of particular interest in the maze. In particular, the corners and areas next to the walls are of particular interest for rodents during early exploratory phases (Drai et al., 2001; Fonio, Benjamini, and Golani, 2009), and our aim was that the occupancy of these areas could be clearly identified. In fact, the occupation of these areas is associated to thigmotaxis that has been identified as a proxy for anxiolytic behavior (Treit and Fundytus, 1988). In Fig. 3.7-3.8, we can see that, for all the three datasets, most of the rodents show a particular preference for staying in corners and close to walls in terms of occupation compared to the case of an exploration performed by a *random decision-making simulated agent* in the same MDP (random-dm in Fig. 3.7-3.8).

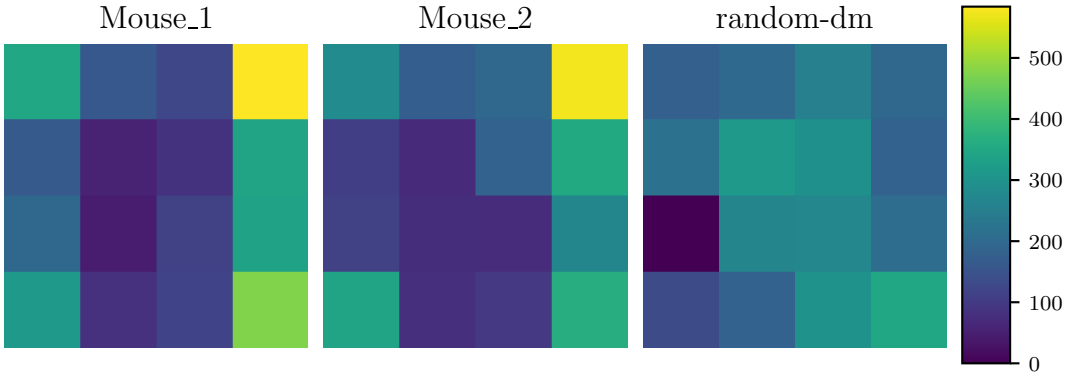
From this point of the thesis, when we refer to a random decision-making agent or to a *random exploration*, we mean the simulated agent or the behavior which is generated by navigating the MDP describing one of the three mazes (u-maze, open square maze, and grid-maze) by having an uniform probability to choose one of the next possible actions $A(cs_t)$ (no behavioral value is assigned to the possible next states). The random decision-maker starts its exploration with the same initial position and orientation of the corresponding rodent and navigates for the same duration (number of discrete steps).

This first data analysis already reflects a behavioural pattern where rodents consider places closer to walls safer when exploring new environments (Tchernichovski, Benjamini, and Golani, 1996; Fonio, Benjamini, and Golani, 2009). Indirectly, this is also a measurement of the animal's anxiety level, which prefers the perceived safer corners of a maze to more open areas. At the beginning of an exploratory session, these areas are considered safer because they can provide more tactile information to the animal and help it to better localize itself in the new environment (Touretzky and Redish, 1996).

On the second hand, a discretization in 600 ms time-steps was applied to all of the data, since the actions that the animals usually did in that time-step were from 0



(A) U-maze



(B) Square open maze

FIGURE 3.7: Occupation maps once the data have been discretized in time and space. The colorbars represent the number of visits of each discretized state, and are different for each maze, given the different time-scales of the experiments. U-maze and grid-maze (Fig. 3.8) experiments have similar durations, while the experiments performed in the square open-maze are longer (as described in Sect. 3.1.1).

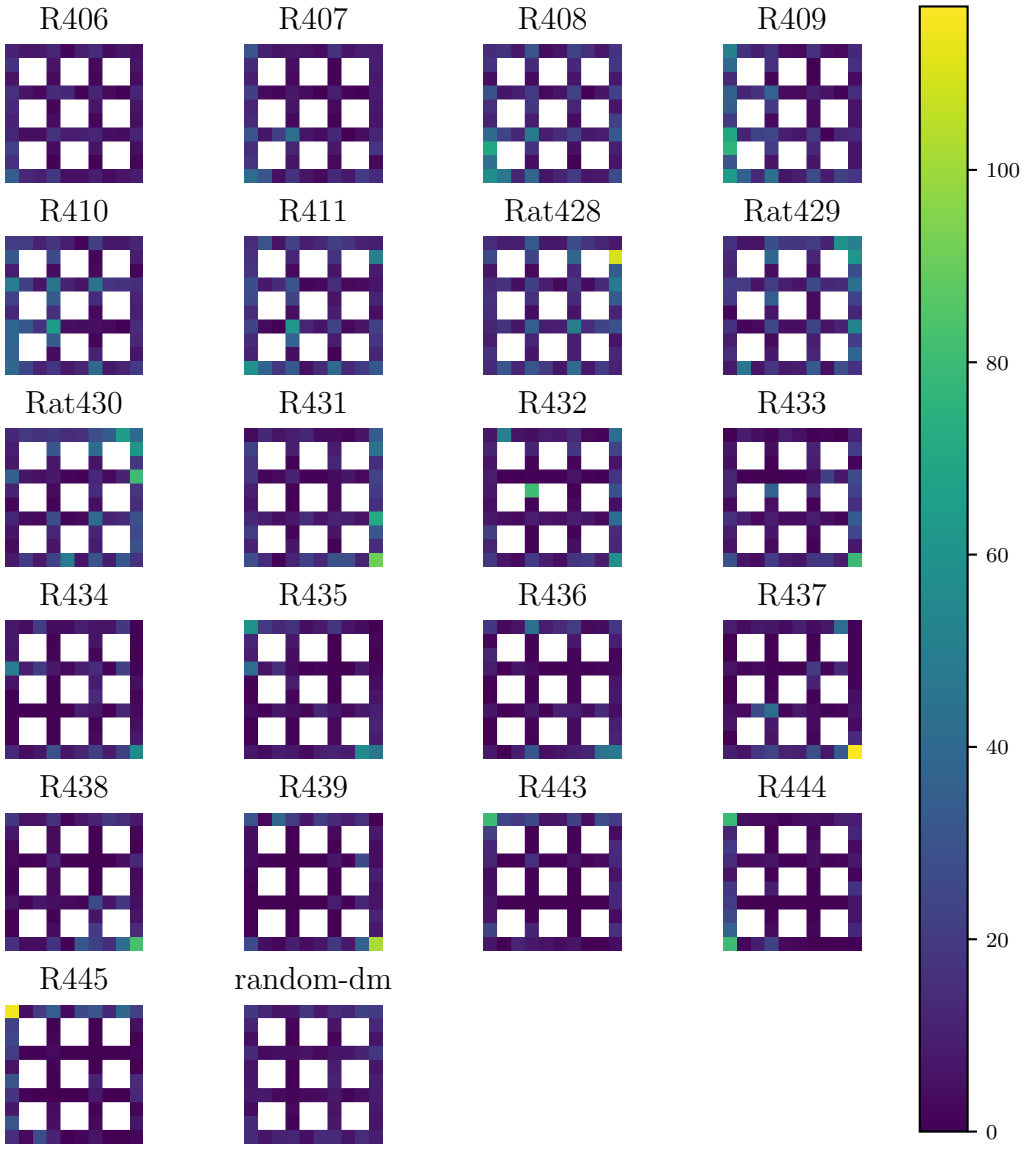


FIGURE 3.8: Occupation maps for the grid-maze dataset once the data have been discretized in time and space.

to 2 tiles away. This means that the 600 ms time interval was approximately correct to describe a decision-making process leading to a maximum of a two tiles away movement. A very similar timescale was also recently identified by Markowitz et al. (2023) as a median duration for the main behavioral sequences of mice spontaneous free exploration.

The identification of the possible next actions is computed for all the datasets by applying the k-means clustering algorithm (Lloyd, 1982) with the k-means++ seeding algorithm technique (Bachem et al., 2016) by considering the relative occupation of the animals in the next time-step. More in detail, the aim is to maximize the distance between a cluster of the most visited tiles and one of the least visited ones, among the tiles chosen for the next step. Since the current tile (the one where the rodent triangle is placed in Fig. 3.9) and the ones immediately in front of that one were occupied in a larger scale compared to the rest of the surrounding tiles, they are a priori considered as belonging to the group of the most occupied nest-steps next tiles. They are not used in the clustering process. Interestingly, for all the datasets, k-means identifies the next possible tiles (gray crosses in Fig. 3.9) symmetrically with respect to the current position and, as expected, showing a strong preference for keeping the current orientation in the mazes where corridors are present (Fig. 3.9 u-maze and Fig. 3.9 grid-maze). Thus, the directional preference gradually decreases from the grid maze (Fig. 3.9 grid-maze) to the squares open-maze (Fig. 3.9 square open maze), passing by the u-maze (Fig. 3.9 u-maze).

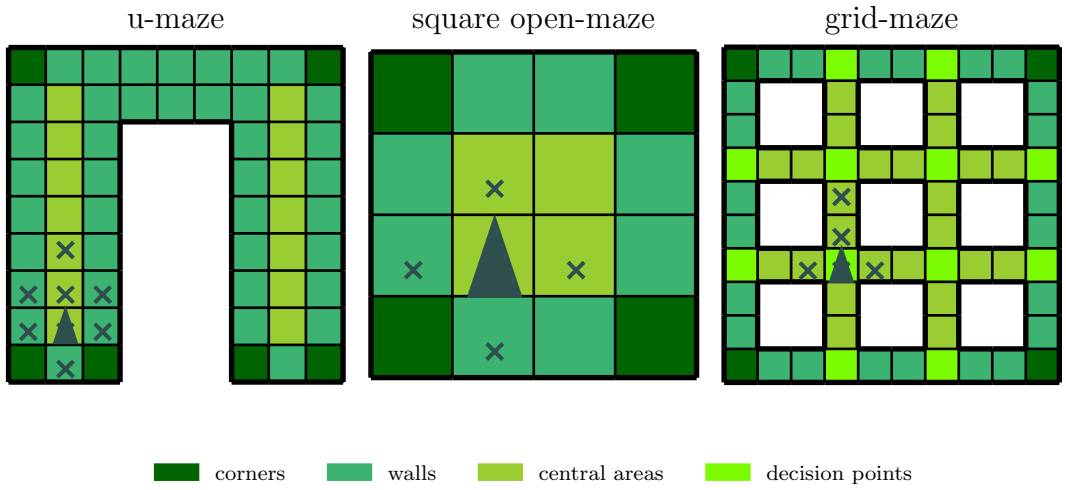


FIGURE 3.9: Mazes' discretization and conversion into Markov Decision Processes. The tiles' color indicates the topological type of tile as described in the legend. Gray triangles represent an example for a rodent's position inside the maze and the gray crosses together with the gray triangle identify the possible next states from the current position.

After identifying the states and actions of the MDPs we will use for our model, we describe its relevant components which give values to the different parts (states) of the maze.

In our proposed model, each one of the possible next states s' (in gray in Fig. 3.9) at time t would have the following value $V_{free\_exploration}(s'_t)$ for free exploration:

$$V_{free\_exploration}(s'_t) = V_{safety}(s'_t) + V_{biomechanical\_cost}(s'_t) + \\ + V_{biomechanical\_persistence}(s'_t) \quad (3.11)$$

and this value is assigned to all the available next states which coincide with a feasible area inside the maze.

The model is intended to generally describe rodent free exploration in novel environments. It contains 9 or 10 parameters (10 in the case of the grid-maze, because it also has decision-points tiles) which are supposed to be optimized to capture the individual behavioural nuances of each animal. In the following description of the model, some parameters will be written in red to highlight the variables that will be optimized for each rodent as explained by using an evolutionary algorithm (Sect. 3.1.3).

The free exploration value $V_{free\_exploration}(s'_t)$ in Eq. 3.11 is composed of three main components:

- the Safety component
- the Biomechanical cost component
- the Biomechanical persistence component

Safety component

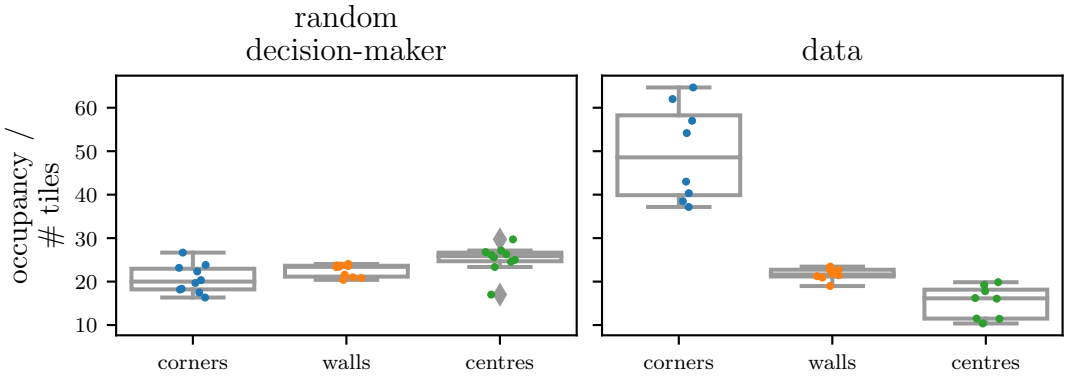
The safety component derives from what has been already observed in the literature: the fact that, when they first explore an environment, rodents tend to spend more time in very familiar and confined areas, such as the home-cage, the corners of a maze, or in areas which are closer to the walls (Tchernichovski, Benjamini, and Golani, 1996; Fonio, Benjamini, and Golani, 2009). Most importantly, our data confirms this preference (Fig. 3.10). In this figure, we represent the occupation of the different tile-types (corners, walls, centres ,and decision-points in the grid-maze, (Fig. 3.10c), over the number of that particular tile-type in a maze (Fig. 3.9).

Just for this safety component, we have two different definitions for the type of maze because the grid-maze, due to its morphology, presents an unique tile-type compared to the others: decision-points (Fig. 3.9).

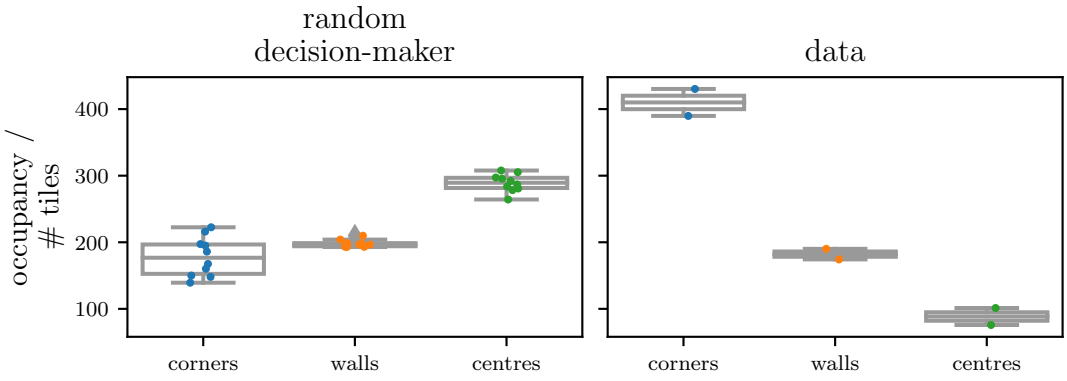
Concerning the first two types of maze, u-maze and square open-maze, the definition of the safety component in our model is based on the constraint of a hierarchical relationship between three classes of tiles which show decreasing level of occupational priority for rodent novel exploration, as shown in Fig. 3.10a and Fig. 3.10b. Corners would hold the maximal priority p_1 as the the safest spots in the maze, followed by walls which would have a priority that is a p_2 ratio of the one from corners. Finally, the central areas would have a p_3 ratio of the walls value. Thus, for modeling this hierarchical relationship observed in the data, we have:

$$V_{safety}(s'_t) = \begin{cases} p_1 & \text{if } s'_t \text{ is an external corner} \\ p_2 p_1 & \text{if } s'_t \text{ s next to an external wall} \\ p_3 p_2 p_1 & \text{if } s'_t \text{ is in a central area of the maze} \end{cases}$$

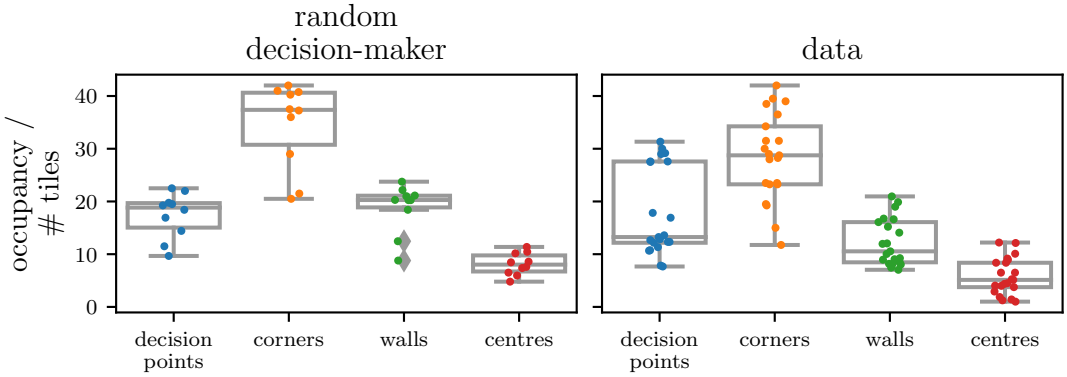
$$\text{with} \quad p1 \in (0, 10], \quad p2 \in [0, 1], \quad p3 \in [0, 1] \quad (3.12)$$



(A) U-maze



(B) Square open maze



(C) Grid maze

FIGURE 3.10: Histogram distributions of the tiles occupation for the data and a simulated random-decision making agent (10 repetitions with data coherent starting points, starting orientations and duration).

As we can see from the Fig. 3.10a and Fig. 3.10b, this hierarchy is not present in the case of random exploration, instead the trend is opposite, suggesting that this occupation bias could be crucial in describing rodent free exploration. The parameters ranges in Eq. 3.12 show the boundary values for the optimization of the parameter of this component. $p1$ presents a larger range, from 0 to 10 because its role is also to weigh the importance of the safety component with respect to the other two. Also the biomechanical cost (Sect. 3.1.2) and biomechanical persistence components (Sect. 3.1.2) have a parameter, ranging from 0 to 10, to measure their importance in the resulting behaviour of a particular rodent. $p2$ and $p3$ vary from 0 to 1 since they represent relative amounts of $p1$ and of $p1p2$ respectively.

Concerning the grid-maze, we also considered the decision points as relevant tile-type. In this case, we consider the relevance of the decision points to be independent from the other types of tiles. This because, even though looking at the data in Fig. 3.10c, it seems that the common trend for the decision points is to have a likelihood to be chosen between the wall areas and the corners, this is not always the case for all the rats (if we look at the points for each rat). For this reason, the extra model parameter $p4$ also ranges from 0 to 10 as $p1$ (Eq. 3.13).

$$V_{safety\_grid\_maze}(s'_t) = \begin{cases} p_1 & \text{if } s'_t \text{ is an external corner} \\ p_2 p_1 & \text{if } s'_t \text{ is next to an external wall} \\ p_3 p_2 p_1 & \text{if } s'_t \text{ is in a central area of the maze} \\ p_4 & \text{if } s'_t \text{ is in a decision-point of the maze} \end{cases}$$

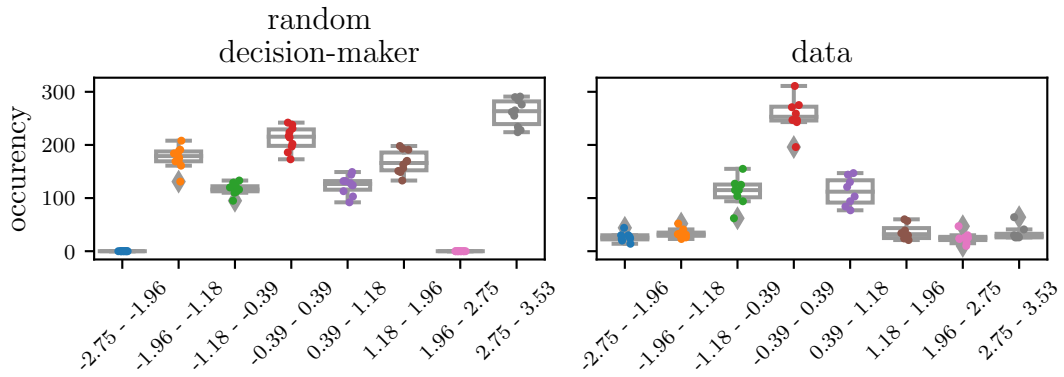
$$\text{with} \quad p1 \in (0, 10], \quad p2 \in [0, 1], \quad p3 \in [0, 1], \quad p4 \in (0, 10] \tag{3.13}$$

This maze morphology is particularly challenging for our model because the corridors are narrow compared to the animals' size (the corridors are one tile large) and present junction points. Thus, we have re-adapted our safety component for this particular case. Interestingly, the hierarchy between external corners, walls and open areas also holds in this dataset; corner areas and external walls seem to be prioritized anyway compared to the maze's internal corridors. The safety component for the grid-maze $V_{safety\_grid}(s'_t)$ is then defined as in Eq. 3.13. Strikingly, for the grid-maze, the same trend in the data also seems to be present in random exploration, implying that this occupation distribution could be derived from the maze morphology and not underlying any particular behavioural pattern.

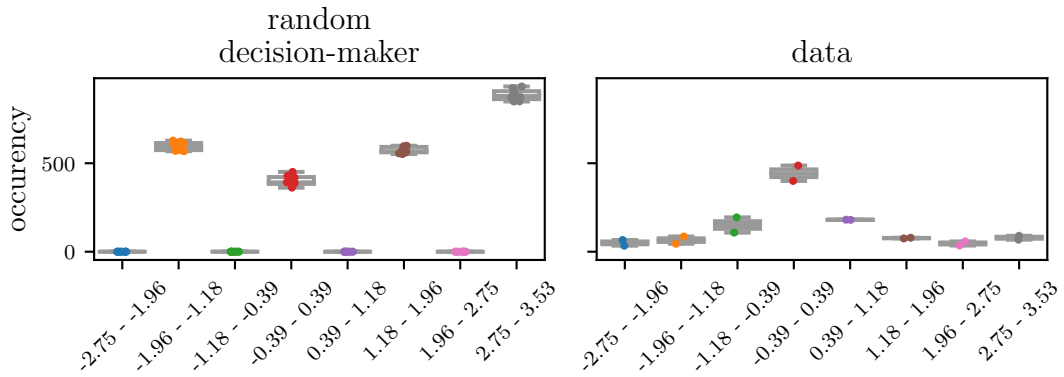
Biomechanical cost component

Instead, the biomechanical cost component describes rodents' directional persistence in *dynamic navigation*. With the term dynamic navigation we will refer to the case where the rodents were moving in the previous considered time sample. This behavioral tendency results from our time discretization of 600 ms. At this time scale, a biomechanical cost of performing high rotations exists for our datasets as showed in Fig. 3.11.

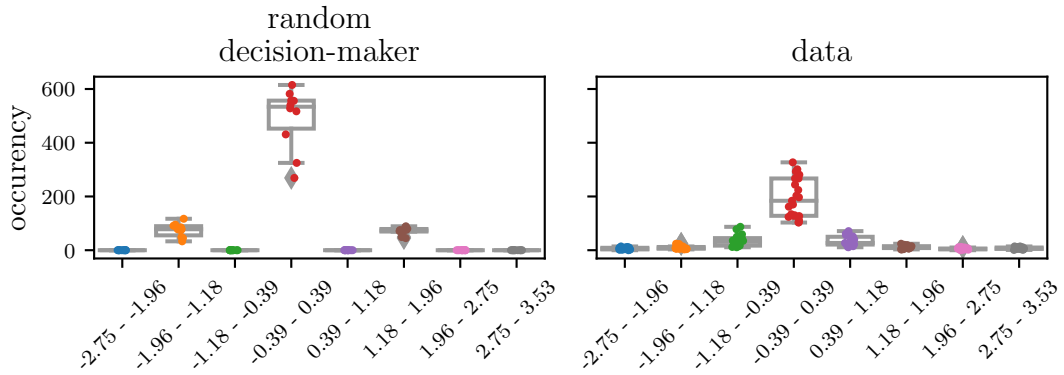
Here, the occurency of a dynamic relative rotations are represented. Bins 1 to 8 correspond to {[-2.75; -1.96], [-1.96; -1.18], [-1.18; 0.39], [-0.39; 0.39], [0.39; 1.18], [1.18; 1.96], [1.96; 2.75], [2.75; 3.53]} radians in relative rotations, so the red distribution correspond to the decision of keeping the same direction of movement and the gray



(A) U-maze



(B) Square open maze



(C) Grid maze

FIGURE 3.11: Histogram distributions of the dynamic relative orientations for the data and a simulated random-decision making agent (10 repetitions with data coherent starting points, starting orientations and duration). The x axis represents rotation intervals in radians.

one to go for the opposite direction. Each relative rotation bin is around 45 degrees large.

The global shape of the bin distribution is similar for all the mazes, but particularly relevant for the u-maze (Fig. 3.11a). Compared to random exploration, the dataset strongly prefers to keep the same direction of motion or narrowly deviate from it. The more the rotation angle, the less the preference for it. Interestingly, we can notice a strong difference also for the choice of going in the opposite direction of motion. In the maze, where the MDP allows for this choice (Fig. 3.11a and Fig. 3.11b), the opposite direction is chosen more frequently than the current one; situation that never happens in the data. Even when the maze is an open environment (Fig. 3.11b), the rodents prefer to go straight. While the MDP for the square open maze constrains the next actions to be one tile front, one back and, one tile $\pi/2$ left and right (Fig. 3.9), we decided to use the same number of bins and bin directions to have results comparable to the data and the other mazes. This implies that there are empty bins corresponding to directions that the MDP agent cannot take. Nonetheless, if we were to reduce the number of bins for the data of the square open maze, the histograms would still be very different: the distribution for a hypothetical bigger bin [1.96; 3.53] would still be significantly lower than the one for [-0.39; 0.39], thus it would still show a preference for going forward. On the contrary, in the case of the random decision maker, going forward is actually the least preferred option. This is because the next possible actions for the grid-maze instead (Fig. 3.11c) are not admitting to take the opposite direction of motion (Fig. 3.9 grid-maze) and for this reason we cannot see the same situation here. Moreover, in this case, just 4 actions are available compared to the 8 possible actions of the u-maze MDP. The maze morphology is more constraining for the animal's decision-making process. This, together with the fact that most of the states of the grid-maze force the rat to keep the same direction of motion, makes the simulated agent also show a very strong bias for moving forward (the red bin in Fig. 3.11c). This bias is stronger in random exploration because, in this case, the simulated agents are more active than the real rats and have just three possible relative directions to account for the agent's rotations.

In this case, we adopt the VonMises function $f(\theta|\mu,\kappa) = \psi\frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)}$ to model the distribution of the relative angular rotations when the animals are moving.

$$V_{biomechanical\_cost}(s'_t) = \begin{cases} 0 & \text{if } s'_t = s_t \\ \psi f(\theta|\mu,\kappa) = \psi\frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)} & \text{if } s'_t \neq s_t \end{cases}$$

$$\text{with} \quad \psi = 1, \quad \text{if } ||s'_t - s_t||_\infty <= 1, \text{and} \quad \psi \in (0,1] \quad \text{otherwise;} \quad \kappa \in (0,10] \tag{3.14}$$

Only the relative rotation, when the rodents are moving, is considered here. We assume that the inertia due to a greater relative rotation is consistently larger when the animals run from one point of the maze to another and not when they just rotate on the same position.

Further, looking at our data dynamics in Fig. A.3, Fig. A.4, Fig. A.5, in particular at the histograms about the tile-distance covered by the animals of the datasets for their next actions, we can derive that the usual number of tiles that the animals cover during a 600 ms time step, is a 1, and sometimes 2 tiles-distance. The tiles which are more than 1 step away from the current one are significantly less occupied in most cases. To capture this behavioural feature in the model, a ψ factor is added to this

component, to scale the preference for taking 2 or more tiles away actions compared to 1 tile away ones and of that its optimization range is from o to 1 (Eq. 3.14.) Thus, the biomechanical cost component for the next possible state s' at t is in Eq. 3.14. κ, which is a measure of the concentration, also represents this component's importance. For this reason, its evolution range is from 0 to 10. If κ is zero, the distribution is uniform, otherwise if κ is large, the distribution is very concentrated around its centre μ that is 0 radians with respect to the previous orientation of motion.

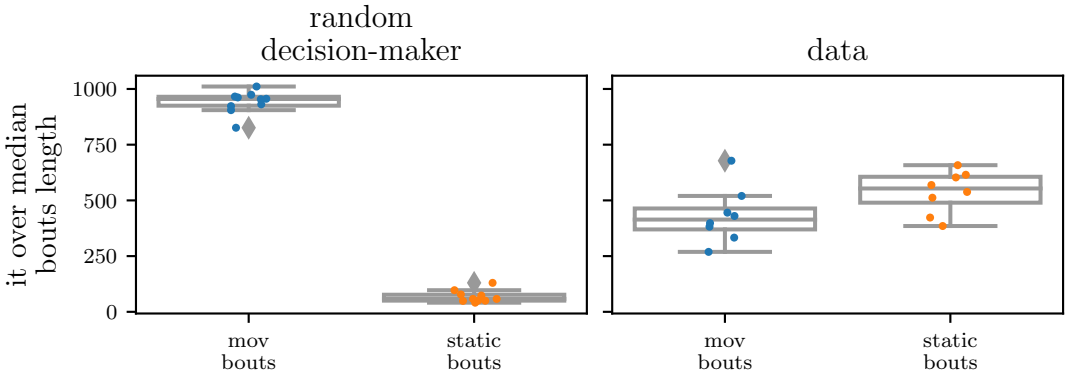
Biomechanical persistence component

The biomechanical persistence component models the exploration motion dynamics of rodents. The design of this component comes from the periodic nature of dynamic and static exploratory bouts in rodent navigation (Tchernichovski and Golani, 1995). This component is important to have a behavioural model that can also capture the exploration's dynamics and does not depend on the specific rodent's position in the current maze, to not over-fit the behavioural characteristics of an individual rodent on the specific maze morphology.

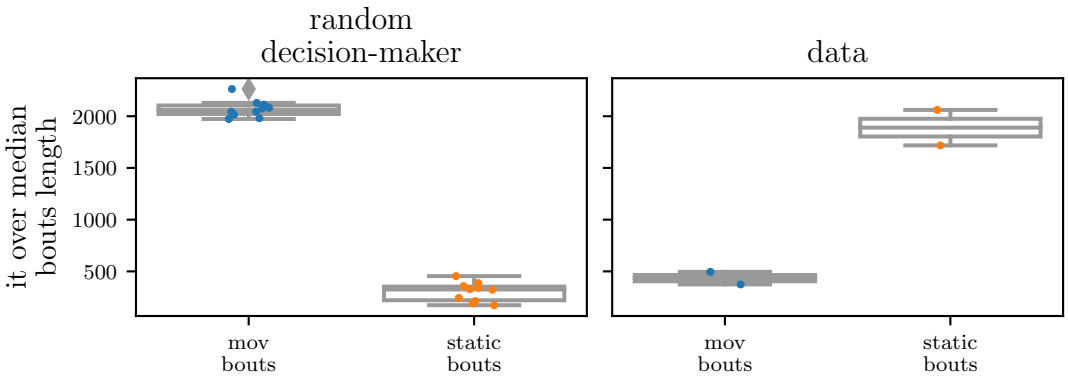
Observing our data, we saw that the series of dynamic and static bouts show significantly longer static bouts in the data than in what we observed if simulate random decision-maker rodents in the same framework (Fig. A.6, Fig. A.7, Fig. A.8). Representing differently this analysis, by computing how long the animals moved or did not move longer than the medians of both the dynamic ans static bouts (black lines in Fig. A.6, Fig. A.7 and, Fig. A.8), we can see a common result in all the three datasets and what it is impressive is that these results correspond to the opposite trend resulting when a random decision-maker is exploring in the same framework (Fig. A.6, Fig. A.7 and, Fig. A.8, random-dm). In fact, for all three datasets, we can see that the static bouts (red) are higher than the moving ones (green) for all the animals, but not for the random decision-makers. The red predominance of these figures suggests that, in the rodent behavior we analyzed, static intervals were usually longer than dynamics ones. An other interesting observation comes from the median length of the bouts for the random decision-maker: it is comparable to the one from the mice, in the u-maze case (black line in Fig. A.6), but lower than the one from most of the animals, for the square open maze and grid-maze (black line in Fig. A.7 and, Fig. A.8). This last consideration implies that, without any modeling constraint, an agent randomly exploring the proposed square open maze and grid-maze MDPs, would more frequently change from static to dynamic bouts and vice versa than the rodents. That's why we decided to describe the exploratory dynamic with this component. The idea behind this component is also to describe rodent internal state in novel exploration. As observed by Tchernichovski and Golani (1995) and Fonio, Benjamini, and Golani (2009), rodents tend to gradually explore new environments by alternating exploratory runs to static periods (in the case of Fonio, Benjamini, and Golani (2009), in form of comebacks to the home cage).

Fig. 3.12 suggests that rodents are usually less active compared to a random exploration in our proposed framework and that becomes more relevant when the exploratory phases are longer, like in the case of the square open maze (Fig. 3.12b).

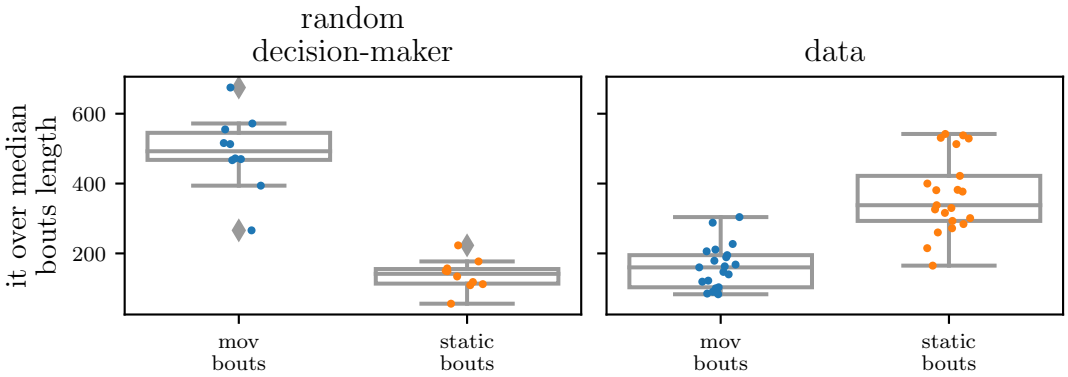
The biomechanical persistence value for one of the next state s' at time t is defined as follows:



(A) U-maze



(B) Square open maze



(C) Grid maze

FIGURE 3.12: Histogram distributions of the time spent moving or not moving more than the median duration of the dynamic and static bouts (black line in Fig. A.6, Fig. A.7, Fig. A.8), for the data and a simulated random-decision making agent (10 repetitions with data coherent starting point, starting orientation and duration).

$$V_{biomechanical\_persistence}(s'_t) = \begin{cases} \begin{cases} bp & \text{if } s'_t \neq s_t \\ W_{nm}bp & \text{if } s'_t = s_t \end{cases} & \text{if } s_t \neq s_{t-1} \\ \begin{cases} bp & \text{if } s'_t = s_t \\ W_m bp & \text{if } s'_t \neq s_t \end{cases} & \text{if } s_t = s_{t-1} \end{cases}$$

$$\text{with} \quad bp \in (0, 10], \quad W_{nm} \in (0, 10], \quad W_m \in (0, 10] \tag{3.15}$$

where the first case represents the persistence of remaining in motion while already moving bp and the urgency to stop while moving $W_{nm}bp$, while the second case represents the persistence of remaining static, always bp, and the urgency to start moving from a static condition $W_m bp$. The bp value is always the same in the two cases since it represents the relevance of this biomechanical persistence component compared to the other two (safety and biomechanical cost), thus its evolution range is from 0 to 10. Then, it possesses other two parameters: the non-moving weight Wnm and the moving weight Wm which represent the importance of changing the animal's motion state, respectively from dynamic toward static and vice versa, to static towards dynamic. In practice, this component models the persistence of a particular motion state of the animal, *i.e.*, dynamic or static, to persist. Wnm and Wm are also being optimized between 0 and 10 because they are not a proportion of other values, but an absolute value that denote an urgency of stopping while moving or, of starting moving while being still.

Decision-making

Finally, in our framework, the simulated agent decides on which, among the available next states, to occupy based on the soft-max distribution of the free exploration values $V_{free\_exploration}(s'_t)$ of these possible next states (Eq. 3.16, Sect. 2.2.1, Daw et al. (2006) and Khamassi et al. (2011)).

$$P(s'_t) = \frac{e^{\beta V_{free\_exploration}(s'_t)}}{\sum_{s \in N(s_t)} e^{\beta V_{free\_exploration}(s)}} \quad \text{with} \quad \beta \in (0, 10] \tag{3.16}$$

Thus, the probability of occupying the state s' at time $t+1$ is $P(s'_t)$, where β is the inverse temperature modulating the exploration/exploitation ratio in the behaviour. If β is low, the agent's decision will scarcely depend on the free exploration model, while if β is large, they will unquestionably rely upon $V_{free\_exploration}$. As for other parameters, β would be optimized between 0 and 10.

3.1.3 Model optimization and results

All the model parameters β, $p1$, $p2$, $p3$, ($p4$ in the grid-maze case), k, ψ, bp, W_m, W_{nm}, written in red in the previous section, are optimised for each rodent by using the multi-objectives evolutionary strategy Non-dominated Sorting in Genetic Algorithms-III (NSGA-III, Deb and H. Jain (2013), Sect. 2.2.5). The evolutionary ranges for these parameters have been explained in the previous section and the larger range is always between 0 and 10. This constraint has been imposed to the optimization process, because the values of the parameters in this range can already generate very diverse behaviours. If we assume a larger range, this diversity is lost

and, for values larger than 10, the parameters saturate and reproduce non distinguishable behaviours.

Fig. 3.13 shows the main dynamics of the NSGA-III parameters optimization process, for the example of Mouse 8 in the u-maze.

The parameter optimization minimizes three objective functions which are related to the three behavioral components of the model *i.e.*, safety, biomechanical cost and biomechanical persistence (Sec. 3.1.2). Each objective represents an error measurement between a behavior characteristic of the data and the same characteristic measured in ten simulations of the model, with the same parameters' configuration. To simulate the behaviour of a particular rodent, with a particular set of parameters, the computation is set so that the simulation of the artificial individual starts from the same position and orientation as the real rodent and explores for the same amount of time. Just in the case of the grid-maze, the starting orientation will always be $pi/2$ radians (like the orientation of the gray triangle in Fig. 3.9, grid-maze). This is because of our modeling constraints. Given the one-tile-large-corridors and the possible next actions identified from the data (Fig. 3.9, grid-maze), an initial orientation, which is not parallel to the corridors directions, would result in a completely static exploration. This static exploration is caused by the fact that the next possible actions would always drive the agent into walls (not available areas), so they would not be chosen and the agent would spend the whole experiment in the same starting position.

The three objectives functions to be minimized are described below.

The safety fitness F_{safety} evaluates the occupancy on the different types of tile (*i.e.*, corners, walls, central areas, and decision points) to make sure that the relevance of safety spots for a particular rodent is respected in its model's set of parameters. We define this metric as the Manhattan distance of the corner and wall occupation in the simulated model and in the data of that particular rodent (Eq. 3.17). The central area occupation is not to be considered since it is already completely defined by the other occupations because the simulated rodent would be exploring for the same time as the real one.

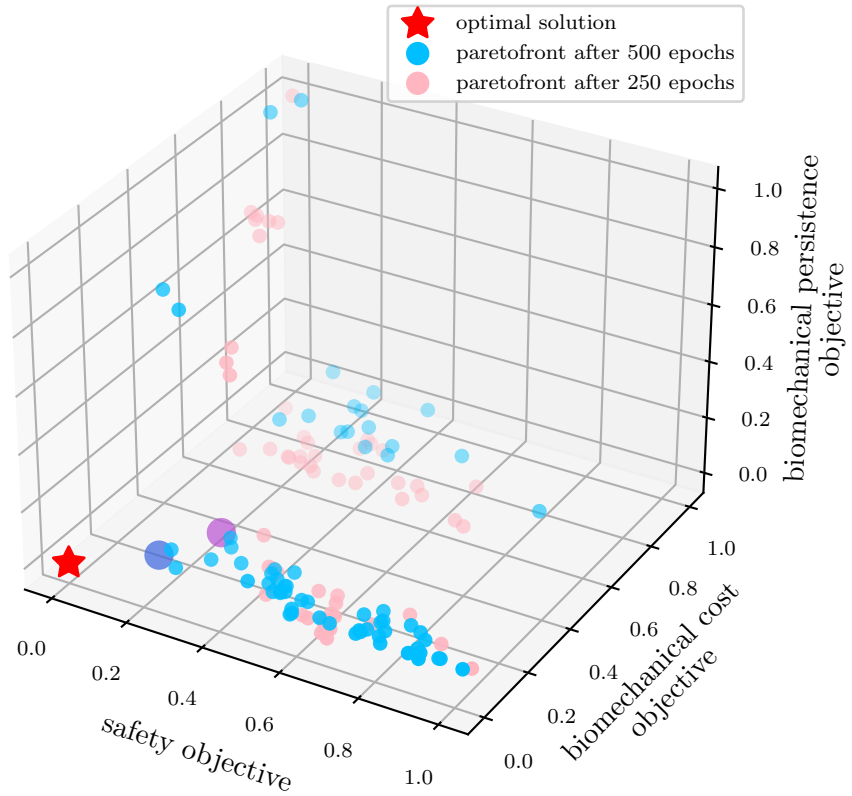
$$F_{safety} = \mid s_{corner_{data}} - s_{corner_{model}} \mid + \mid s_{wall_{data}} - s_{wall_{model}} \mid \tag{3.17}$$

with s_{corner} being the number of iterations where the agent was in the corner areas and s_{wall} the number of iterations where the agent was next to walls (Fig. 3.9). Since, in the case of the grid maze, we also have the decision points (Fig. 3.9, grid-maze) the definition of the safety objective will include also the decision points tiles, so that the following equation completely determines the central areas occupation:

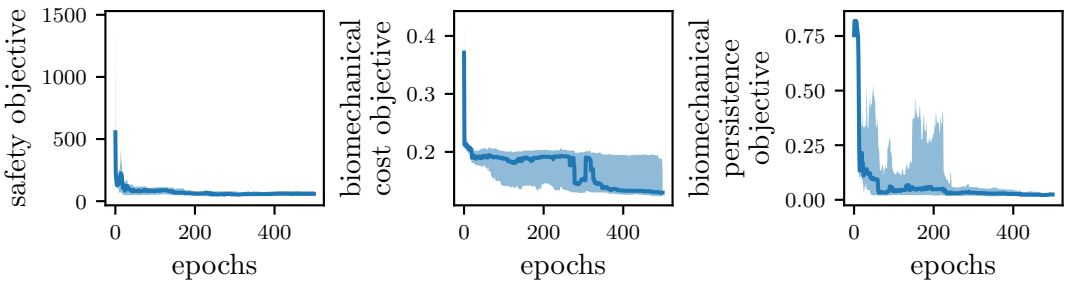
$$\begin{aligned} F_{safety\_grid\_maze} = \mid s_{corner_{data}} - s_{corner_{model}} \mid + \mid s_{wall_{data}} - s_{wall_{model}} \mid + \\ + \mid s_{decision\_point_{data}} - s_{decision\_point_{model}} \mid \end{aligned} \tag{3.18}$$

with s_{corner} being the number of iterations where the agent was in the corner areas, s_{wall} the number of iterations where the agent was next to walls, and $s_{decision\_point}$ the number of iterations where the agent was in the decision-points (Fig. 3.9, grid-maze).

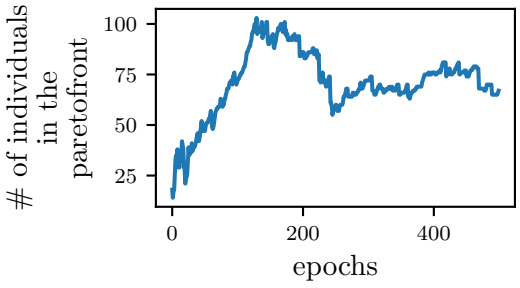
The second objective evaluates the directional persistence of the rodent. In this case, the normalized distance between the histograms describing the relative orientation distributions for the simulated free exploration model and the animal is computed. By computing the *normalized* distance the evaluation of this distance is



(A) Pareto front



(B) Objectives



(C) Pareto front size

FIGURE 3.13: Evolution dynamics for Mouse 8 in the u-maze.

performed more on the shape of the histograms than on the individual height of the bins, that is what we are the most interested in optimizing.

$$F_{biomechanical\_cost} = \left|\left| \frac{Bdata}{|\ Bdata\ |} - \frac{Bmodel}{|\ Bmodel\ |} \right|\right|$$

$$\text{with} \quad F_{biomechanical\_cost} = 0, \quad \text{if } |\ Bdata\ | = 0 \quad \wedge \quad |\ Bmodel\ | = 0, \quad \wedge$$
$$F_{biomechanical\_cost} = 1, \quad \text{if } F_{biomechanical\_cost} > 1 \tag{3.19}$$

where B is a function that counts the number of observations that fall into each of the disjoint categories (known as bins) and create the histogram, respectively for the data $Bdata$ and for the model $Bmodel$. So, if k is the total number of bins, in our case 8 (Sect. 3.1.2), and n is the total amount of observations, we will have that $n = \sum_{i=1}^{k} B_i$ and $|B|$ represents the cardinality of the histogram B.

Finally, the biomechanical persistence objective evaluates the exploratory dynamic of the simulated set of parameters for the model. In this case, the distance between the data's behavior and the one from the model is computed as in the biomechanical cost case (Eq. 3.19). Thus, $F_{biomechanical\_persistence}$ is the difference between vectors containing the times the simulated model and the data are moving or are static more than the median length of all the bouts.

$$F_{biomechanical\_persistence} = \left|\left| \frac{l\_b_{data}}{|\ l\_b_{data}\ |} - \frac{l\_b_{model}}{|\ l\_b_{model}\ |} \right|\right| \tag{3.20}$$

with $l\_b = (l\_mov\_b, l\_not\_mov\_b)$ and l_mov_b being the sum of the lengths of all the moving bouts which are longer than the median bouts length and $l\_not\_mov\_b$ being the sum of the lengths of all the not moving bouts which are longer than the median bouts length.

After a literature and empirical search, we select and apply the same genetic algorithm hyper-parameters for all the optimizations (Tab. 3.1).

max # gen	# ind	CXPB	MUTPB	RPs
500	50	0.8	0.01	12

TABLE 3.1: Hyper-parameters for NSGA-III. Here we present the maximum number of generations, max # gen, the population size, # ind, the cross-over probability, CXPB, the mutation probability, MUTPB, and the number of reference points, RPs for the pareto-front (Sect. 2.2.5).

Once the evolution process reaches its last 500th generation, a set of sub-optimal, equally dominant pareto-individuals is found. At this point, to identify the best set of parameters among the last pareto-individuals, the Chebyschev distance (Eq. 3.21, Cantrell (2000)) between the optimal point and all the normalized solutions in the pareto-front is computed.

$$Chebyschev\ distance\ (ind) = \max_{obj} |\ OptimalSolution_{obj} - ind_{obj}\ | \tag{3.21}$$

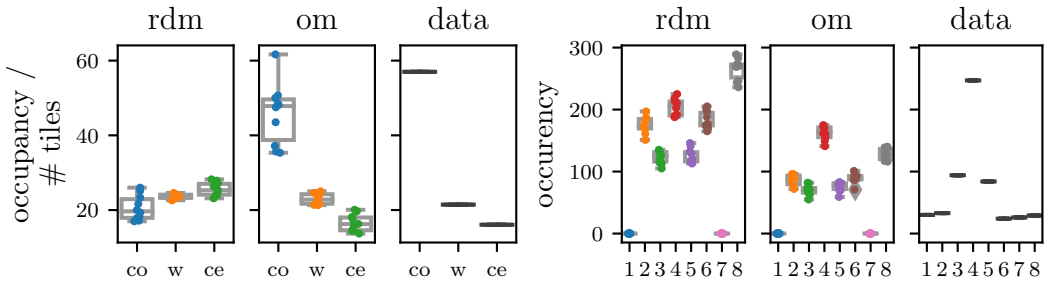
In Fig. 3.13a two examples of paretofront for the individual Mouse 8 in the u-maze are shown. The red star indicates the $OptimalSolution$ where all the objectives are minimimzed to zeros. In this point, all the three objectives for evaluating the three behavioural components are equal to the their value for the data of that particular rodent. The other dots in the figures shows how the paretofront evolves from generation 250 to generation 500; the largest dots indicate the best solutions for the two represented epochs, identified with the Chebyschev distance equation (Eq. 3.21). Fig. 3.13b shows instead a separate, but a more complete view of the three objectives being minimized over the evolutionary process, always for Mouse 8 in the u-maze. The three objectives are plotted in their corresponding value ranges. They converge to a minimized value, close to 0. In particular, the case biomechanical persistence objective consistently decreases and stabilizes its variance in the last epochs. Then, Fig. 3.13c shows, always for the same rodent, the number of pareto sup-optimal individuals identified at each epochs. Even if the size of the paretofront or the diversity of its individuals is not a stopping criterion in our implementation of NSGA-III, monitoring if the number of pareto-individuals stabilizes for several generations could indicate that the current set of pareto-individuals has converged to an optimal set of solutions.

After selecting the estimated best individuals at generation 500 (the large blue dot in Fig. 3.13a), we validate the behavior of this optimal model against random exploration and the data. Fig. 3.14 shows the behavioural comparison among the data, 10 repetitions of random exploration (rdm), and 10 repetitions of the optimized model (om), in terms of tiles occupation (A), directional preferences (B) and exploratory dynamics (C).

Regarding our behavioural measurements, the optimized model remarkably grabs the behavioral traits of the particular individual. The biomechanical cost metric has been the most difficult to be fitted by the optimization. However, the shape of the relative orientation distributions is robustly closer to the data compared to the distribution obtained by the random decision-maker. Looking at the overall maze occupation in this comparison (Fig. 3.14d), it is noticeable that the optimized model's occupations show a strong preference for corners and walls closer to the ones observed in the data. In the design of the proposed model there is no interest in trying to reproduce particular asymmetries that the data could show in the occupation of the two corridors or the exact same trajectories that the rodents followed. That is why the optimized model does not prefer to occupy the exact same corner as the one preferred by the mouse.

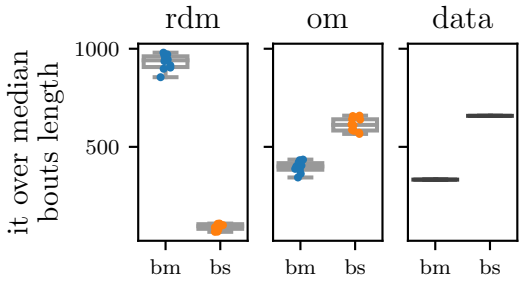
By performing a more global analysis of our results, the winning models for all the rodents are analyzed. Fig. 3.15, Fig. 3.16 and Fig. 3.17 show a statistical analysis to evaluate how better the optimized model can capture the behavioural components of the data compared to random exploration. In all the three figures, the distributions represent the computation of the three objectives (*i.e.,* safety, biomechanical cost and biomechanical persistence) for ten repetitions of the optimized model (blue) and of the random exploration (orange). Looking at the three objectives for each animal, we can see the behavioral difference between the data and the optimized free exploration model (blue distributions) or random exploration (orange distributions).

The results in Fig. 3.15 encouragingly show that the optimized version of the free exploration model can represent the three behavioural components significantly better than random exploration (just one exception for the safety objective of Mouse_1 that shows an unusually homogeneous exploratory behavior compared to the rest of the animals of this dataset, Fig. 3.7a, Mouse_1). We can say that the three objectives are significantly minimized compared to the behavior generated by the random

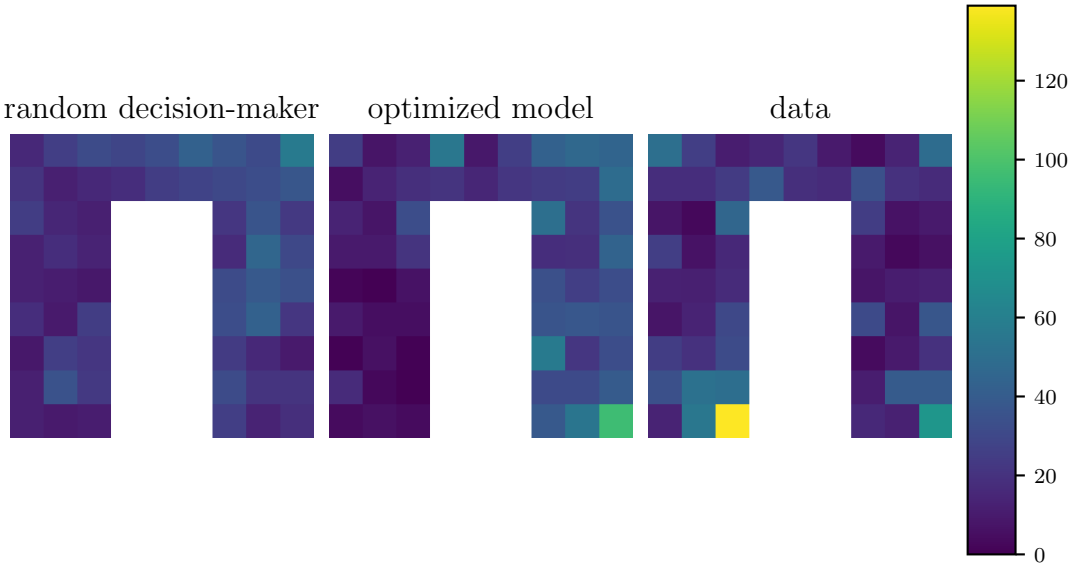


(A) Safety metric

(B) Biomechanical cost metric



(C) Biomechanical persistence metric



(D) Occupancy map

FIGURE 3.14: Optimizated model (om) behavior in comparison to the data and random exploration (rdm); example for Mouse 8 in the u-maze. A) Safety metric (occupancy for corners (co), walls (w) and central areas (ce)). B) Biomechanical cost metric (bins 1 to 8 correspond to {[-2.75; -1.96], [-1.96; -1.18], [-1.18; -0.39], [-0.39; 0.39], [0.39; 1.18], [1.18; 1.96], [1.96; 2.75], [2.75; 3.53]} radians in relative rotations). C) Biomechanical persistence metric (moving (bm) and static bouts (bs) over the median bouts length).

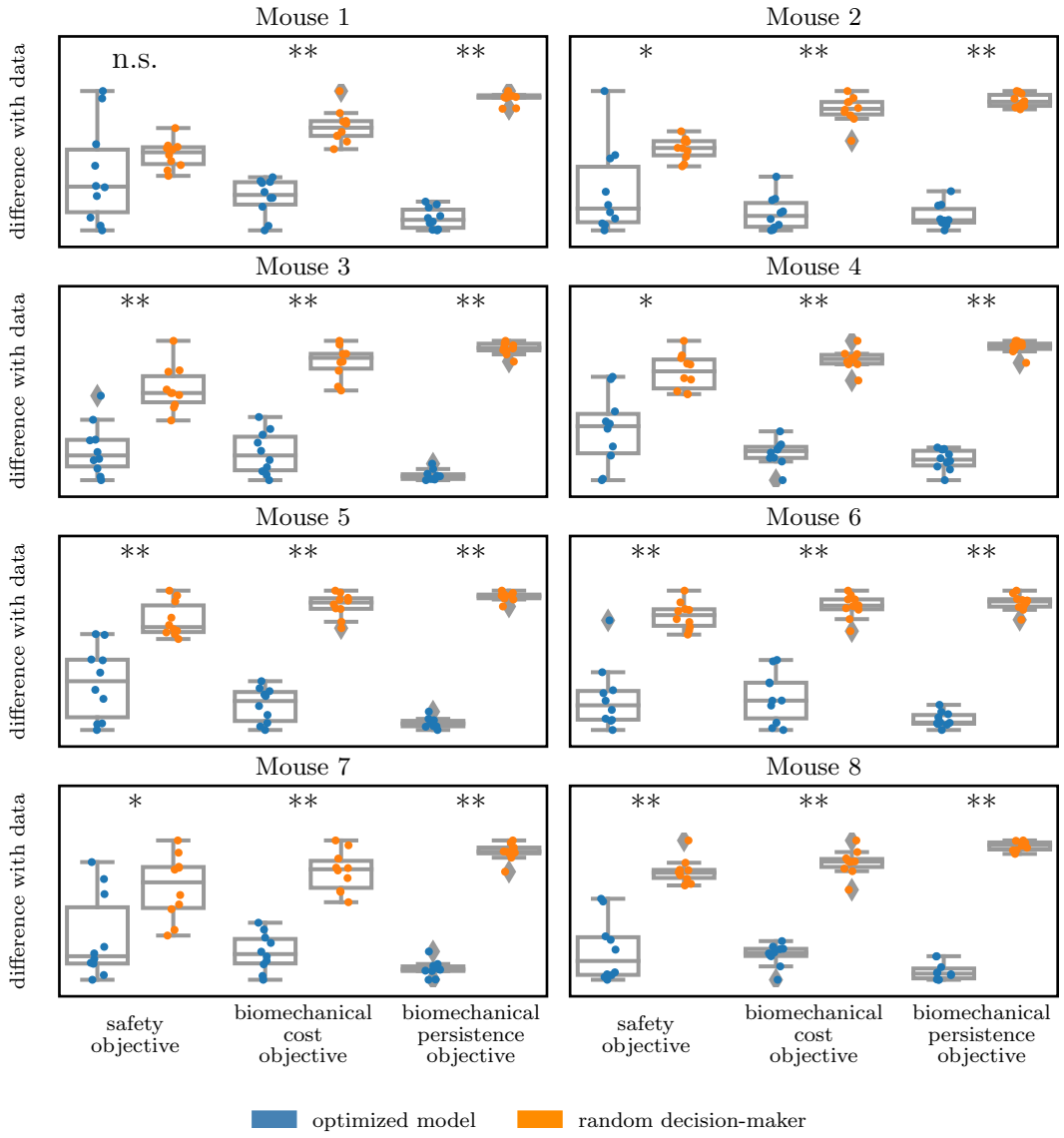


FIGURE 3.15: U-maze comparative statistical analysis on the safety, biomechanical cost, and biomechanical persistence objectives for the selected optimized models and the corresponding random exploration. Each sub-figure represents the results for each mouse agent in terms of behavioral difference with the data. ** indicates that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise.

decision-making agent for 23/24 total objectives. Even though, the general trend to have a hierarchy among occupation of corners, walls and central areas holds for this u-maze dataset, Mouse_1 shows this same hierarchy, but weaker than the other mice in the dataset. Also, the safety objective is the the only objective for which the difference with random exploration is sometimes not strongly significant (Mouse_2, Mouse_4, and Mouse_7, which have just one *). Looking at these results, a possible improvement for the model could be to optimized the preference for each type of tiles (corners, walls, and central areas), without imposing the hierarchy. This would make all the rodents' exploration models optimize their tiles' preferences, even if

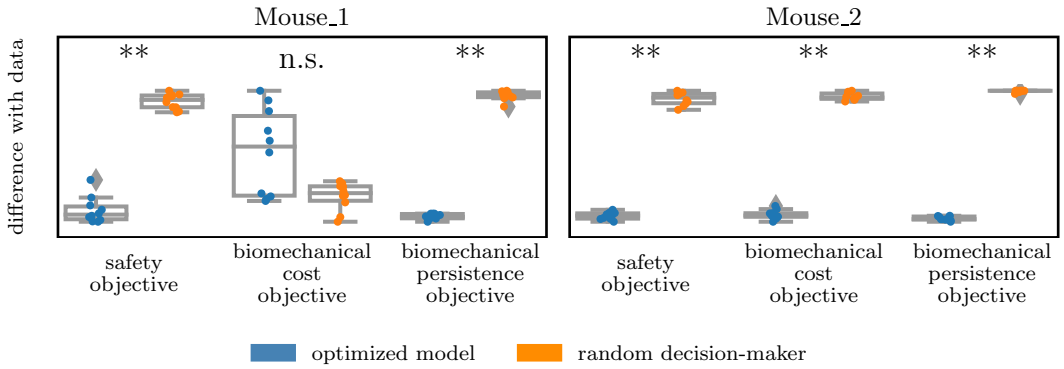


FIGURE 3.16: Square open-maze comparative statistical analysis on the safety, biomechanical cost, and biomechanical persistence objectives for the selected optimized models and the corresponding random exploration. Each sub-figure represents the results for each mouse agent in terms of behavioral difference with the data. ** indicates that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise.

they are different from the general trend of the dataset.

Concerning the optimization results in the square open-maze (Fig. 3.16) the optimal model is also able to strongly significantly better represent the behavioral trends of the data, in general (in 5/6 objectives). In this case, the biomechanical cost objective has not properly been optimized for Mouse_1, probably due to the more uniform nature of its exploration (Fig. 3.4, Mouse_1), to the limited actions range derived from the identification of the most suitable next actions for this dataset, and to the open structure of the maze (Fig. 3.9, square open maze). These factors make this component less predominant when these rodents' behavior is simulated; thus, it is more difficult for the evolutionary algorithm to optimize it. These results imply that the contribution of the biomechanical cost component in the free exploration model should be reconsidered. It should be adapted to the definition of the possible next states of the MDP of a particular maze. In this case an adaptation of the number of bins to describe the relative rotations of the animal should be constrained to be equal or smaller to the number of the possible next states, as for the u-maze case, and targeting the same orientations that the model allows for exploration.

Finally, Fig. 3.17 demonstrates that the only behavioural component which is strongly significantly present in this large rat dataset is the biomechanical persistence one. As we will discuss in Sect. 3.1.4, in the context of our proposed MDP's definition, random decision-making can surprisingly generate a tile occupation that is statistically close to the one we have from the data. For this reason, it is foreseeable that the optimized exploration model's difference with the data in the safety objective is sometimes non-significant different from the random exploration one (this happens in 6/21 cases). In two cases, the random exploration's safety objective is also significantly closer to the data than the optimized model (for rat Rat430 and R432).

Concerning the biomechanical cost objective, the situation is similar but, this time, there are 14/21 cases where the distance with the data of this objective for the optimized model is non-significantly different from the one obtained with random exploration. Also here, there are two cases, for rats Rat428 and R439, where random

exploration is capturing significantly better the biomechanical cost distribution of the data than the optimized free exploration model. As for the square open maze, here the range of motion is more constrained than the case of the u-maze (Fig. 3.9, grid-maze) thus, these results suggest that also for this dataset, it would be interesting to analyze the data and adapt the number of bins in the relative rotations histograms to the same rotations allowed in the conversion of this maze and dataset to the MDP.

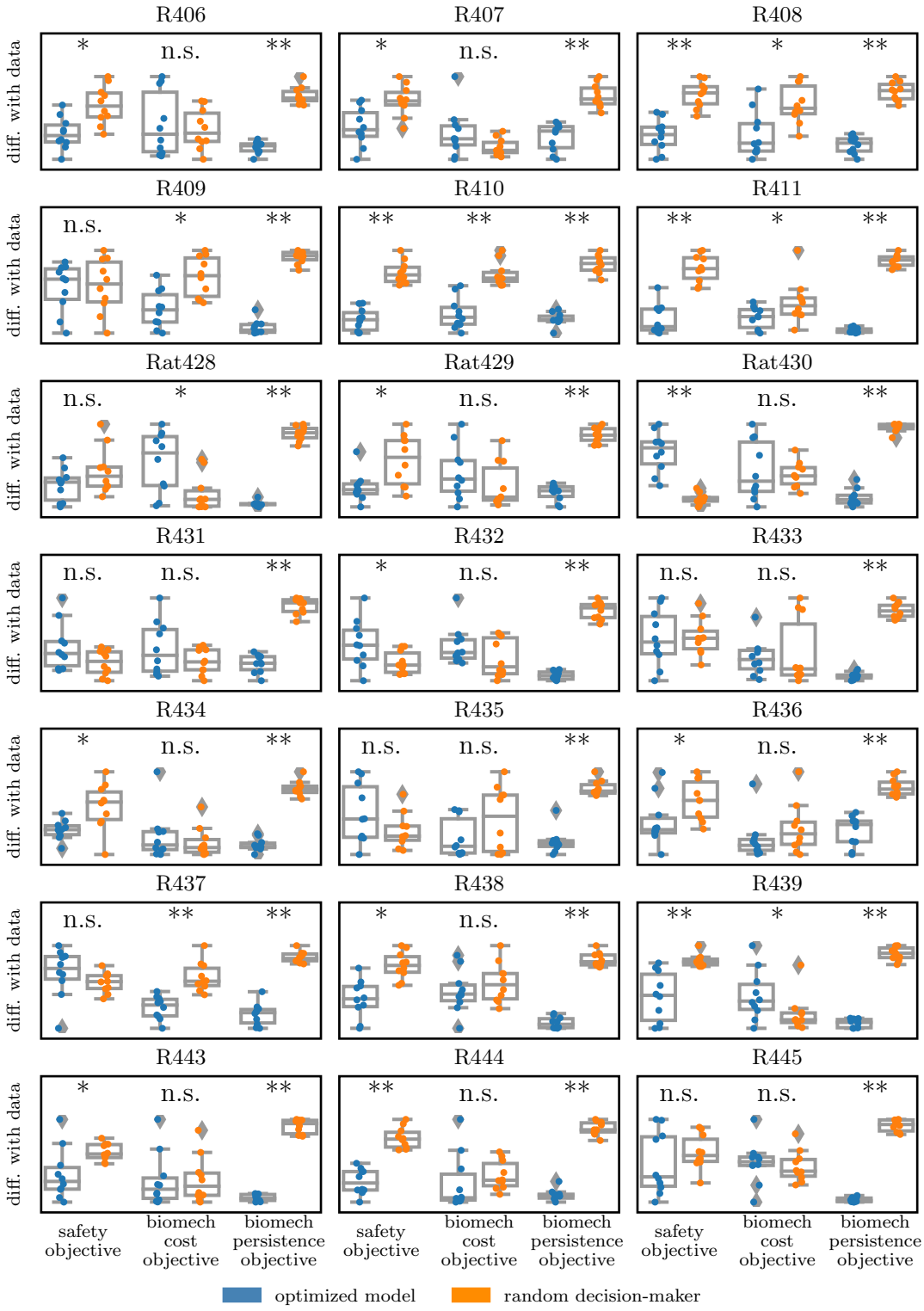


FIGURE 3.17: Grid-maze comparative statistical analysis on the safety, biomechanical cost, and biomechanical persistence objectives for the selected optimized models and the corresponding random exploration. Each sub-figure represents the results for each rat in terms of behavioral difference with the data. ** indicates that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise.

3.1.4 Discussion

This chapter proposes a new design for a generalized value-based decision-making model for rodent free exploration. This computational model is data-driven, meaning that its concept has been inspired not only by the literature (Sect. 2.1.1) but also by new and original rodent behavioral data. The results presented are gathered only on the final iteration of the design, which was, in the beginning, based uniquely on the u-maze dataset since it is the very first data we had. Previous iterations included other behavioral components, whose description is not reported here for brevity.

In Tab. 3.2, we report a statistical analysis done on the distributions of the behavioral measurements of the data (u-maze and grid-maze) and ten repetitions of the simulated random decision-making agents in the same framework and the same conditions.

		u-maze	grid-maze
safety	decision-points		0.095
	corners	0.00016	0.0016
	walls	0.28	0.036
	centres	0.00016	0.51
biomechanical cost	1	0.00041	1.6e-05
	2	0.00093	9.4e-06
	3	0.37	7e-06
	4	0.0054	3.8e-05
	5	0.43	7.1e-06
	6	0.00093	1e-05
	7	0.00041	1.6e-05
	8	0.00091	6.7e-06
biomechanical persistence	bouts mov	0.00016	1.2e-05
	bouts stop	0.00093	6.5e-05

TABLE 3.2: Comparative statistical analysis between the behavioral features of the data and the simulated random decision-making agent in the same conditions and MDP framework. For each dataset (u-maze and grid-maze) and each distribution corresponding to the bins of the measurement of the three behavioral components (Fig. 3.10, 3.11, and 3.12 respectively), a Wilcoxon-Mann-Whitney comparison test is performed. Here, we report the p-values for each comparison, and the blue gradient decreasingly shows non-significant statistical difference (dark blue) and statistical difference; p-values < 0.05 (medium blue) and p-values < 0.001 (light blue).

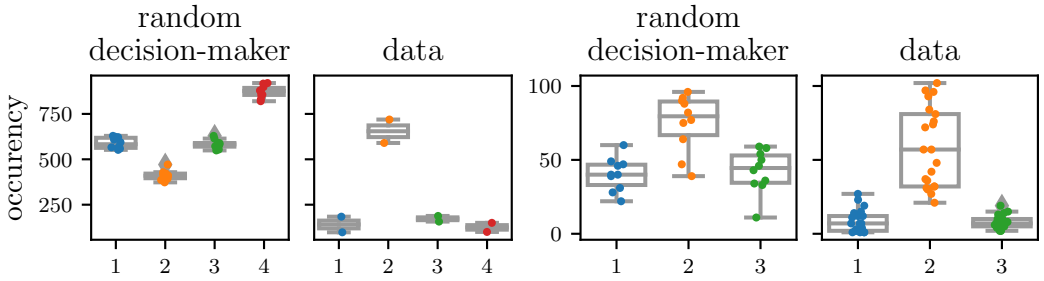
These analyses are separately performed on each dataset and are related to the same results presented in Fig. 3.10 - 3.11 - 3.12 (u-maze and grid-maze). We do not perform the statistical tests for the case of the square open maze since the data has a distribution of just two samples (Mouse_1 and Mouse_2, in Fig. 3.10 - 3.11 - 3.12, square open maze). Almost all the data distributions are statistically different from the ones derived from the random exploration, highlighting that the free exploratory behavior of these animals shows characteristics that are far from haphazard navigation. Even though the datasets represent the behavior of different rodents in different mazes and for a different time duration, the relevance of these behavioral patterns, in contrast to an indiscriminate exploration, interestingly persists. In more detail, the biomechanical persistence measure results indeed as the most relevant behavioral

feature across the two analyzed datasets because it is always significantly different from the random decision-makers distributions (Fig. 3.12, Tab. 3.2).

Concerning the safety measure, it is interesting that, even if the strong definition of a hierarchy between corners, walls, and central areas exists in the u-maze, the occupation of the spots next to the wall is not different from the one captured by random exploration. This suggests that the maze's morphology can already define the data's walls' occupation during habituation (Fig. 3.10, Tab. 3.2). Concerning the grid-maze, despite the large amount of available data (21 individuals) compared to the other two datasets (8 individuals for the u-maze and 2 individuals for the square open-maze), the occupation of the four types of tiles does not differ from the random exploratory case both in the decision points' and the central areas' occupancy. Nevertheless, we think it is interesting to look at the occupation of the decision points absent from the other two datasets. It has been observed that rats present crucial cognitive activity in corridors' intersections (Johnson and Redish, 2007). In particular, neural ensembles in CA3 have been recorded to represent locations swept forward to cover the animal's possible future paths. This active inferential process could imply more time spent in that area for the animal, particularly when the environment is novel. In fact, based on the actual number of junctions in the grid-maze, decision points represent the second most relatively occupied spots, second just to the corner spots that strikingly seem to represent safer locations for the animal's first navigation in a novel environment, even though the whole environment is composed by narrow corridors. So, we decided to keep the decision points tile type as a key element of the maze descriptor even if the distribution of its occupancy is not significantly different from the one in random exploration.

Finally, the biomechanical cost measure is statistically different from the one of a random decision-maker for most of the bins and for all the two datasets. Here, the distribution of some bins appears to be statistically comparable to the one from random exploration (for bins 3 and 5 in the u-maze). However, let us look at the shapes of the ensemble of the distributions representing the relative orientations. It is easy to see that the shape of the orientations for that data is not respected in random exploration for no dataset (Fig. 3.11). Therefore, even though we report this analysis for the sake of clarity, in this particular case, it is probably better to directly look at the shape of the distribution of the relative orientations (as done for the model optimization process and the other analyses in Sect. 3.1.3, in particular, Eq. 3.19) instead of performing a bin-to-bin comparison.

Based on the above data analysis, we have proposed a new data-driven model able to generate a decision-making process resulting in a free exploratory behavior significantly closer to one of the real rodents than the one generated from consistent random exploration. All the rodents got at least one (over three) behavioral component significantly closer to the data than random exploration. Furthermore, 24/31 individuals got two behavioral components that are significantly or strongly significantly better captured by the optimized model than by random exploration. Strikingly, for all the individuals, the biomechanical persistence component is strongly significantly better captured by the optimized free exploration model than by random exploration. This suggests that, at the temporal and spatial scale of the model we propose, the alternation between longer static bouts and shorter moving bouts (Fig. A.6 - A.7 - A.8) is a behavioral characteristic that is shared among the free exploratory behavior of all rodents we have analyzed. To do so, the general framework we propose (Sect. 3.1.2) is thought to be valid for all rodents, with the possibility to describe the behavioral shades of each animal by optimizing a set the 9 (10 in case of the grid-maze) model parameters (Sect. 3.1.3). Our results show that the proposed



(A) Square open-maze (B) Grid-maze

FIGURE 3.18: Relative directional preferences of the animals and of a simulated random decision-maker if the directional bins are designed on top of the possible discrete actions available in the maze MDP (Fig. 3.9). Blue bins indicate turning $\pi/2$ radians left, yellow ones going straight, green ones going $\pi/2$ radians right, and red ones turning π radians.

model can capture the identified behavioral exploratory features of each rodent significantly better than random exploration, with some exceptions for which we have discussed the possible motivations and suggested possible model modifications to deal with them (Sect. 3.1.3). A similar approach to comparing the optimized model to a random exploration model has also been adopted by Rosenberg et al. (2021) to study spatial navigation and learning in mice.

As we can see from Fig. 3.15, Fig. 3.16, and Fig. 3.17, some behavioral components are easily captured by the optimized model. In general, the biomechanical cost component is the one that apprehends the worst of the animals' behavior. This is due to the model's constraints, particularly the fact that the definition of the bins for the relative rotations suits the original u-maze data, which first inspired the design of the free exploration model. However, it does not consider the difference between its possible next actions and the ones of the other datasets. That is why we propose a future adaptation of the bins definition based on the possible next actions of the dataset and MDP. This modification can be easily integrated into the model since it does not need new parameters to be optimized for the biomechanical cost component, and the biomechanical cost objective can remain unchanged. Looking at Fig. 3.18, we can already appreciate that the histogram distribution of the relative directions of both the square open-maze and the grid-maze have a more significant distribution. In these cases, the relative rotations' distributions are closer to a Von-Mises (the distribution we used to model the biomechanical cost component in our model) than if we keep the eight-bins-histogram like in the u-maze case (Fig. 3.11).

Fig. 3.18a shows that the data distribution is almost opposite to the one generated by random exploration if we create relative rotations bins centered in the next possible states (for the two datasets where the biomechanical cost objective was more challenging to be optimized, Fig. 3.16 - 3.17). This was not the case with the eight bins' relative rotations distributions in Fig. 3.11 (square open-maze). Fig. 3.18b shows that, whether we consider just the actual points where either the animal or the simulated agent has a choice to turn, meaning the decision points of the maze, we can notice a difference between the two histogram distributions. In particular, rats prefer to go straight more than the random exploratory agent. This underlines that, even in this very peculiar maze morphology, at this temporal and spatial scale, a biomechanical cost behavioral component that privileges the maintenance of the same direction of motion exists, as for the other examined mazes. This final data

analysis suggests that keeping the histogram distribution of the same size as the possible actions in the MDP could improve the model's power to describe rodent behavior, particularly its biomechanical cost component.

In conclusion, the proposed value-based model could be generalized for different situations (namely free exploration and spatial learning) by adding or removing behavioral components. For each type of exploration (free or goal-directed), the number of components should be kept to the lowest number that can reproduce the significant behavioral patterns of the available data (Sect. 2.2, A. Collins and Khamassi (2021)).

In the next section, the free exploration model will be generalized and evaluated in data-driven cases of spatial learning.

3.2 A reinforcement learning-based model on the role of hippocampal replay in spatial positive and negative learning

Free exploratory behavior is not the only condition characterizing navigation. When exogenous conditions or stimuli that have an emotional valence for the animal become part of the scenario, its behavior changes. We can refer to this exploratory situation as *conditioned exploration*.

In rodent behavior, the two most studied conditioning stimuli concern sustenance and survival and can be of opposite emotional strength (among them Morris (1981) and Girardeau, Inema, and Buzsáki (2017)). Both result in an induced goal-directed exploration when the aim can be either to approach a rewarding stimulus (food, odor for example) or to avoid it (aversive odor, sound, shock, air puff).

Even though aversive stimuli can be modeled as emotional conditioning as per the positive case, but with an opposite sign, two different neural circuits are involved in processing these two opposite valence conditions (Yacubian et al., 2006). Nevertheless, the amygdala and the orbitofrontal cortex (OFC) play an important role in encoding and updating information concerning reward and punishment (Baxter and E. A. Murray, 2002; Pickens et al., 2003; Wrase et al., 2007). Concerning negative conditioning in rats, it has been observed that periaqueductal gray (PAG) is relevant for fear conditioning (Canteras and Goto, 1999) and its lesions suppress freezing behavior in response to negative stimuli. In contrast, the conditioned suppression behavior is preserved (Amorapanth, Nader, and LeDoux, 1999). More recent observations have also pointed out the involvement of PAG in fear memory (Watson et al., 2016) and classified the dorsal PAG (dPAG) as the center which coordinates survival response and the ventral PAG (vPAG) as the part in charge of encoding innate and learned fear behaviors. Palminteri et al. (2015) observed human fMRI data during a task where the subjects were asked to maximize the reward and minimize the penalties. They observed that, after value contextualization, the ventral striatum (VS), devoted to encoding the reward signal, responded to successful penalty avoidance, suppressing the activity of the anterior insula, which usually encodes negative conditioning. Talking about reward-based navigation, it has been commonly recognized that in humans VS received massive projections from dopaminergic midbrain neurons and encode important information for reward-based learning (Delgado, 2007; Daniel and Pollmann, 2014). Moreover, it has been suggested that the VS is also involved in the neural mechanisms at the base of the computation of the reward prediction error (Khamassi and Humphries, 2012). Medial prefrontal cortex and dorsal hippocampus activity have been also found to be correlated with performances in

trained mice after they have learned a reward-based conditioning task (Le Merre et al., 2018).

As introduced in Sect. 2.1.4, instrumental conditioning gathers all the situations where reinforcement or punishment are used to either increase or decrease the probability that a behavior will occur again in the future.

Many research studies have been conducted for studying instrumental conditioning in rodents during navigation tasks and have disclosed the importance of hippocampal reactivations in these contexts and also in goal-directed navigation (Ólafsdóttir, Bush, and Barry, 2018; Cazé et al., 2018). Furthermore, hippocampal replay's crucial role in memory consolidation makes that their disruption causes relevant degradation of goal-oriented spatial learning (Girardeau et al., 2009). Thus, hippocampal reactivation should play an important role in instrumental spatial learning, whether the conditioning is positive or negative. However, for obvious ethical reasons, spatial learning associated with aversive stimulation has been studied to a smaller extent than positive spatial learning. This justified under-representation of these types of studies strongly motivates the research towards designing computational models at different levels (such as neuronal, network level, and behavioral) for making it possible to predict the observation from these experiments with the least possible data.

A recent work by Wu et al. (2017) showed awake hippocampal reactivations representing the animal's path to a punishment location after it has experienced a foot shock in a particular area of the maze, although the animal does not re-enter the punishment area again. In positive conditioning, the animal spends more time in the rewarding areas. Thus, it is more complex to assess if hippocampal reactivations of the reward areas are more related to the emotional relevance or to the recency of the experience. On the contrary, if a preference in reactivating place cells concerning punishment areas exists, this suggests and confirms the hypothesis that hippocampal reactivations can replay emotionally relevant information and not just indiscriminate recent spatial experience. To support this idea, Girardeau, Inema, and Buzsáki (2017) recorded hippocampus and amygdala coordinated reactivations during non-REM sleep after aversive stimulation by an air puff. Also, in this case, the co-activation of the hippocampus and amygdala was prevalent in the trail corridor leading to the aversive conditioning.

Nevertheless, to the extent of our knowledge and excluding the recent work by Bryzgalov (2021), comparison studies investigating possible differences in the initiation and relevance of hippocampal reactivation during positive or negative conditioning have not been presented yet. Additionally, to our knowledge, a computational model accounting for these differences is still lacking. In this chapter, by extending the modeling capabilities of the already presented model (Sect. 3.1) to goal-directed exploration, and by optimizing the model on new data, we aim to predict if such differences exist, by simulating the contribution of simple replay mechanisms in our model.

Sect. 3.2.1 describes the conditioned experimental protocol and analyses the data through the results obtained in Sect. 3.1.1 and a new proposed behavioral metric for emotional conditioning. Then, in Sect. 3.2.2, we describe how we simulate the different phases of the experiment and, in particular, how we integrate the conditioning value to the previous proposed free exploration model Sect. 3.1.2,. In this section, we explain also how we adopt Reinforcement Learning (RL) to simulate the spatial assignment of emotional values to the different areas of the maze. Moreover, we explain how we predict the preference for the presence of replay-inspired RL mechanisms to observe a simulated behavior comparable to the ones measured on the

animals' data by Karim Benchenane and Dmitri Bryzgalov. We then explain which parameters of the model we are interested in optimizing and how, and also what sort of insights concerning the role of hippocampal replay in aversive and positive spatial learning we can deduct from our results (Sect. 3.2.3). Finally, we discuss our results based on the current literature on opposite valence learning in rodent spatial navigation and suggest what experimentalists and computational neuroscientists can do to test our model predictions in terms of future studies.

3.2.1 Behavioral data

The dataset we are using for this thesis chapter is also available to us thanks to our collaboration inside the RHiPAR project funded by the CNRS 80'PRIME Program (Sect. 1.3). As for the free exploration data in the u-maze (Sect. 3.1.1), these data were recorded during the experiments dedicated in studying the role of hippocampal reactivation in aversive and rewarding experience (Bryzgalov, 2021).

The data we will employ in this chapter concerns six new C57BL6jR mice, navigating in the same u-maze of 1m x 1m size, already described in Fig. 3.1. The technique for recording the animals' position is the same one that has been described in Sect. 3.1.1.

In this conditioned exploration protocol, the data are organized in different sessions. Each mouse experiences:

- **Pre-conditioning sessions**: the animal is free to explore the u-maze as in a habituation phase;
- **Conditioning sessions**: while navigating in the maze, the animal is subjected to an exogenous stimulation, always when entering or remaining in a specific part of the maze (every 6 seconds);
- **Post-conditioning sessions**: the animal re-enters the u-maze after the conditioning phases and it is now free to explore the environment without external stimuli or constraints.

The mice experienced all these phases either with a positive or a negative intracranial electrical stimulation. A PulsePal stimulator has performed both stimulations (Sanworks, NY, USA). Each stimulation was a train of 13 biphasic 1 ms short pulses with an interstimulus interval of 8 ms (125 Hz) (Bryzgalov, 2021). For aversive and rewarding stimulation, the dorsolateral periaqueductal gray matter and the medial forebrain bundle were targeted respectively. Fig. 3.19 shows the experimental protocol of the data that are available to us.

Each experimental phase has different sessions with a total duration of 16 minutes for the pre-conditioning sessions and the post-conditioning ones and 32 minutes for the conditioning sessions. Between the conditioning phase and the post-conditioning one, the animals slept for 2 hours.

Even though the pre-conditioning phase is divided into shorter navigation sessions, the data reflects the same relevant behavioural components, identified in Sect. 3.1 (Fig. 3.20).

Fig. 3.20 suggests that the pre-conditioning phase can be modeled as a habituation phase, using our proposed free exploration model (Sect. 3.1.2). First of all, the data were sampled at 600 ms as in the previous case (Sect. 3.1) and the u-maze discretized in the same way (Fig. 3.9, u-maze), since it is the same maze that has been used for recording the previous data (Sect. 3.1.1). As expected, by applying the same

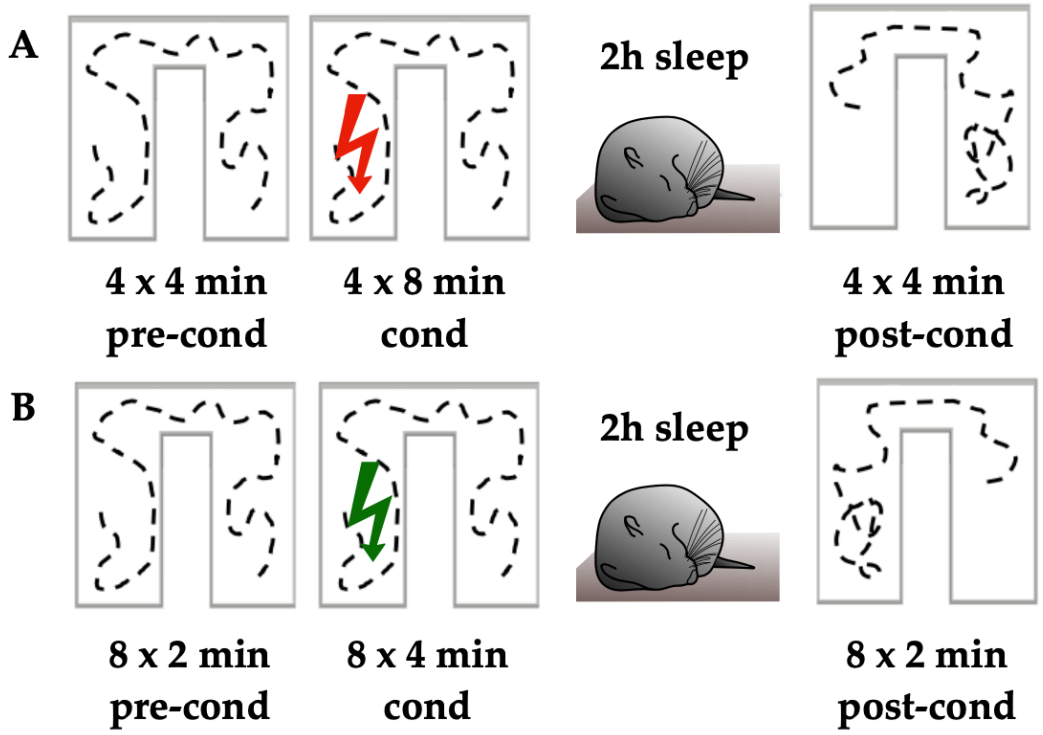


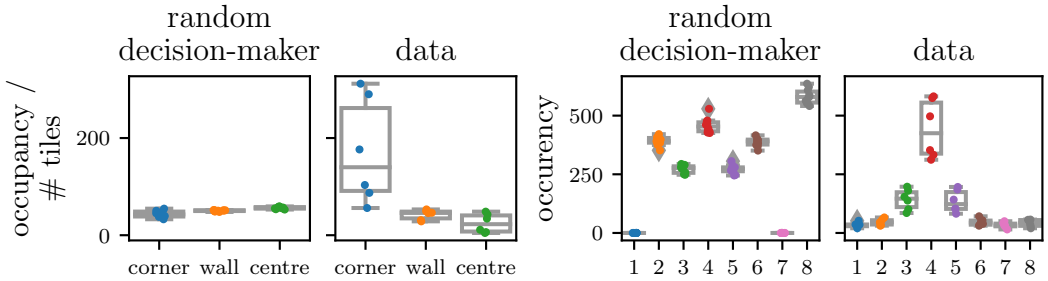
FIGURE 3.19: Experimental protocol. A) Aversive experimental protocol: pre-conditioning, conditioning and post-conditioning phases. The red arrow indicates the area where the aversive stimulation has been delivered. B) Positive experimental protocol: pre-conditioning, conditioning and post-conditioning phases. The green arrow indicates the area where the positive stimulation has been delivered. Figure adapted from Bryzgalov (2021) and Girard (2021).

clustering algorithm on the data from the pre-conditioning phases, the relevant actions for the definition of the Markov Decision Process (MDP) are the same ones alrady identified for the previous u-maze date (gray crosses and triangle in Fig. 3.9, u-maze).

Concerning the post-conditioning exploration, we can observe a very biased occupation of the u-maze compared to the pre-conditioning phase due to the emotional stimulation. As it has been done in Bryzgalov (2021) , we divide the u-maze into seven different sub-areas, as represented in Fig. 3.21a. We analysed the occupacy of these seven sub-areas in the pre-conditioning and the post-conditioning sessions, in case of aversive and positive stimulation (Fig. 3.21b and Fig. 3.21c respectively).

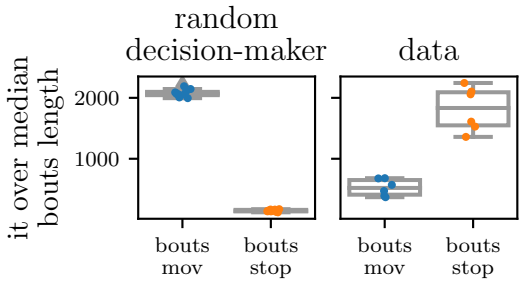
These histogram distributions represent how much the animals approach or avoid (respectively for the positive and negative conditioning) the areas where the stimulation was experienced. We decided to keep the division in these seven areas, even though the occupation of some of them is not significantly different from the one in free exploration, because they equally divided the maze area in sub-areas large as the stimulation area. Thus, the conditioning metric can more homogeneously scale the occupancy of the maze in a sort of linear distance to the location of the stimulus.

Interestingly, we can notice that the pre-conditioning occupations of the maze in the seven sub-areas are very similar between aversive and positive stimulation experiments (Fig. 3.21). The preference for the first, fifth and seventh sub-areas is stronger than the occupation of the remaining sub-area with corners, the third one.



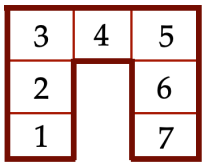
(A) Safety metric

(B) Biomechanical cost metric

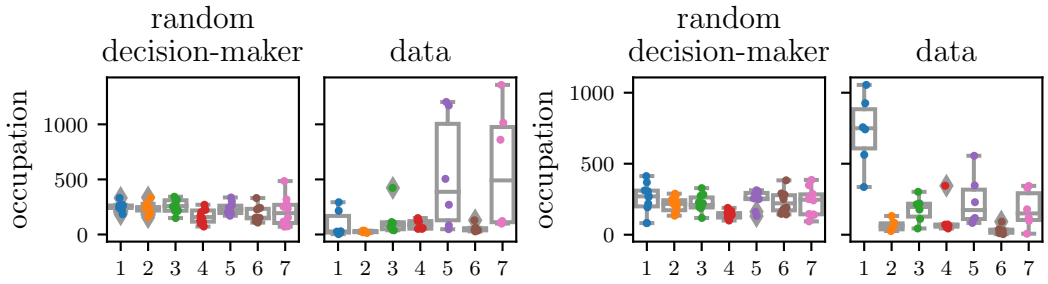


(C) Biomechanical persistence metric

FIGURE 3.20: Behavioral measurements on random exploration and on the pre-conditioning data in the u-maze. B) Bins 1 to 8 correspond to {[-2.75; -1.96], [-1.96; -1.18], [-1.18; 0.39], [-0.39; 0.39], [0.39; 1.18], [1.18; 1.96], [1.96; 2.75], [2.75; 3.53]} radians.



(A) 7 sub-areas



(B) Aversive stimulation

(C) Positive stimulation

FIGURE 3.21: Occupation of the 7 subareas in the case in the pre-conditionig and post-conditioning phases.

The evidence that the sub-areas with corners are the most occupied by the animal is an observation we expect since corners are usually the most occupied spots in the maze (Fig. 3.20a). This suggests they could be perceived as the safest places in the maze or the ones that can provide the largest amount of information, without exogenous stimulations (Berlyne, 1950; Golani, Benjamini, and Eilam, 1993). Concerning the post-conditioning sessions, as expected, there is an increasing preference for the two corner areas (fifth and seventh ones) on the opposite corridor of the aversive stimulus, and a very strong preference for the first sub-areas when the positive stimulation is instead given to the rats. These observations suggest that the positive stimulation has been experienced many times. Its location has been well consolidated in the animals' spatial memory (Fig. 3.21c, right), as well as the negative stimulation experience, since the most visited sub-areas lies on the opposite corridor of the maze (Fig. 3.21b, right).

3.2.2 Exploration model

The free exploratory behavior we observed in the dataset described in the previous chapter (Sect. 3.1.1), can be seen as a specific case of a more general exploration model. Indeed, the animals did not experience any externally induced emotional reaction in that exploratory phase, so we can talk about *free exploratory behavior*. Nevertheless, external factors, such as brain stimulations triggering an emotional response, can deeply affect this free exploration.

Here, we propose that a general exploration model during rodent navigation is one that adds the emotional value to the list of values that we proposed in the free exploration model (*i.e.*, the values of the biomechanical cost component, safety component, and biomechanical persistence component). This section will introduce the most general model we propose for rodent exploration. As in the previous chapter, all the parameters written in red are the parameters that are then optimized to describe the conditioning behavior of the animals.

As already said in the previous section, the modeled MDP used here will be the same as for the u-maze dataset of the previous chapter, due to the consistency of the measurements we took on the two datasets. Thus, each one of the possible next state s' (gray triangle and crosses in Fig. 3.9, u-maze) has a value $V_{exploration}$ at time t which consists of the value that is assigned by the three free exploration components $V_{free\_exploration}$ plus the contribution of the conditioning component $V_{conditioning}$, weighed by W_c with respect to the other free exploration component (Eq. 3.2.2).

$$V_{exploration}(s'_t) = W_c V_{conditioning}(s'_t) + V_{free\_exploration}(s'_t)$$

$$\text{with} \quad W_c \in (0, 10] \tag{3.22}$$

The conditioning component $V_{conditioning}$ is weighed by the parameter W_c which modulates the relevance of the learned conditioned values with respect to the other behavioral components (safety, biomechanical cost and biomechanical persistence, in Sect 3.1.2), that already have parameters to modulate their relative contributions in the behavior, $p1$, κ, and bp, respectively). Thus, the evolutionary range of W_c is also from around 0 to 10 to be consistent with the range of the weight parameters of the other components.

The conditioning component is learned through multiple conditioning sessions, the same the real mice had, where the agent learns the conditioning values by following a reinforcement learning (RL) rule. Since it has been commonly accepted that a good framework for modeling instrumental learning is RL (Glimcher, 2011; A. Collins and Khamassi, 2021), we propose to model the spatial learning process, associated with the emotional stimulation, by using RL. The values learned by this new RL component are then assigned as conditioning values to the maze's next states. Fig. 3.22 gives an overall view on how the conditioning experiments from which we have our data, are modelled.

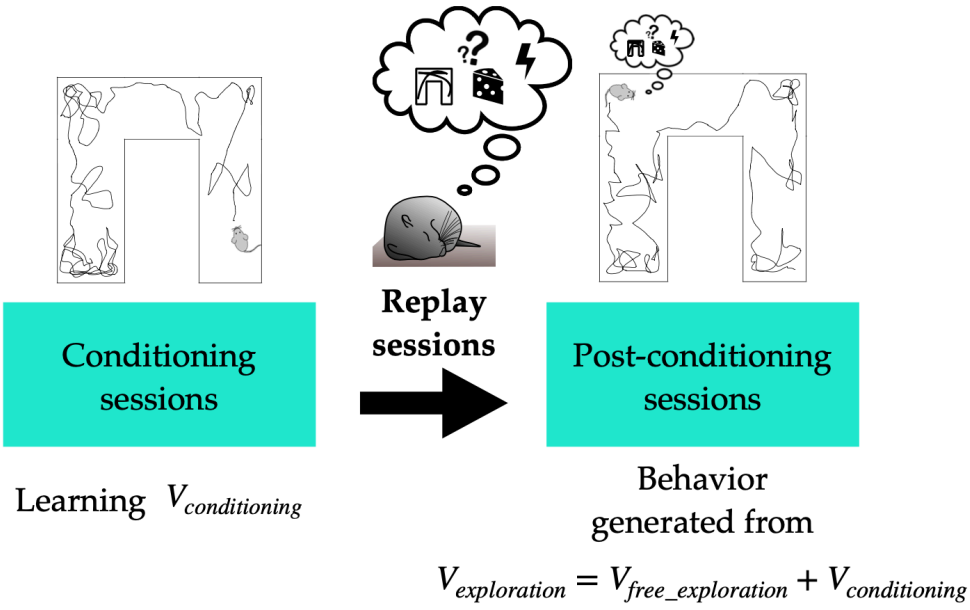


FIGURE 3.22: Scheme of the modeling paradigm for the conditioning experiments. On the left, the simulated agent replicates the same discretized trajectories that the corresponding mouse did during the conditioning session and, simultaneously, the conditioning values for the maze's state $V_{conditioning}$ are learned. After conditioning, the sleep hippocampal reactivations are simulated by the agent updating the $V_{conditioning}$ by performing unordered off-line replay. Finally the post-conditioning sessions are simulated by having the agents making its own decision, based on $V_{exploration}$. Figure adapted from Girard (2021).

Our proposal for learning the conditioned components is to have two different formulations based on the emotional valence of the stimulation, to reflect the existing neural circuitry duality between aversive and positive instrumental learning.

By considering the maze's states as a simplified representation of how the place cells encode the different locations of the maze in the rodent's hippocampus, the proposed model predicts the values that the animals would estimate at each of these locations. Thus, this estimation is based on the replication of the real experience of the stimulation and subsequent behavior that the animals had during the conditioning phases. Indeed, on the left part of Fig. 3.22, we show that the modeled mouse follows the same trajectories covered by its real counterpart through the conditioning sessions and in the meantime learns the $V_{positive\_conditioning}$ for the visited states.

As anticipated above, the model has two different q-learning formulations depending on the valence (sign) of the conditioning. If the stimulation is positive, we apply the following RL rule to the state s, to learn its $V_{positive\_conditioning}(s)$:

$$V_{positive\_conditioning}(s_t)_{t+1} = V_{positive\_conditioning}(s_t)_t + \alpha[R(s_{t+1}) + \gamma \max_{s'_t} V_{positive\_conditioning}(s'_t)_t - V_{positive\_conditioning}(s_t)_t]$$

$$\text{with} \quad \alpha \in [0,1], \quad \gamma \in [0,1] \tag{3.23}$$

Here, the value of the current state is updated based on its relative likelihood to get the agent to a positive stimulation R in the next future (discounted by a factor γ, between 0 and 1), compared to the other possible next states. Besides, α is also a parameter, evolving between 0 and 1, which modulates the learning speed of the algorithm, as in the standard q-learning formalization (Eq. 2.4 and Sect. 2.2.1).

Q-learning is one of the most used off-policy learning rules in RL, due to its efficiency in finding near-optimal policy with very little information on the reward and on the experimental environment. Compared to the most classic formulation of q-learning, where the updates of the *q-values* are performed on state-action couples, here the learning rule performs the updates just on the states. This formulation is more relevant in our proposed value-based decision-making model, since we assign values to all the possible next states also for the other components (safety, biomechanical cost, and biomechanical persistence). The q-learning rule we have applied in our model does not need to consider a set of possible next actions and learn their relationship with the state of the maze. This choice is also a consequence of our MDP modelization in which we have both continuous and discrete states. From a discrete state and a relative action, it is impossible to determine what would be the next state, as this would require adding an orientation to the discrete formulation. Conversely, it would be possible, albeit complex, to perform the classic q-learning on the continuous states and relative actions, but this would not align with the rest of the discrete state evaluation. If the conditioning has an opposite valence, meaning it is aversive, it computationally represents a reward of -1, instead of 1, and it is called punishment P. With an aversive stimulation, animals need to know which areas to avoid and, in particular, propagate this negative signed knowledge. To model the negative conditioning learning phase, we use the same q-learning rules, explained just above, with a negative conditioning P, instead of the positve reward R. We average the update of the chosen next state s_{t+1} on the minimum, instead of the maximum, value among all the other possible next states s'_t (Eq. 3.24).

$$V_{negative\_conditioning}(s_t)_{t+1} = V_{negative\_conditioning}(s_t)_t + \alpha[P(s_{t+1}) + \gamma \min_{s'_t} V_{negative\_conditioning}(s'_t)_t - V_{negative\_conditioning}(s_t)_t]$$

$$\text{with} \quad \alpha \in [0,1], \quad \gamma \in [0,1] \tag{3.24}$$

In this case, the learning parameters to optimize (α and γ) will be the same as in the positive case (Eq. 3.23), and with the same evolutionary ranges. By using the minimum instead of the maximum, the learning process would be the same as the

one in the positive conditioning case, just with a propagation of negative state values through the maze.

Laventure and Benchenane (2020) reviewed the theoretical bases of hippocampal reactivations during sleep and reported many experiments that have been conducted on the topic suggesting that sleep sharp waves and ripples (SWRs) not only improve the animal localization, by replaying pure spatial content, but they also represent place–reward (or place–punishment) associations (Atherton, Dupret, and Mellor, 2015). Thus, we propose to model the conditioning experimental protocol with an unordered model-free replay phase (Sect. 2.2.4) in between the conditioning and post-conditioning phases, to be consistent with the research for sleep replay conducted in the work by Bryzgalov (2021) (Fig. 3.19). After all the conditioning sessions have been simulated, the agent performs a number $\#rs$, which will be optimized in the range $[0, 10]$, of replay sessions. This means that the exploration model has also the possibility of not performing offline replay at all, if a good combination of the learning parameters with $\#rs = 0$ is found and it is able to replicate the post-conditioning behavior of the mouse. To be consistent with the evolutionary range of the other parameters, the maximum number of times that all the conditioning sessions could be shuffled and replayed is 10. We consider 10 sessions a sufficiently large replay budget to consistently boost the learned states' values propagation. Each replay session consists of shuffling the entire memory buffer containing all the transitions (s, a, s', r) done by the agent during the conditioning sessions, and replay them one by one. In this way, the values of all the states s, contained in the transitions (s, a, s') will be updated and the positive or negative states-values in the maze would be modified as if the agent had virtually explored the maze again. On the one hand, we decided to model the sleep offline reactivation as unordered replay of the past experience for simplicity, because we do not have more detailed information about the type of hippocampal reactivations when we designed the model. Thus, we do not want to make constraining assumptions on the type of replay, accordingly avoiding the need to use more complex model (such as model-based, prioritized sweeping, or trajectory sampling replay strategies) to describe them. On the other hand, from a RL-computational point of view, Cazé et al. (2018) suggest that the noisy dynamics of asleep hippocampal reactivations can be appropriately modeled by unordered model-free replay.

The replayed transitions (s, a, s', r) do not directly represent an equivalent of place cells in the hippocampus, but a compact and discretized version of parts of the animals' previous spatial experience, in the timescale of a decision-making step in the exploration model. They are coherent with the classical format used to represent spatial experience and its associated values in MDP and RL.

The last part of the whole conditioning experimental protocols concerns the post-conditioning sessions (Fig. 3.22). In these post-conditioning sessions, the mice agents are simulated again as decision-making agents. Their decisions and the resulting behavior is based on $V_{exploration}$ (Eq. 3.2.2). $V_{exploration}$ depends, on one side, on the optimized free exploration parameters (β, $p1$, $p2$, $p3$, κ, ψ, bp, Wnm, and Wm, Sect. 3.1.2) that have been indentified on their pre-conditioning sessions and that are not involved in the optimization process for the conditioning behavior (Sect. 3.2.3). On the other side, it depends on the $V_{conditioning}$ values that have been learned during the conditioning sessions and updated during the offline replay phase. In this case these four parameters, α, γ, $\#rs$, and Wc are then optimized to generate a conditioned maze occupancy significantly similar to the one from the data. As showed in Fig. 3.19, all the exploratory phases are divided into multiple sessions. In our simulation, we simulate agents that go through all the separated sessions and then

concatenate their trajectories to evaluate the related behavioral metrics and compare them to the metrics' values in the data, where the same procedure is done.

3.2.3 Model optimization and results

The pre-conditioning exploratory sessions are considered here as habituation phases and, since the behavioral characteristics existing in our previous u-maze seem to exist also in these new data (Fig. 3.20), we optimized the free exploration model parameters on these pre-conditioning sessions, by using the same evolutionary strategies and hyperparameters we used in the previous chapter (Sect. 3.2.3). On the contrary, given that the conditioned behavior is influenced by three different phases, conditioning exploration, sleeping phase and post-conditioning exploration, the parameters concerning these three sessions must be optimized simultaneously. Fig. 3.23 shows a scheme for the two-step optimization performed on this dataset.

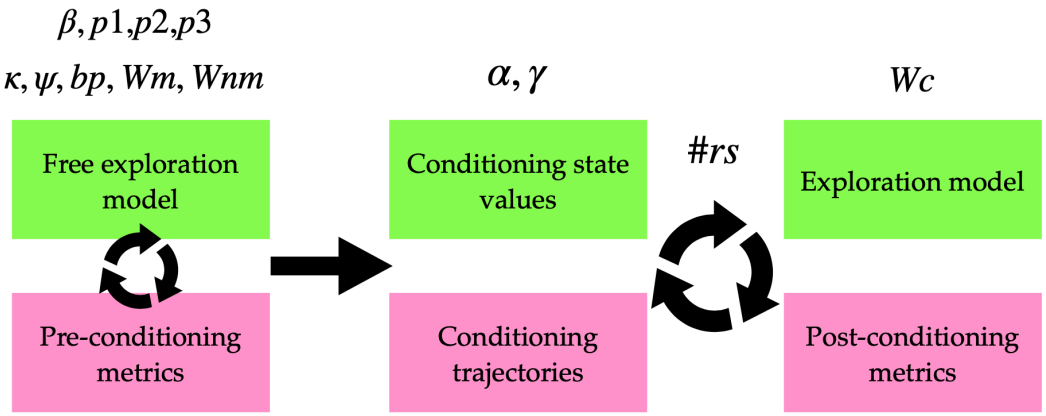


FIGURE 3.23: Scheme showing the parameters that are optimized at the different phases of the simulated experiment through the NSGA-III (Deb and H. Jain, 2013), in the free exploration model case, and through the CMA-ES algorithm (Hansen, 2006), in the conditioned exploration model one. All the variables listed in black out of the boxes are the parameters to optimize in that particular phase. The green boxes indicate where the above parameters are used; either to create a particular behavior in a computational model (free or conditioned) or to learn the conditioning values for each state. The pink boxes indicate which information is taken from the data to evaluate (in the case of the metrics) or generate (in the case of the trajectories) the simulated behaviour.

First, both pre-conditioning phases (the one before the aversive stimulation and the one before the positive one) are considered as a unique set of free exploratory behavioral sessions that are used to optimize the nine parameters of the free exploration models: β, $p1$, $p2$, $p3$, κ, ψ, bp, Wnm, and Wm (Sect. 3.1.2). Once the free exploration model parameters have been optimized on the three behavioral metrics that characterized our proposed model for free exploration: the safety, the biomechanical cost and the biomechanical persistence objectives (Fig. 3.24), the optimization on the conditioned exploration phases can be performed.

As expected, since the free exploration model was able to robustly capture the behavioral components of the other dataset we have on the u-maze, the results are quite solid also in this case (16/18 objectives are strongly significantly closer to the

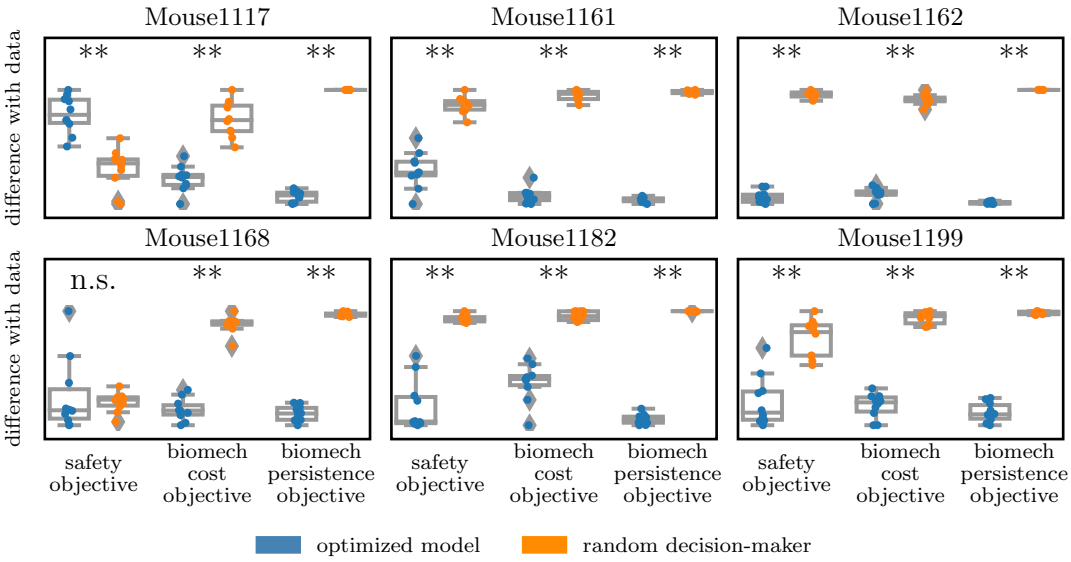


FIGURE 3.24: Comparative statistical analysis on the safety, biomechanical cost, and biomechanical persistence objectives for the selected optimized models and the corresponding random exploration. Each sub-figure represents the results for each mouse agent in terms of behavioral difference with the data. ** indicates that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise.

data than random exploration and, for all the mice, there are at least 2/3 objectives which are strongly significantly closer to the data than random exploration).

Moving forward to the novelty of this chapter, in the second optimization phase, our objective is to generate a model able to describe the changed occupancy of the maze that we observed in the data (Fig. 3.21), as it indirectly reflects the influence of the stimulation (positive or negative) that they have experienced. In this case, the distance between the data masurement and the one obtained from the model is computed as in the biomechanical cost case (Eq. 3.19), as a normalized difference between vectors containing the occupations of the seven sub-areas in the simulated model and in the data.

$$F_{conditioning} = \left|\left| \frac{Cdata}{|\,Cdata\,|} - \frac{Cmodel}{|\,Cmodel\,|} \right|\right| \tag{3.25}$$

where C is a function that counts the number of observations that fall into each of the disjoint categories (known as bins, each bin for one of the seven subareas, Fig. 3.21a) and creates the histogram, respectively for the data $Cdata$ and for the model $Cmodel$.

Since, in this case, there is just one behavioral measurement whose value we want to optimize by minimmizing the difference between our model and the data, we used a state-of-the-art evolutionary strategy for one objective optimization: CMA-ES (Sect. 2.2.5, Hansen (2006)). For this optimization with CMA-ES, we have used the hyperparameters in Tab. 3.3.

max # gen	# ind	σ_0
500	50	0.5

TABLE 3.3: Hyper-parameters for CMA-ES. max # gen is the maximum number of generations, # ind, the population size, and σ_0, the initial standard deviation (Sect. 2.2.5).

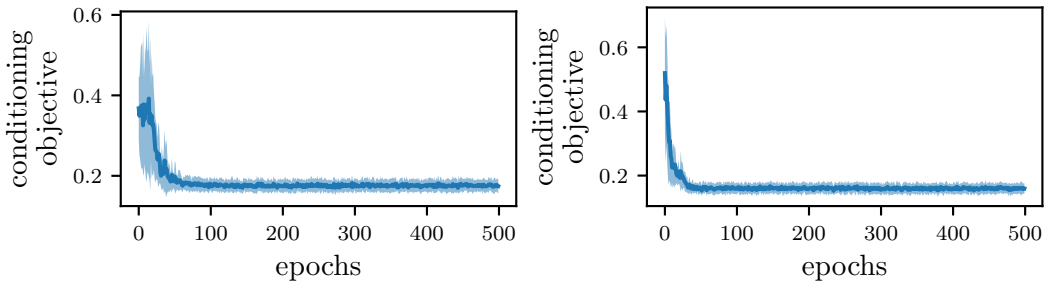
We kept the same maximal number of epochs and the same number of individuals per epochs as we used for the NSGA-III in the previous chapter (Tab. 3.1). The selection for the best individual, *i.e.*, the best set of model parameters, is based on the minimal objective value found through the evolution.

As for the previous chapter, the optimized parameters are the ones in red in the section above and they are four: the weight of the conditioning component inside the exploration model Wc, the learning rate α, the discount factor γ, the number or replay sessions $\#rs$ (Fig. 3.23). Fig. 3.23 shows that α and γ are optimized while the agent is learning the conditioning values of the states of the maze. This learning process is based on the animal's trajectories during the conditioning phases. Furthermore, $\#rs$ optimizes the number of times the whole previous conditioned experiences (in the form of unordered transitions backups) is offline replayed by the animal. Finally Wc describes the relative importance of the conditioning component of the exploration model with respect to the other three components (safety, biomechanical cost and biomechanical persistence) in describing the behavior of the rodents (in terms of the 7 sub-areas occupancies, Fig. 3.21a) during post-conditioning.

Fig. 3.25 shows the results of the two exploration models' parameters optimization (negative and positive conditioning) for Mouse1168 (Fig. A.9-A.10-A.11-A.12-A.13 show the same results for all the other mice in the dataset).We decided to pick this particular mouse to show the behaviour of an average case of the results we obtained from our optimization (Fig. 3.27). Moreover, Mouse1168 is an interesting case to analyze because the optimized number of replay sessions $\#rs$ is very different in the positive (2) and in the negative case (10). Also, its α values are reasonably high (respectively around 0.5 and 0.7 in the positive and negative stimulation case) and Wr is instead smaller (between 0 and 2 in both cases) than for the cases of the other mice. This means that most of the contribution in learning the states' values comes from $\#rs$ and α (Fig. 3.29).

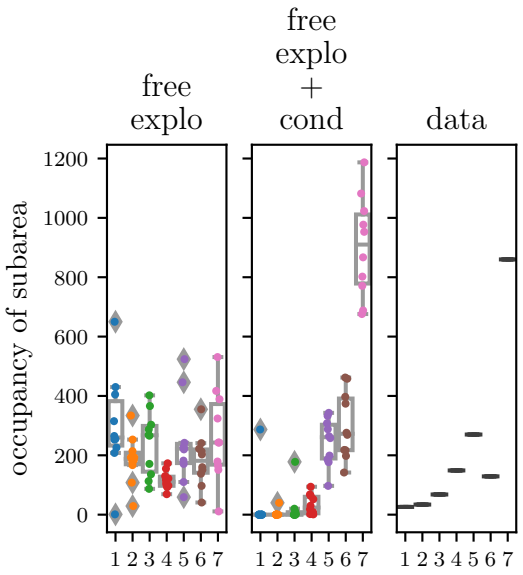
In particular in Fig. 3.25c-3.25d, we want to compare the post-conditioning occupacy of the maze for the two optimized models (free explo and free explo + cond) for the same mouse to prove that the conditioning component and the possible addition of replay session is beneficial to capture the post-conditioning behavior of the animals.

The conditioning objectives, describing the distance in subareas occupancy distribution between the data and the model, have been minimized for both the case of aversive (Fig. 3.25a) and positive (Fig. 3.25b) stimulation. The increasing occupancy of the locations which are far away from the negative stimulation (pink distribution) is better captured by the optimized exploration model (free explo + cond) than by the corresponding version of the optimized free exploration model (free explo), considering the statistics over ten repetitions of the same model parametrizations (Fig. 3.25c). Considering the positive stimulation, since in the case of this particular Mouse1168, there is not a stronger occupation preference for the positive stimulation area (blue error-bar), we can notice that the median of the highest bars are closer to the optimized free exploration model (free explo) values than to the optimized exploration model (free explo + cond) ones (Fig. 3.25d). Even though the optimized

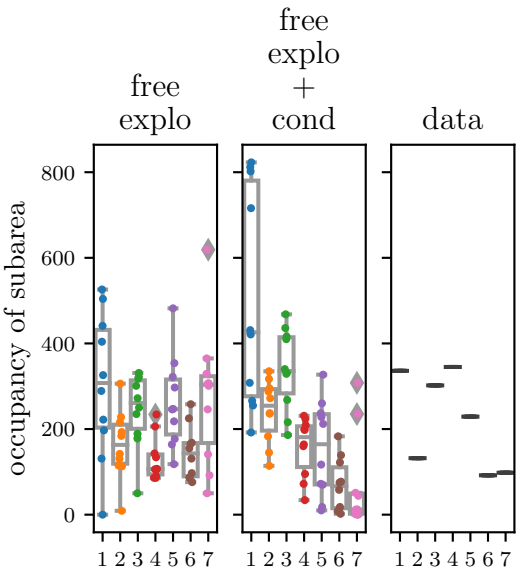


(A) Aversive stimulation

(B) Positive stimulation



(C) Negative stimulation



(D) Positive stimulation

FIGURE 3.25: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1168.

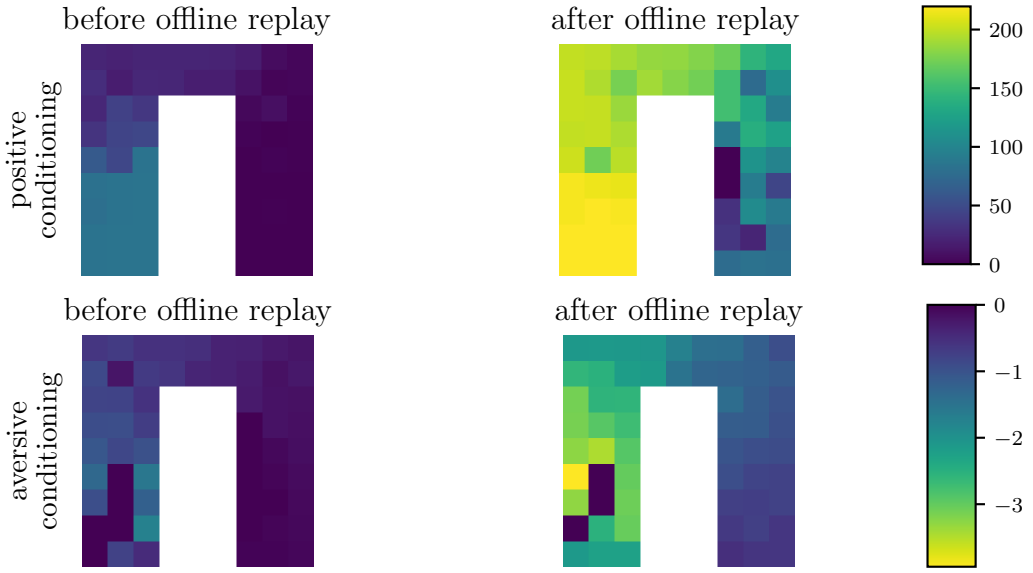
exploration model (free explo + cond) is producing a distribution which is more biased towards the rewarding area than the one observed in the data, the decreasing shape of the data distributions towards the areas far away from the positive stimulation is more reflected in the results from the free explo + cond model than in the ones from the optimized free exploration model (free explo) which are slightly biased towards the reward area. Infact, looking at one example of the occupancy map for Mouse1168 during all the positive post-conditioning sessions (Fig. 3.26c), we can see that the free explo + cond model's occupation is more biased towards the first sub-area compared to the data. However, the free explo model shows no strong preference for one of the two corridors. In particular, this occupancy map shows more occupation in the corridor opposite the positive stimulation.

Here we can remark the strong preference for spending time in corners that characterizes the free exploratory behavior of rodents. As expected by the more convincing replication of the data occupancy in Fig. 3.25c the occupancy of the corridor opposite to the negative stimulation is clearly more matched by the free explo + cond model than by the free explo model which shows a more homogenous exploration (Fig. 3.26b).

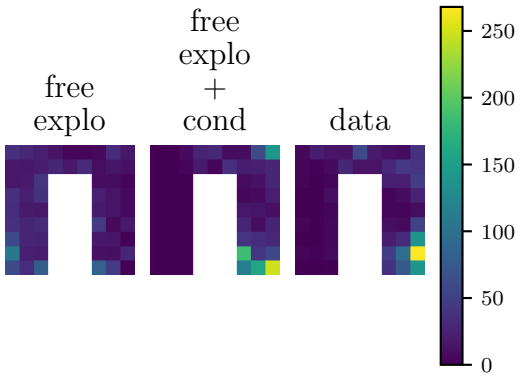
Fig. 3.26a represents an example of evolution and propagation of the conditioning states value $V_{conditioning}$ before and after the offline replay sessions. As explained before in the model description (Sect. 3.2.2), the learning process for positive conditioning assigns positive values to the experienced states. In contrast, the one for negative conditioning assigns negative ones and we represent them with opposite heatmaps to highlight the propagation of the absolute value of the states. For both cases, after the replay sessions, the states-values are increased and propagated all over the end of the corridor opposite to the stimulations. Nevertheless, we note that positive values are two orders of magnitude higher than the negative ones. This large difference is due to the number of times Mouse1168 experienced the stimulation: for over 140 seconds in the positive case, and for less than 4 seconds in the negative one (Fig. 3.30).

By having an overall look at the optimization results for the parameters of the exploration model for the all the mice, the obtained results for the best individuals promisingly show that the exploration model significantly better captures the biased occupancy of the maze in post conditioning than the best individual optimized for free exploration (Sect. 3.1.2). Over ten simulations of the same models, Fig. 3.27 shows that for the majority of the mice (all of them, but negative conditioning for Mouse1161), the global optimized exploration model significantly decreases the error the optimized free exploration model has in describing the conditioned occupancy of the maze in the post-conditioning sessions. More importantly, there are no cases where the free exploration model is significantly better at capturing the post-condition occupancy of the maze than the complete exploration model. This means that, in the case of the negative conditioning of Mouse1161, the optimization has not managed to bring significant improvements to the free exploration model to fit the post-conditioning behavior of the animal, but the performance are comparable.

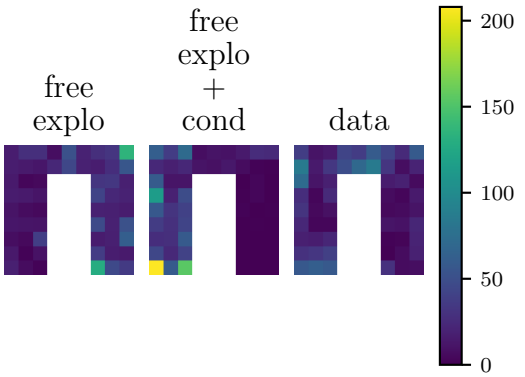
Once we have verified that, in the majority of the cases (11/12 considering both the positive and the negative conditioning in Fig. 3.27), the optimization algorithm can find a proper model for the post-conditioning occupancy of the seven areas, our interest is in analysing if any relationship exists between the optimized number of replay sessions and the experimental and behavioral characteristics of the animals. To assess and valid the proposed learning strategies (Sect. 3.2.2), we investigate if the exploration model can reflect some phenomenological relationships observed in the analyses of our collaborators (Bryzgalov, 2021).



(A) Conditioning states values

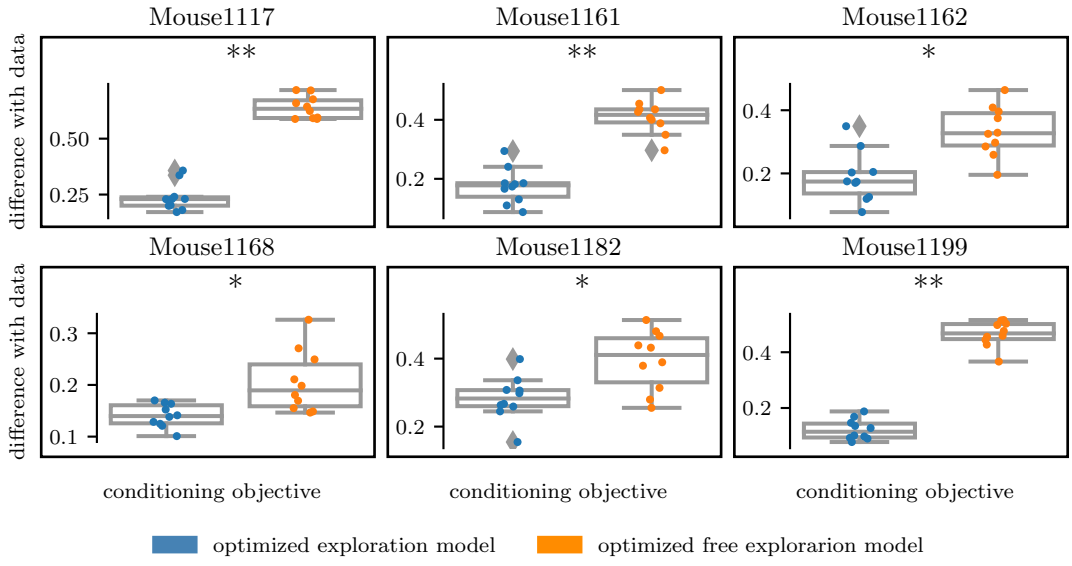


(B) Aversive stimulation

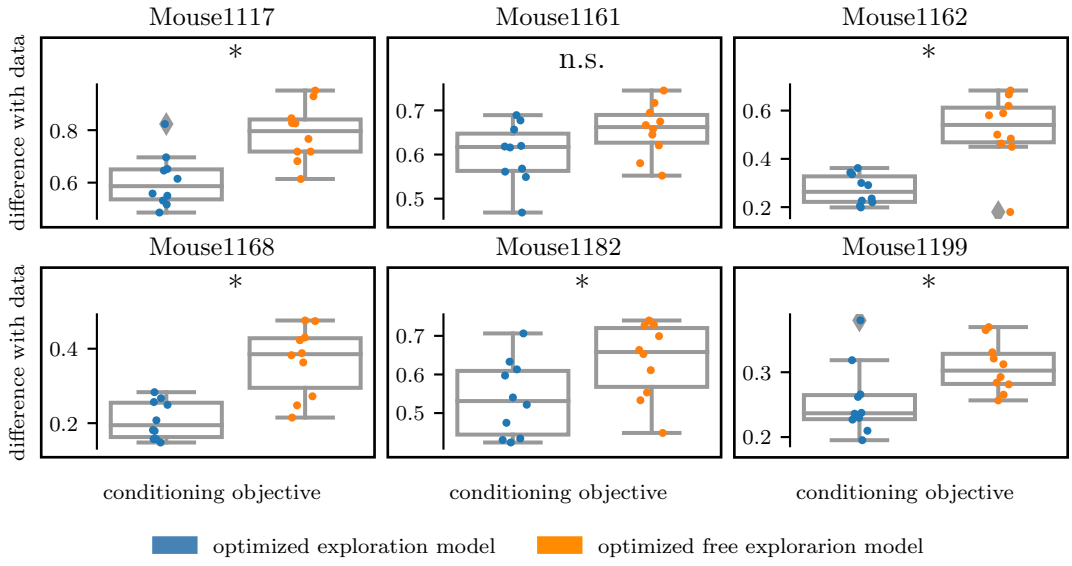


(C) Positive stimulation

FIGURE 3.26: Example of learning and behavior for the optimized exploration model for Mouse1168. A) Learned states values maps before and after the replay sessions: positive and negative conditioning cases. B) Comparison of the post-conditioning occupancy map among the optimized free exploration model (free explo), the optimized model (free explo + cond) and the data.

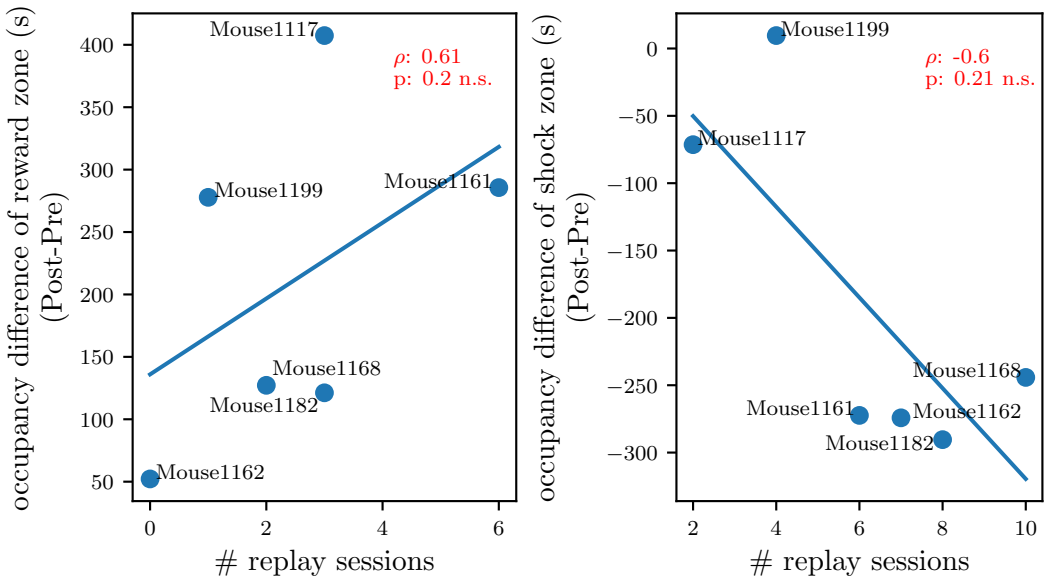


(A) Positive conditioning



(B) Negative conditioning

FIGURE 3.27: Comparative statistical analysis on the conditioning objectives for the selected optimized models and the corresponding free exploration model. Each sub-figure represents the results for each mouse agent in terms of behavioral difference with the data. This difference is expressed as the conditioning objective (Eq. 3.25). ** indicates that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise.



(A) Positive stimulation (B) Negative stimulation

FIGURE 3.28: Linear correlation between the number of replay sessions in the optimized model and the occupancy difference to enter in the stimulation zone in the post-conditioning exploration data. ρ is the Spearman correlation coefficient and p is the p-value for this correlation test.

To investigate the role of hippocampal reactivation in spatial learning, it is interesting to look for correlations between the amount or type of hippocampal reactivations, and the change in the behavior before and after learning. In the case of our experiments, since the pre-conditioning behavior is significantly different to the post-conditioning one, in term of spatial occupation of areas close the stimulation (Tab. A.1), it would be interesting to test if a correlation between number of replay sessions and occupancy of the stimulation areas exists. Following the interesting observations and results obtained by Bryzgalov (2021), we decided first to perform the same analyses they did on the data, by replacing the explained variance (EV) of ripples with the optimal number of replay sessions for a particular mouse (Fig. 3.28). Measuring the EV is a common approach to identify reactivations. It measures how much the activity patterns observed during post task rest or sleep can account for wakeful activity patterns during the task (Kudrimoti, C. A. Barnes, and McNaughton, 1999; Ólafsdóttir, Bush, and Barry, 2018).

Regarding the correlation between the number of replay sessions and the difference in occupancy of the shock zone between post- and pre-conditioning (Fig. 3.28b), we found a similar Spearman's ρ and p coefficients than the one found by Bryzgalov (2021)'s data (Fig. 5.5-L, Spearman's ρ is 0.64 and p is 0.14). The difference between the occupancy times of the shock zone in the post- and in the pre- conditioning sessions goes negative for almost all mice because the time spent in the shock areas after the negative stimulation is lower than the one spent there before it. Then, we look at the replay-occupancy relationship in the case of the positive conditioning experiment (Fig. 3.28a). Interestingly, the more replay session are required by the optimized exploration model, the longer is the corresponding occupancy of the stimulation areas. We found this tendency slightly stronger in this positive case than in the negative one, and it would be interesting to further investigate this prediction either computationally or experimentally. In conclusion, these results suggest that in

both the positive and the negative stimulation case, the replay activity could predict and be correlated to the level of occupancy and avoidance of the stimulation area respectively.

3.2.4 Discussion

In this chapter, a generalization of the free exploration model we proposed in Sect. 3.1.2 is presented and validated against a similar but new dataset. This generalization extends the definition of state-value to the case where it can be learned by the agent-environment-conditioning interactions through a reinforcement learning rule. To investigate and make predictions on the possible role of hippocampal replay in spatial learning, the optimization process for the parameters of our model includes the possibility of performing offline model-free replay. This replay session happens after the animal has finished experiencing the conditioning and just before it re-enters the maze for the post-conditioning phase.

The optimization of the parameters of the exploration model was able to get a lateralized occupancy of the maze, which is significantly closer to the one presented in the data than the one we can have by using just the free exploration version of the same model. Our results predict that all the mice (except one, Fig. 3.29, $\#rs$), if they have the possibility to perform replay sessions, need at least one of them to happen, for better fitting the post-conditioning occupancy of the data. Interestingly, very similar correlations to the ones presented in the Bryzgalov (2021) data analyses exist between the number of replay sessions and the difference in the time spent in this zone between post- and pre-conditioning (Fig. 3.28b).

Our research interest in this part of the thesis was to investigate and try to predict with our model if hippocampal reactivations could play a differential role or have a different saliency when the emotional valence of the conditioning changes. The proposed model is based on one of the most simple and classic RL mechanisms to reproduce experience reactivations. We propose to start investigating the hypothetical existence of these differences by using the simplest possible computational mechanism for replay. To answer this question, we did some analyses on the parameters of the best individuals, which resulted from the optimization process either in the positive or the negative conditioning case (Fig. 3.29).

The most relevant learning parameters, $\#rs$ and α, result in statistically higher values in the case of negative conditioning (n) than in positive conditioning (p). Following the q-learning rule (Eq. 3.23 and Eq. 3.24), the states-values propagate mainly depending on the coordinated action of these parameters. Naturally, Wr is also important in the states-values learning because it represents the relative weight of the conditioning component with respect to the other components. In support of the common trend for the other parameters, it is also increasing its value from positive to negative conditioning for 5/6 mice, even though the two overall distributions, p, and n, are not significantly different. The parameter which probably weighs more in the states-values learning is the number of unordered replay sessions $\#rs$ since its modulation multiplies the contribution of the other parameters (α, γ and, Wr) by increasing the number of time when the agent learns (updates the states-values) on its overall past experience.

With the possibility to optimize α, γ, Wr and, $\#rs$, it is interesting that the optimization finds solutions (set of model's parameters) which almost always (11/12 mice considering both the positive and the negative stimulation experiments) need a number of replay sessions greater or equal to 1 (Fig. 3.29). This suggests that,

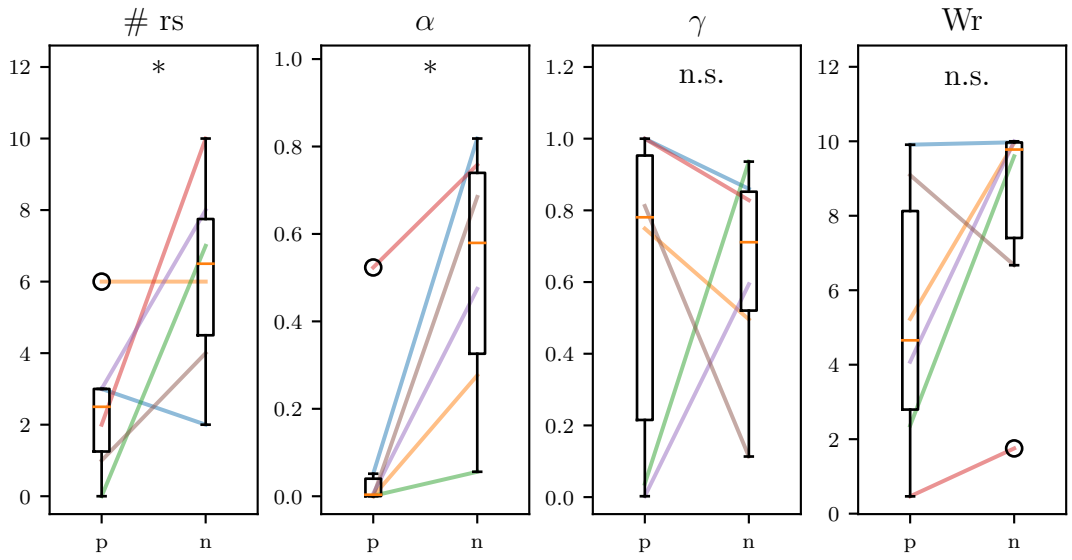
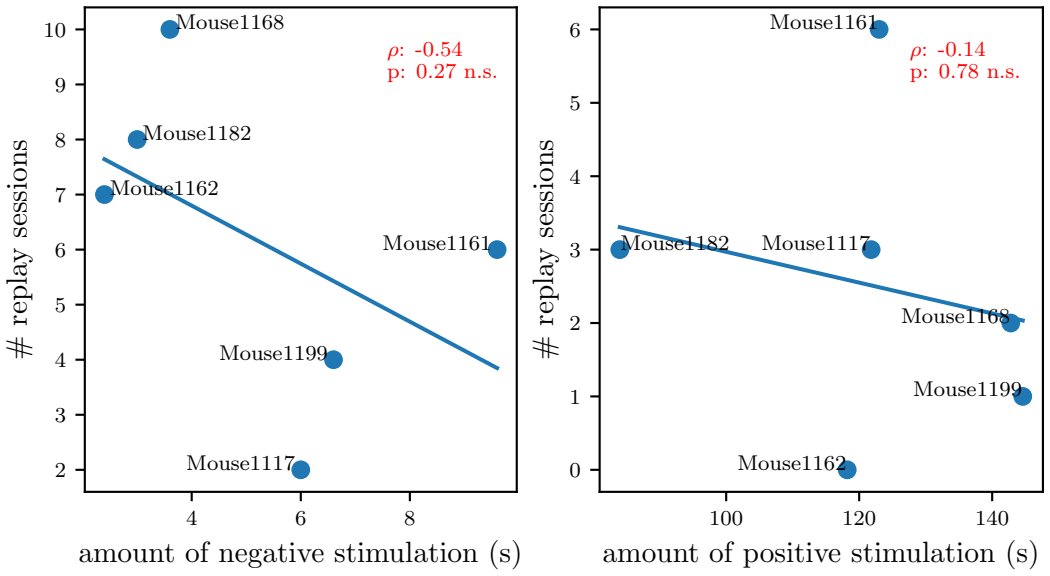


FIGURE 3.29: Statistical analysis on the exploration model parameters for the best individuals found by CMA-ES, in the case of the positive stimulation data (p) and negative stimulation ones (n). *#rs* indicates the number of replay sessions, α the learning rate, γ the discount factor, and finally *Wr* the weight for the conditioning component. * means that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test between the distributions of the model parameters in positive and negative stimulation is lower than 0.05 otherwise it is non-significant (n.s.).

from the proposed computational point of view of our model, longer replay sessions are preferred to higher values of the learning rate α, for example, to amplify the learning response of the stimulation sessions. If we assume that the RL is a proper framework to model instrumental behavior, spatial learning, and hippocampal reactivations and also that we appropriately use the evolutionary optimization strategy. In that case, we can infer that increasing the number of replay sessions *#rs* is more computationally efficient than increasing the learning rate α, to replicate the animals' occupancy of the maze. The relevance of replay sessions in replicating the post-conditioning behavior of the animals is also evident from the fact that, in the aversive stimulation case, the winning set of parameters still have *#rs* values greater or equal to 2, despite showing α values significantly higher than the ones in the positive stimulation's case.

Looking at the whole set of the learning parameters, it is striking that a significant increase of α and *#rs* exists when the model needs to fit the post-conditioning behavior after the aversive stimulation, compared to the positive one. This suggests that, from the MF-RL perspective of the model, offline replay is significantly more important when the valence of the conditioning is negative. Furthermore, in our results, replay sessions are needed more when the animal has experienced very few emotional stimulations (Fig. 3.30a).

These results make us predict that the amount and content of the information replayed during offline hippocampal reactivations are not purely linked to spatial experience. In fact, considering that all the mice explore for the same duration, we observe the urgency for more replay sessions when the negative stimulation is less experienced (Fig. 3.30a). This suggests that when there is more emotionally relevant



(A) Aversive stimulation

(B) Positive stimulation

FIGURE 3.30: Linear correlation between the amount of stimulation received by the mice and the number of needed replay sessions to represent the post-conditioning sub-areas occupancy of the maze. ρ is the Spearman correlation coefficient, and p is the p-value for this correlation test.

negative content, rodents need fewer replay sessions to consistently avoid the shock area. However, the same correlation is not present in the case of positive stimulation (Fig. 3.30b).

Despite the observation of replay during NREM sleep between the conditioning and post-conditioning experimental sessions (Bryzgalov, 2021), in Sect. A.1 we report some supplementary results, where we optimized the same exploration model presented in this chapter but with offline replay sessions disabled. We found out that our optimization procedure can find behavioral results (in terms of post-conditioning occupancy of the sub-areas) which are comparable to the ones of the optimized model with replay in most of the cases (9/12, in Fig. A.1). The results for the states-values propagation and the post-conditioning occupation of the maze for all the 10 repetitions of all the mice for the two models (with or without replay) are reported in Fig. A.14-A.25. We can see further analyses we have performed, always in Sect. A.1, showed that a tendency exists for the learning rate α to have higher values in the no replay model than in the one with replay (Fig. A.2b, but in particular in Fig. A.2a). Even though we have just a few mice behavioral data to analyze and the difference between the two distributions of α is not significant, these supplementary results suggest that the absence of the possibility of offline replay sessions leads the model optimization to compensate for them with higher learning rates. In case of complex scenarios, like in the case of our conditioning sessions where the agent replicates the animal's behavior, higher α can eventually hinder the convergence of the TD learned states values and lead to unstable behaviors (R. S. Sutton and A. G. Barto, 2018). That is why reinforcing learning with offline replay could be more efficient and robust than increasing the learning rate. It is also the solution that the great majority of the optimized exploration models, with the possibility to have replay sessions, choose (11/12 optimized models cases in Fig. 3.30). Considering these other results

implies that offline reactivations are not necessary to fit the post-conditioning occupation of the maze in either the positive or the negative conditioning. However, they are computationally preferred as a mechanism to enhance the states' values learning. Adding to the model the possibility to perform offline replay is adding computational cost to the model, but not modeling complexity since the backups done during these sessions are just unordered MF updates of the same TD-learning rule of the conditioning component. Thus, we consider it interesting trying to apply and compare the performances of these two versions of the model (with or without replay) on more spatial learning rodent datasets, where the resulting exploration is even more biased than the one of our data (data in Fig. 3.25,-A.9-A.13). Also, Fig. A.2c interestingly shows that a trend for higher values of α and Wr in the negative conditioning case, compared to the positive conditioning one, also exists in the case of an exploration model without replay. The same preferences have also been observed in Fig. 3.29 for the conditioning exploration model with replay. These other results confirm that, from the RL computational point of view of our model, stronger learning dynamics are needed in negative conditioning to fit the spatial occupational preferences of the mice.

The amount of stimulation received by the mice is much smaller in the negative cases than in the positive ones (two orders of magnitude smaller, axes x in Fig. 3.30), as we expect from external conditions which cause aversive emotional reactions in these animals. To recap, Fig. 3.30 implies that hippocampal replay is more important for replicating the lateralized occupation of the maze we see in Fig. 3.21b when the stimulation is negative. Also, the amount of replay sessions needed in this case is inversely proportional to the aversive stimulation received by the animal (Fig. 3.30a). In other words, our data-driven model suggests that when the exogenous conditioning is a punishment, animals obviously do not want to experience it many times, but however they want to have a robust avoidance of the punishment area. The same model we propose, with two opposite signed-learning rules, optimized on the animals' behavior, needs significantly higher learning rate α, and longer replay sessions $\#rs$, when the stimulation is negative (Fig. 3.29). As far as we know, this computational hypothesis, as raised here, concerning the significance of longer hippocampal offline reactivations periods in negative conditioning, has not been directly investigated yet. Aversive experimental protocols are more difficult to address for ethical reasons and for the freezing aversive reactions that they induce, particularly in rodents. Freezing reactions, happening just during negative conditioning and not during positive one, make it difficult to directly compare the exploratory dynamics of two equal experiments with opposite valence stimulations.

Also concerning freezing behavior, but on a different note, Bryzgalov (2021) documented an interesting phenomenon concerning rodent freezing reflex in the case of aversive learning. They observed that a small subpopulation of dorsal hippocampal interneurons strongly decreases its firing rate during freezing behavior. They have also remarked that awake SWRs happen mostly in the most visited maze locations, which are also the ones where the animals exhibit freezing behavior. Thus, they hypothesize a strength and rate modulation of SWRs during freezing, even though the exact dynamics of this phenomenon have not been clarified yet. Once additional information on the locations and timing of freezing behavior is given, an interesting expansion of the model would be to model the contribution of awake replay during the conditioning phases (in particular, during freezing behavior, in the aversive stimulation case).

As a future perspective of these results, the model could be improved to describe

the spatial characteristics of hippocampal reactivations. To do that, it will be interesting to use an RL prioritized sweeping replay algorithm able to produce spontaneous replay activity of the experiences that were the most unexpected (Aubin, Khamassi, and Girard, 2018). This model-based replay algorithm would suggest predictions on the location of the hippocampal reactivation that could then be validated by comparing the model observations to the analyses on the location of the explained variance of the hippocampal SWRs recorded and analyzed in Bryzgalov (2021) and other similar works.

To summarize the contributions of the neuroscientific chapter of this thesis:

- We have designed a computational free exploration model for rodents that can simulate the behavior of an agent whose decision-making process reflect the main observed behavioral preferences of the animals, such as safety, biomechanical cost, and biomechanical persistence;
- We have validated such a model on new rodents' datasets;
- We have extended the proposed model to account for spatial learning scenarios by considering two parallel learning rules for opposite valence conditioning;
- Without prior assumptions, the optimisation of the parameters of our exploration model on the animals' behavior subsequent to positive and negative conditioning suggests that negative stimuli are perceived to have an higher valence than positive ones and that more offline reactivations are required.

In the next chapter, we will discuss the contribution of hippocampal reactivations-inspired RL mechanisms, similar to the one adopted in this chapter, in enhancing robotic goal-directed navigation. In particular, we will discuss and test multiple RL replay mechanisms (namely model-based and model-free) to identify the most promising strategies once the experiments get closer to a realistic scenario. Finally, a brief description of a robotic demonstration, conceived and realized in our laboratory, is given to illustrate the advantages of immersive real robotic navigation tasks to efficiently explain hippocampal reactivations, also from a neuroscientific perspective.

Chapter 4

Scientific contributions in machine learning and robotics

From RL-based models of hippocampal reactivations that investigate rodents' goal-directed behavior in navigation tasks, the focus passes on to the contribution of the same strategy and more complex ones in robotic goal-directed navigation. Sect. 4.1 evaluates and analyses multiple Model-Free- and Model-Based-RL reactivation strategies in robotic contexts, looking at their implication when moving from theoretical simulations to experiments on real robotic platforms.

Finally, Sect. 4.2 presents and discusses the conceptualization and design of a demonstration for dissemination purposes concerning hippocampal reactivations for robotics.

4.1 Model-based and model-free replay mechanisms for reinforcement learning in neurorobotics

Experience replay has been modeled and used in artificial intelligence (AI) to allow agents to reuse past experience and learn more efficiently (Lin, 1992) since before the adoption of the reinforcement learning's theory to model the computational contribution of dopaminergic activity (Schultz, Dayan, and Montague, 1997) and hippocampal reactivations, particularly in rodents (Khamassi and Humphries, 2012). Since these first research studies in AI, many mechanisms, including unordered, reverse-ordered, and prioritized memory buffers, have been tested in different tasks. They showed different learning properties which are suitable for specific experimental scenarios (*i.e.*, simulation or robotic experiments, static or dynamic task, Cazé et al. (2018)). In Sect. 4.1.2 - 4.1.3 - 4.1.4, we test different neuro-inspired RL control architectures in navigation tasks of increasing complexity. In particular, we take inspiration from the instrumental conditioning dichotomy between habitual and goal-directed learning systems in the brain (Sect. 2.1.2) to test the coordination of model-based and model-free replay strategies in neurorobotics' navigation tasks. Since, the brain mechanisms which orchestrate hippocampal replay are still unclear, the intent of the following study is also to raise new neuroscientific hypotheses starting from our simulation and robotic results on which replay technique turned out to be more efficient in distinctive scenarios and situations.

Eventually, in Sect. 4.2 we illustrate TaVAR, a new robotic device to explain to the general public and real-time visualize how classic RL algorithms work in a goal-oriented navigation task. The proposed platform shows the audience which spatial information the robot uses to learn an optimal behavior to get to the desired location and also how replaying past spatial transitions can speed up this learning. This

visual and robotic set-up could also ease the explanation of the computational mechanisms that are thought to be behind hippocampal reactivations in neuroscience.

Sect. 4.1.1 - 4.1.2 - 4.1.3 - 4.1.4 - 4.1.5 present a reformated version of a published paper: Massi, E., Barthélemy, J., Mailly, J., Dromnelle, R., Canitrot, J., Poniatowski, E., Girard, B. and Khamassi, M. (2022). Model-Based and Model-Free Replay Mechanisms for Reinforcement Learning in Neurorobotics. Frontiers in Neurorobotics, 16.

4.1.1 Introduction

For a reinforcement learning (RL) agent (R. Sutton and A. Barto, 1998; Lin, 1992), experience replay consists in storing in (episodic) memory a buffer containing a series of observations (*i.e.,* a quadruplet composed of: the previous state, the action, the new state, the reward), and periodically replaying elements from this buffer to bootstrap learning during offline phases (*i.e.,* between phases where the agent acts and samples new observations in the real-world) (Fedus et al., 2020).

Several important parameters have an impact on the performance of RL agents with experience replay, such as the size of the memory buffer (Zhang and R. S. Sutton, 2017), the relative time spent learning from replay versus the time spent collecting new observations in the world (Fedus et al., 2020), or whether to shuffle the memory buffer and uniformly sample elements from it or to prioritize elements as a function of their associated level of surprise (*e.g.,* the absolute reward prediction error associated to a given quadruplet observed from the environment) (Moore and Atkeson, 1993; Peng and Williams, 1993; Schaul et al., 2015).

To our knowledge, these replay techniques have their origin in the 1990s, when Long-Ji Lin at Carnegie Mellon University proposed solutions to enable RL reactive agents (*i.e.,* model-free (MF) agents such as Q-learners (Watkins, 1989)) to bootstrap their learning process in large dynamic (non-stationary) discrete simulation environments (Lin, 1992). One of the investigated solutions was to use the Dyna-Q architecture (R. S. Sutton, 1990) to learn action models and use these models to sample hypothetical actions. Another tested solution consisted in storing the agent's experience in a memory buffer and replaying it to bootstrap learning. Interestingly, one of the main results was that the best performance was obtained by reversing the order of the replay buffer, what we will call *backward replay* (*i.e.,* replaying first the most recent observation, then the second-to-last one, and so on until the oldest observation). This is because each time the buffer contains a rewarding observation, replay leads to increasing the value of the action performed in the previous state, followed by replaying precisely that previous state at the next iteration (because the buffer is in reverse order), and thus increasing the value of the preceding action, and so on. Consequently, single processing of the memory buffer results in reward value propagation from rewarding states along the whole sequence of actions that the agents had experienced to get the reward.

In parallel, other researchers further investigated the efficiency of model-based (MB) techniques to sample hypothetical actions rather than replaying experienced actions from a memory buffer. One example is called *prioritized sweeping* and consists in replacing uniform model sampling with a prioritization that depends on the absolute value of the reward prediction error (Moore and Atkeson, 1993; Peng and Williams, 1993). While model-based methods can be conceived as ways of planning, thus different from model-free learning, they can nevertheless be seen as an

alternative way to perform offline Q-value updates. Further, there is a mathematical equivalence between the sequence of Q-values obtained with model-based updates and with model-free methods with replay (Seijen and R. Sutton, 2015). This is why throughout this paper, we will discuss both *model-based* and *model-free replay*, in the sense that they represent alternative offline reactivation mechanisms to update action values. We will refer to model sampling as *Simulation Reactivations* and sampling from a memory buffer as *Memory Reactivations*.

Strikingly, neuroscience research has found that the mammalian brain also seems to perform some experience-dependent reactivations of neural activity, particularly in a part of the brain called the *hippocampus* (M. A. Wilson and McNaughton, 1994). These reactivations occur either when an animal is sleeping (Ji and M. A. Wilson, 2007) or during moments of quiet wakefulness between trials of the task (Karlsson and Frank, 2009). Most importantly, these reactivations play an instrumental role in learning and memory consolidation since blocking these neural reactivations leads to impaired learning performance (Girardeau et al., 2009; Ego-Stengel and M. A. Wilson, 2010; Jadhav et al., 2012), while new memories can be created by stimulating reward circuits during these reactivations (De Lavilléon et al., 2015).

The computational neuroscience literature has recently compared the different replay techniques from machine learning with the properties of hippocampal replay recorded experimentally (Pezzulo, Kemere, and Van Der Meer, 2017; Cazé et al., 2018; Mattar and Daw, 2018; Khamassi and Girard, 2020). Interestingly, the reactivation of a sequence of states experienced by the animal during the task sometimes occurs in the same *forward* order, and sometimes in *backward* order (Foster and M. A. Wilson, 2006; Diba and Buzsáki, 2007). Nevertheless, a large part of hippocampal reactivations occur in apparently random order, and the underlying computational principle remains to be explained (see for instance the proposal of Aubin, Khamassi, and Girard (2018)). Moreover, computational investigations recently found that prioritized sweeping can also explain some properties of hippocampal reactivations (Cazé et al., 2018; Mattar and Daw, 2018). However, it is not yet clear whether a single unified computational principle can explain hippocampal replay, or whether the brain alternates between different types of replay (backward, shuffled, prioritized / model-free versus model-based) in different situations (sleep versus quiet wakefulness, depending on the difficulty of the task, the level of noise/uncertainty).

Thus, a new field of neurorobotics research is currently dedicated to integrating offline reactivations in the reinforcement learning processes to improve and speed them up. As mentioned above, this focus on offline reactivations is inspired by the machine learning techniques created in the 90s and now commonly used in DeepRL, as well as by the neuroscience results on hippocampal reactivations and the probable cohabitation of model-based and model-free RL systems in the brain. With robotic applications as an aim, these contributions must bridge the gap between perfectly controlled discrete state simulations and real embodied robotics experiments in continuous environments. The goal of this research is to understand which replay techniques give the best learning performance in different situations (constrained corridor-based versus open maze environments; non-stationary goal locations and maze configurations) and whether robotic tests lead to different conclusions than simple perfectly controlled simulations (physical versus abstract simulations, autonomous state decomposition by the robot, noisy perception). For instance, a recent neural network-based simulation of a rat maze task highlighted that shuffled experience replay was required to break the temporal data correlations, to learn a neural internal world model (Aubin, Khamassi, and Girard, 2018). Notably, while neurorobotics research during the last 20 years had already studied hippocampus

models for robot navigation (Arleo and Gerstner, 2000; Fleischer et al., 2007; Dollé et al., 2008; M. Milford and G. Wyeth, 2010; Caluwaerts et al., 2012; Jauffret, Cuperlier, and Gaussier, 2015), to our knowledge the impact of different types of replay on the performance of these models has only recently started to be investigated.

In this paper, we illustrate this line of research by presenting a series of numerical simulations of laboratory mazes (used to study rat navigation in neuroscience) as benchmark tasks for robotic learning. These simulations are presented in order of increasing complexity toward real-world robotic experiments. At each step of this presentation, we simulate and compare different replay techniques in either model-free or model-based RL agents. We discuss the properties of these simulations, how they contribute to improving learning in robots, and how they can also help generate predictions for neuroscience.

4.1.2 Simulation of individual replay strategies in a predefined discrete state space

This section presents a series of numerical simulations in a simple deterministic maze task with predefined state decomposition. The task mimics the multiple T-maze of Gupta et al. (2010), where rats have to follow constrained corridors and make binary decisions (go left or go right) at specific T-like decision points (Fig. 4.1a). This will enable us to illustrate the properties of different replay methods in the same conditions as the perfectly controlled simulations usually performed in computational neuroscience work. Then in the following sections, we will study what happens in more open mazes where moreover, the robot will autonomously build its state decomposition.

The work presented in this section contains two main differences with our previous computational neuroscience simulations of the multiple T-maze task (Cazé et al., 2018; Khamassi and Girard, 2020)[1]: (1) in previous work, following experience replay techniques in machine learning, we had allowed the agent to perform a series of replay iterations after each action; here, because it would be energy- and time-consuming for a robot to stop after each action, we allow the simulated robot to perform replay only at the end of the trial, while it is waiting for the subsequent trial at the departure state; (2) we had simulated a version of *model-based (MB) prioritized sweeping* where the memory buffer contained one element per state; here, we test whether it is also efficient to have an element for each (state,action) couple, thus filling the memory buffer with multiple elements for the same state (as long as they represent different actions).

Methods

We simulate the multiple T-maze task as a Markov Decision Problem (MDP), where an agent visits discrete states $s \in \mathcal{S}$, using a finite set of discrete actions $a \in \mathcal{A}$. States represent here unique locations in space, equally spaced on a square grid (Fig. 4.1a), a piece of information expected to be provided by place cell activity in the hippocampus (O'Keefe and Dostrovsky, 1971). The actions allowed the agent to represent moves in the four cardinal directions: north, south, east, and west. During the first 100 trials, the reward will always be located on the left arm. Then during the successive 100 trials, the reward will be on the right arm, and the agent will have to adapt its decisions accordingly.

[1]The updated code for these simulations is available at https://github.com/MehdiKhamassi/RLwithReplay

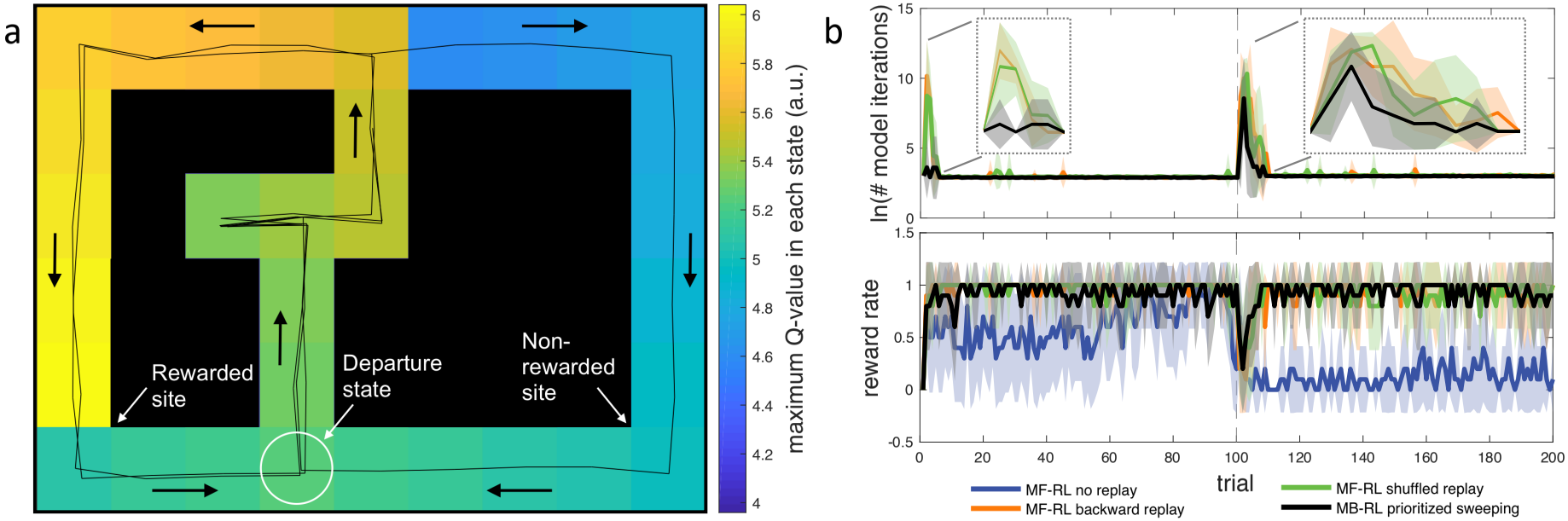


FIGURE 4.1: A) Discrete state-space simulations in the multiple T maze task Cazé et al., 2018; Khamassi and Girard, 2020. The reward is on the left side for 100 trials and then shifted to the right side for the next 100 trials. In the present simulations, replay is only allowed in the departure state before starting the next trial. Despite this constraint, the figure shows that after only 3 trials (2 correct / 1 error), the MF-RL algorithm with backward replay has already learned a full gradient of Q-values across the maze. B) Comparison of the performance (reward rate) and computation time (Napierian logarithm of the number of iterations during replay phases) for 4 different algorithms. The thick lines represent the average, and the area around represents the mean square error. The figure illustrates that MF-RL without replay requires 60-70 trials to reach optimal performance, and does not manage to adapt to the change in reward location within only 100 trials. All the other algorithms perform similarly in terms of reward rate: rapid increase in performance, brief drop in performance after the change in reward location, fast re-increase of performance afterward. These algorithms mainly differ in the required duration of the replay phases: MF-RL with random replay and MF-RL with backward replay both show a strong peak in the number of replay iterations after the change in goal location. The state-based version of MB-RL prioritized sweeping shows a smaller peak.

Here we simulate three model-free reinforcement learning algorithms and one model-based one: MF without replay, MF with backward replay, MF with shuffled replay, and MB prioritized sweeping (Table 4.1).

	MF no replay	MF backward replay	MF shuffled replay	MB prioritized sweeping
α	0.2	0.2	0.2	-
γ	0.99	0.99	0.99	0.99
β	3	3	3	3
ϵ	-	0.001	0.001	0.001
N	-	54	54	54

TABLE 4.1: Algorithm parameters used to generate the results in this section. They have been taken from Cazé et al. (2018) without retuning. α is the model-free (MF) learning rate. γ is the discount factor. β is the inverse temperature in the softmax for decision-making (Equation 4.2). ϵ is the threshold for Q-values convergence during replay. N is the maximal size of the episodic memory buffer.

For each Markovian state-action couple (s, a) in the environemnt, MF-RL agents use Q-learning (Watkins, 1989) to learn the Q-value of performing action a from state s, as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{4.1}$$

Where $R(s,a)$ is the reward obtained from the environment when performing (s,a), and s' is the arrival state after executing action a in state s.

At the next timestep, deciding which action to perform is computed by drawing the next action a from a probability distribution given by the softmax Boltzmann function applied to the Q values:

$$P(a|s) = \frac{e^{\beta Q(s,a)}}{\sum_{i \in \mathcal{A}} e^{\beta Q(s,i)}} \tag{4.2}$$

With $\mathcal{A}$ being the set of all the possible actions from state s and β being the inverse temperature parameter that regulates the compromise between exploration and exploitation: the closer to zero, the more the differences between the Q-values will be attenuated, and thus the more the selection will be uniform (hence exploratory); conversely, large values (that can go up to infinity) will enhance the contrast between the Q-values and will thus favor exploitation of the largest one.

In *MF-RL backward replay* and *MF-RL shuffled replay* and for all the other RL replay algorithms tested in this section and the next one (Sect. 4.1.3), we enable the agent at each timestep to store in a memory buffer the quadruplet describing the current observation: the previous state s from which the agent performed action a, the resulting state s' and the scalar reward r obtained from the environment (1 when the rewarding state has been reached, 0 elsewhere). This memory buffer progressively increases in size, timestep after timestep, but is limited by the maximal size N (N being chosen to correspond to the number of states in the environment, see Table 4.1). When the maximal size has been reached, adding a new element to the buffer is accompanied by throwing away the oldest element.

When the agent has finished the current trial and reaches the departure state again, a replay phase is initiated where at each replay iteration one element from

the buffer is processed and the corresponding Q-value is updated following Equation 4.1. This is repeated until the sum of variations of Q-values over a window of N replay iterations is below a certain replay threshold ϵ, which indicates that the Q-values have converged and do not require to be updated anymore.

In the *MF-RL backward replay* algorithm (Lin, 1992), at the beginning of a new replay phase, we reverse the order of elements in the buffer and then perform replay iterations following the procedure explained above. In the *MF-RL shuffled replay* algorithm, we shuffle the buffer elements before starting the replay phase.

We also test a model-based algorithm where the learning process aims at building a world model, *i.e.,* a model of how the perceived world changes when actions are taken. This model is conventionally composed of a transition function and a reward function. The transition function $T(s,a,s')$ represents the probability of observing s' next, if action a is taken while in state s. In the present discrete case, it is built by storing the number of times each (s,a,s') triplet was encountered and dividing by the number of times (s,a) was experienced, as shown in the equation below:

$$T(s,a,s') = \frac{V_N(s,a,s')}{V_N(s,a)} \tag{4.3}$$

where $V_N(s,a)$ stands for the number of visits of state s when action a is then chosen and $V_N(s,a,s')$ is the number of transition from state s to state s', having performed action a. The reward function $R(s,a,s')$ represents the average reward signal experienced when effectively performing the (s,a,s') transition. For the *MB-RL prioritized sweeping* algorithm that we simulate here (Moore and Atkeson, 1993; Peng and Williams, 1993), we add to each element in the memory buffer the absolute reward prediction error Δ measured when experiencing (s,a,s',r) in the world. This Δ can also represent the magnitude of change in $Q(s,a)$, which resulted from this observation. The memory buffer is sorted in decreasing order of Δ, thus giving a high priority to be replayed to elements representing surprising events in the world that resulted in important revisions of Q-values. In fact, Mattar and Daw, 2018 has formally shown that deriving *Expected Value of (Bellman) Backup* (in other words, an expected value of doing a replay) leads to maximizing a *gain* term, which is higher for transitions that have been associated to larger reward prediction errors (hence larger surprise) when the agent was experiencing the real world.

During the replay phase of *MB-RL prioritized sweeping*, we start by considering the first element (s,a) of the buffer with the highest Δ. We use the world model learned by the agent to estimate the virtual reward r and arrival state s', and then apply one iteration of the *Value Iteration* algorithm (R. Sutton and A. Barto, 1998) to update the Q-value of (s,a), where k are all the possible actions starting from the arriving state s':

$$Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} T(s,a,s') max_{k \in \mathcal{A}} Q(s',k) \tag{4.4}$$

From Equation 4.4, we can compute the new Δ for the couple (s,a) and reinsert it within the memory buffer with Δ as the new priority level. Finally, we use the world-model to find all possible predecessors of (s,a), *i.e.,* couples (s'',a''), which according to the model, enable the agent to reach state s. Because the predecessors of a given state s can be challenging to determine in a stochastic world, Moore and Atkeson, 1993 propose to consider as predecessors all the states s'' which have, at least once in the history of the agent in the current task, performed a one-step

transition $s'' \rightarrow s$. The priority associated with a predecessor s'' can thus be the corresponding absolute prediction error Δ_{pred} and determines in which position it will be inserted in the memory buffer, as introduced by Peng and Williams, 1993. The replay phase then continues by processing the next element in the buffer with the highest priority level, and so on, until one of the stop conditions described above is met. For the sake of terminological clarification, what we call here a replay phase for an MB algorithm corresponds to an inference phase. This is because *MB-RL prioritized sweeping* does not replay memorized past experience but rather generates Simulation Reactivations (SimR) through model sampling combined with the Value Iteration algorithm described above. Thus, to transpose from MF to MB the replay phase stop conditions described above, the size of the replay budget N (which could also be called an inference budget in the case of MB) represents here a maximum number of iterations that can be inserted in the prioritized memory buffer and replayed through the Value Iteration algorithm.

Results

With the two changes that we made here compared to Cazé et al. (2018) and Khamassi and Girard (2020) (*i.e.*, (1) only allowing the simulated robot to do replay at the end of the trial when reaching the departure state, and (2) storing distinct (state, action) couple in the memory buffer for *MB-RL prioritized sweeping* rather than a single element per state), we found consistent performance results and only a difference in terms of a reduced computational cost during replay phases, which we describe below.

Figure 4.1b shows that the three algorithms with replay (*i.e.*, *MF-RL backward replay*, *MF-RL shuffled replay*, and *MB-RL prioritized sweeping*) quickly reached the optimal reward rate of 1 at the beginning of learning and then experienced only a brief drop in reward rate after the change in reward location at trial #100. In contrast, *MF-RL without replay* took longer to reach the optimal rate (approx. 60 trials) and barely re-increase its reward rate within 100 trials after the change in reward location. So the first conclusion is that any replay technique is equally valuable for enabling fast learning in such a simple maze task with predefined state decomposition.

The second interesting observation has to do with the transient and nearly discrete increases in replay time that are produced in responses to task changes (Figure 4.1b). All replay techniques enable the agent to avoid spending time performing replay during most of the task. Moreover, they all show a sharp increase in replay time after a change in reward location. Importantly, this property was also confirmed in our previous work, where replay was not restricted to the trial's end but allowed in any state of the task (Cazé et al., 2018). Thus it is interesting to note that such a way to generate replay events is not only compatible with neurobiological data (Cazé et al., 2018; Mattar and Daw, 2018) but also shows properties that could be useful for autonomous robots: bursts of replay could be used by the robot as a way to detect new task conditions automatically (but here the robot does not need to label these events explicitly; it just needs to adapt and maximize reward). The rest of the time, the agent starts each new trial without pausing, as if not showing any hesitation, similar to what is classically observed in well-trained rats in similar tasks (Gupta et al., 2010).

In addition, it is interesting to compare the duration of replay phases between the different replay techniques. While there is no difference in the average number of replay iterations after the change in reward location at trial #100 (Figure 4.1b), *MB-RL prioritized sweeping* performs drastically fewer replay iterations than *MF-RL*

backward replay and *MF-RL shuffled replay* during the initial learning phase (first 5-10 trials of the task). Now that we restricted these algorithms to perform replay only at the end of each trial, rather than after each action during the trial, *MB-RL prioritized sweeping* performs even fewer replay iterations than what we previously obtained in the same task (Cazé et al., 2018), without affecting its reward rate. The new proposal to restrict replay to the inter-trial interval thus seems promising for real robots. In Dromnelle, Renaudo, et al. (2020) (where we had not implemented any replay mechanism yet), the robot took a few seconds after each trial to return to the departure state. This short moment seems ideal to let the algorithm perform replay without affecting the robot's performance during the trial.

In the next section, we keep these principles and compare the same replay algorithms in a more open environment where the robot autonomously learns to decompose the task into discrete states, to verify that these algorithms still perform well under these more realistic conditions.

4.1.3 Simulation of individual replay strategies with an autonomously learned state decomposition

The neural activity of hippocampal place cells is often observed as showing transients and increases after surprising events (Valenti, Mikus, and Klausberger, 2018). During maze navigation, surprising events mainly occur at locations in the environment that are associated with positive or negative outcomes. From these locations, reverse replay, in particular, could reinforce spatial learning by occurring during awake periods after the spatial experiences (Foster and M. A. Wilson, 2006). They can potentially reinforce the surprising experience by propagating the outcome of the event to states that the animal has encountered on its way to the reward or punishment site. Moreover, rewarding states are also very likely to initiate replay activity in the hippocampus, to enhance the memory consolidation of novel information (Michon et al., 2019). During these events, the reactivation of the hippocampus neural activity is thought to be initiated by rewarding outcomes to bind this unexpected positive experience to the events that preceded it (Singer and Frank, 2009).

One of the first and most relevant experimental protocols to study these and other phenomena related to spatial navigation learning in rodents is the *Morris Water Maze* (Morris, 1981). In this work, rats were introduced to a circular pool with opaque water and removed only after reaching a hidden platform located just below the water's surface. Even though the rats could not see the platform, they could still localize it spatially. This was found even in cases where their starting point changed within the pool, thus indicating a robust spatial memory.

In this section, the same MF-RL and MB-RL replay strategies (Memory Reactivations (MemR) and Simulation Reactivations (SimR), Sect. 4.1.2) are tested in a more realistic robotic set-up, where the discretization of the environment in multiple Markovian states is autonomously performed by the robot [2]. Similarly to the experiment in Section 4.1.2 and to what has been experimentally observed by Foster and M. A. Wilson (2006), the replay phase takes place once the agent has reached the reward state to enable offline learning of Q-values, as previously done by Mattar and Daw, 2018. Neurobiologically, even though *Vicarious Trial and Error (VTE)* plays an essential role in animals' reasoning and decision-making (Tolman, 1939; Redish, 2016), it usually happens in uncertain moments, such as at beginning of the experiment, at decision points or surprising spots (Cazé et al., 2018; Khamassi and

[2]The code for these simulations is available at https://github.com/esther-poniatowski/Massi2022

Girard, 2020) and can also be unconsciously constrained by the attempt to limit the opportunity cost (Keramati, Dezfouli, and Piray, 2011).

This aspect is particularly crucial for the robotic experiment because it allows the agent to spend the Inter Trial Interval (ITI) updating the Q-table based on replay of its past experience. Usually, this time interval does not require expensive computations for the robot, since it does not need to make any decision on its way back to the starting position, and by replaying past experience, the learning speed could be enhanced without losing important experimental time.

The research question addressed whether MF-RL or MB-RL replay strategies could enhance spatial learning for artificial agents and robots. We found it interesting to first test our proposed algorithm in a simulated version of an experimental task (Morris, 1981) and eventually investigate if there were any differences between replaying reverse sequences of actions, random transitions, or the most surprising transitions, similar to what has been done in Section 4.1.2.

Like in the previous section, the presented simulated experiment investigates the role of diverse replay strategies relative to a changing reward condition. Moreover, the aim is also to investigate whether replays are relevant when transitions between the states of the task are stochastic. These simulations thus bring us to more realistic robotic experiments in stochastic and dynamical environments.

Materials and Methods

To study the implications of offline learning in spatial navigation, from rodent behavior to robotics, we have first investigated the role of two MF- and one MB-RL replay techniques (as in Sec. 4.1.2) in a circular maze, consistent with the original Morris water maze task (Morris, 1981) in terms of environment/robot size ratio. Then, the learning performances of the analyzed replay techniques are discussed in two main conditions:

- A deterministic version of the task, where an action a performed in a state s will always lead the robot to the same arrival state s' with probability 1.
- A stochastic version of the task, where performing action a in state s is associated with non-null probabilities of arriving in more than one state.

Learning algorithm and replay

As in the previous series of simulations (Sec. 4.1.2), here the simulated agent is learning using either classical *MF-RL Q-learning* (Watkins (1989), Eq. 4.1) or *MB-RL prioritized sweeping learning* (Moore and Atkeson, 1993; Peng and Williams, 1993). The values of their parameters (learning rate α and the discount factor γ) are shown in Tab. 4.2.

The first implementation of offline learning techniques that we tested is the *MF backward replay*. Similar to the double T-maze experiment in Sect. 4.1.2, the offline learning phase happens once the agent has reached the reward state, which indicates the end of a trial. All along with the trial, the Q-values $Q(s,a)$ of the state-action couple (s,a) are updated with Eq. 4.1, and once the rewarding state has been reached, they are updated again in reverse order, starting from the reward state. These backward sequences can be up to N updates long if the agent has gained enough past experience and stored it in its memory buffer. The reverse sequences are then replayed until the sum of variations of Q-values over the last replay repetition is below a certain replay threshold ϵ (Tab. 4.2). Given the size of the environment (36 states), these

N long backward replay sequences can also involve experiences that happened during the previous trials of the same agent (*i.e.*, during the previous attempts to get to the reward). In this way, the robot can transfer the acquired knowledge through different trials and learn more efficiently.

The second replay strategy that has been tested is *MF shuffled replay*; in this case, in the ITI, the internal values $Q(s,a)$, are randomly ordered and then updated by Eq. 4.1. As for the *MF backward replay*, the memory buffer that is accessible to initiate the reactivations keeps in memory the latest N transitions (Tab. 4.2). Also, in this case, the agent can benefit from the experience acquired during the latest trials and learn to extract more general and valuable knowledge from its recent and uncorrelated past actions (because of shuffling). The ITI replay phase lasts until the sum of the Q-values is converged under an ϵ value given in Tab. 4.2.

As for Sect. 4.1.2, we compared the learning performance of the above-explained MF replay strategies to an MB prioritized sweeping algorithm (Moore and Atkeson, 1993; Peng and Williams, 1993). The implementation of the latter is the same as described in Sect. 4.1.2, and the convergence criterion is reached when the prioritized replay buffer, which can be maximum N transitions long, is empty. **The experimental set-up and implementation**

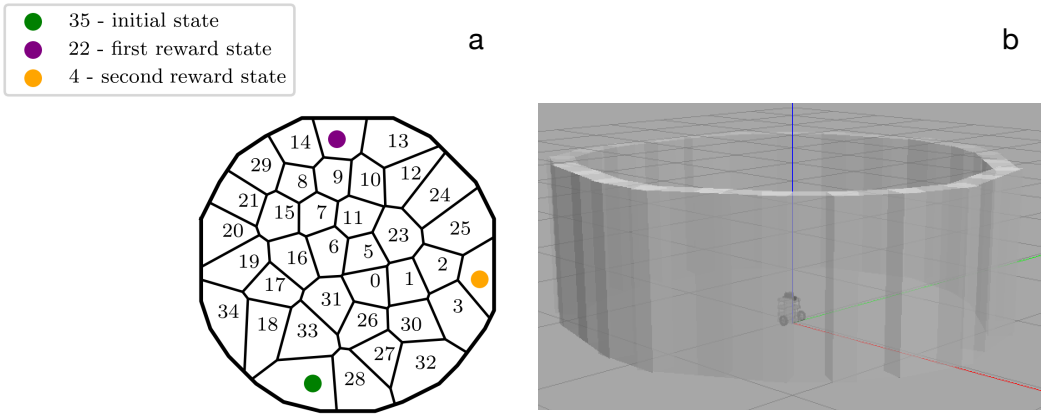


FIGURE 4.2: Description of the experimental set-up. A) map of the discrete states of the maze, identified by the robot during the exploration on Gazebo. The initial state and the two rewarding states are also highlighted. B) the ROS Gazebo simulated Turtlebot 3 in the center of the circular environment.

The simulated experimental set-up intends to replicate a Morris water maze task: the agent is introduced in a new circular environment and has to learn how to reach a particular location associated with a positive reward (Morris, 1981). In our set-up, the agent is a Turtlebot3 Burger, simulated with the Robot Operating System (ROS) middleware and the Gazebo simulation environment (Quigley et al., 2009). The water maze is represented as an empty circular arena surrounded by high walls (Fig. 4.2b).

The robot discovers and defines the different discretized areas in the maze by autonomously navigating within the environment. Despite the odometry and the laser sensor being installed on the robotic device, the acquired space representation is allocentric. This is an emergent property of the automatic clustering process

when applied to robot sensor data in a task where the robot can only move in a horizontal plane, as found in previous neurorobotics work (Caluwaerts et al., 2012). The robot, in fact, explores by selecting between 8 directions of motion defined in the environment's global reference frame, and its current position and orientation are also elaborated in the maze reference frame. This allocentric description of the robot movements and the states of the maze is possible thanks to a re-mapping of the relative position of the robotic agent and the discretized states to the reference coordinate system of the map. This is possible thanks to the 360 Laser Distance Sensor of the robotic platform, combined with the use of a classical SLAM technique. Note that such an allocentric space representation is compatible with neurophysiology (hippocampal place cell activity) and can also be combined with egocentric representations to account for a variety of experimentally observed animal behaviors during navigation tasks (Khamassi and Humphries, 2012). The discrete MDP, presented in Fig. 4.2a, is obtained thanks to a Rao-Blackwellized particle filter that builds grid maps from laser range data (Grisetti, Stachniss, and Burgard, 2007). The simulated implementation of this Simultaneous Location and Mapping Algorithm (SLAM) on ROS Gazebo is called *GMapping*.

This state decomposition process makes the robot able to immediately create new states if necessary. However, in our work, the aim was to create the finest and most robust possible discretization of the maze to be then employed in all the simulation experiments where we tested the different replay strategies. As observed by Khamassi, 2007; Chaudhuri et al., 2019; Benchenane et al., 2010, rats could re-explore the whole maze every day before doing a learning task, and that could reflect their need to rapidly acquire and stabilize a state representation before starting an extra learning process.

For these reasons, the robot performs a long autonomous exploration phase to acquire its state representation before starting the learning phase. During the first 48-minute-long exploration in Gazebo, the SLAM algorithm estimates the current robot coordinates. Whenever the robot is more than 15 cm away from any existing state, the algorithm creates a new state, whose reference position is the current. This results in a Voronoi partition of the space, composed here of 36 states (Fig. 4.2a). This 15 cm state radius was chosen to be similar to the robot footprint of 13,8 x 17,8 x 19,2 (L x W x H, cm). The action space $\mathcal{A}$ instead contains 8 homogeneously distributed directions of motion, defined with respect to the world reference frame (same as for Sect. 4.1.4, Fig. 4.8a top right).

Then we ran another free exploration of the arena by the simulated Turtlebot3 robot to automatically learn the transition probabilities $p(s'|s,a)$ that can be approximated from randomly executing different actions a in different states s, and observing the arrival state s'. This second free exploration phase was chosen to be 5357-action long, the same duration as for the results that will be presented in Sect. 4.1.4. Lesaint et al., 2014 found that when an agent was progressively learning its transition function during the task, the RL model was better at accounting for rat behavior than a model with a prior given transition function.

In practice, the transition probabilities autonomously learned by the robot during free exploration in Gazebo are stochastic: the same action a performed in the same state s can lead to more than one state with non-null probabilities. For instance, moving north from state #31 alternatively leads to states #0,5,6,16 and even sometimes to state #31 itself when the robot initiated its movement from the bottom part of this state (Fig. 4.2a). Such stochasticity results from several properties: (1) because the states autonomously decomposed by the algorithm are not evenly

distributed; (2) because the experiments are performed in a simulated physical environment, which includes frictions between the robot's wheels and the floor, and where the robot sometimes moves too close to the walls, thus triggering its obstacle-avoidance process, hence resulting in a different effect of the same action performed without obstacle-avoidance.

The actual level of uncertainty of the stochastic version of the task is displayed in Fig. 4.10a, where each state s has an entropy $H_{env}(s)$ computed as in the equation 4.5, where $\mathcal{A}$ is the set of all the possible actions a from state s, s' are all the possible arrival states from the original state s and $p(s'|s,a)$ is the probability that the agent arrives in state s' after starting from state s and performing action a:

$$H_{env}(s) = \max_{a \in \mathcal{A}} \sum_{s'} -p(s'|s,a) \log_2 p(s'|s,a) \tag{4.5}$$

Finally, in order to obtain a deterministic version of the same task from these autonomously learned transition probabilities $p(s'|s,a)$, for each (state,action) couple (s,a), we search for the state s' with the highest probability of arrival (*i.e.*, $s' = \underset{x \in S}{\text{argmax}}[p(x|s,a)]$), and set $p(s'|s,a) = 1$ while setting $p(s''|s,a) = 0$ for all other states $s''(s'' \neq s')$. The deterministic version of the task consists, in fact, in the simplification of the interaction between the robot and the environment, meaning that the trajectories that the robot can cover in the same environment are reduced. To quantify the simplification of the resulting MDP, we have performed an analysis of the trajectories which have been taken by the four different algorithms in the two different environments. We compute the pairwise Fréchet distance of these trajectories to the optimal one, found by following a greedy optimal policy. Fig. 4.3 shows this analysis during the first half of the experiment when the reward is fixed in state #22. The results from this analysis show that, for all the adopted strategies, in the stochastic environment (Fig. 4.3b), the sparsity of the trajectories around the optimal path is generally higher compared to the same deterministic case Fig. 4.3a). To assess the difference among these distance distributions, we did a Kruskal-Wallis H-test (Kruskal and Wallis, 1952), founding them significantly different from each their corresponding distribution in the other environment. The conversion of the environment in a deterministic MDP is then intrinsically limiting the level of exploration of the agents, resulting in two very different scenarios. However, it is crucial to investigate this transition, given our intent to study the role of RL replay strategies in robotic navigation, from a theoretical to a more realistic robotic outline.

To replicate a non-stationary task similar to the one in the original experiment (Morris, 1981), we changed the reward location from state #22 to state #4 at trial 25. We tested the learning performances of the agent with four different replay strategies (no replay, MF backward replay, MF shuffled replay and MB prioritized sweeping) and in two different environments: a deterministic and a stochastic version of the task.

Results

To assess the actual contribution of the tested replay strategies to the learning process of the described spatial navigation task, an unbiased learning rate α_{best} has to be found. Since α_{best} could be different depending on the unpredictability of the MDP which simulates the task (*i.e.*, deterministic or stochastic), we simulated 100 robotic agents performing 50 trials to get to the rewarding states, for a set of uniformly distributed α values between 0 and 1 (Fig. 4.4). For each value of α, we looked at the

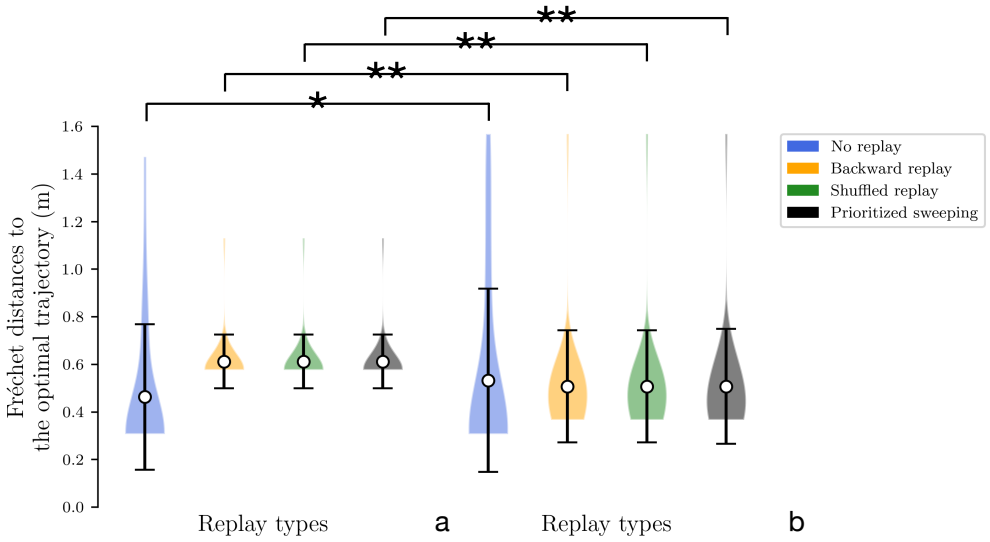


FIGURE 4.3: Analysis to investigate the level of the sparsity of the explored trajectories by the agent. The Fréchet distance has been computed for the first half of the simulation. ** stands for p-value lower than 0.001 and * for p-value lower than 0.05. A) The extension of the Fréchet distance to the optimal trajectory in the deterministic case for all the algorithms. B) The same extension of Fréchet distance in the stochastic environment.

average value $\overline{action(\alpha)}$ along the trials, with $action(\alpha)$ being the number of actions needed by the robot to get to the rewarding states. This value is computed for both the deterministic (Fig. 4.4a) and the stochastic worlds, considering the entirety of the experiment, and the minimization of the sum of these two values is used to identify the final α_{best} (Fig. 4.4b and Tab. 4.2) as described in the equation below:

$$\alpha_{best} = \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}}(\overline{action_{deterministic}(\alpha)} + \overline{action_{stochastic}(\alpha)}) \tag{4.6}$$

where $\mathcal{A}$ is the set of tested α values.

Once identified the most appropriate value for the learning rate α, the following four replay conditions have been tested in the task:

- *MF-RL no replay*
- *MF-RL backward replay*
- *MF-RL shuffled replay*
- *MB-RL prioritized sweeping*

and the other relevant parameters for the experiment are described in Tab. 4.2.

The main results are shown in Fig. 4.5. The four different RL algorithms (no replay, backward replay, shuffled replay, and prioritized sweeping) are compared in terms of the number of model iterations to get to the rewarding state (Napierian logarithm of the first, median, and third percentiles over the behavior of 100 robotic agents). The task changes at trial #25 when the reward switches from state #22 to state #4 (Fig. 4.2).

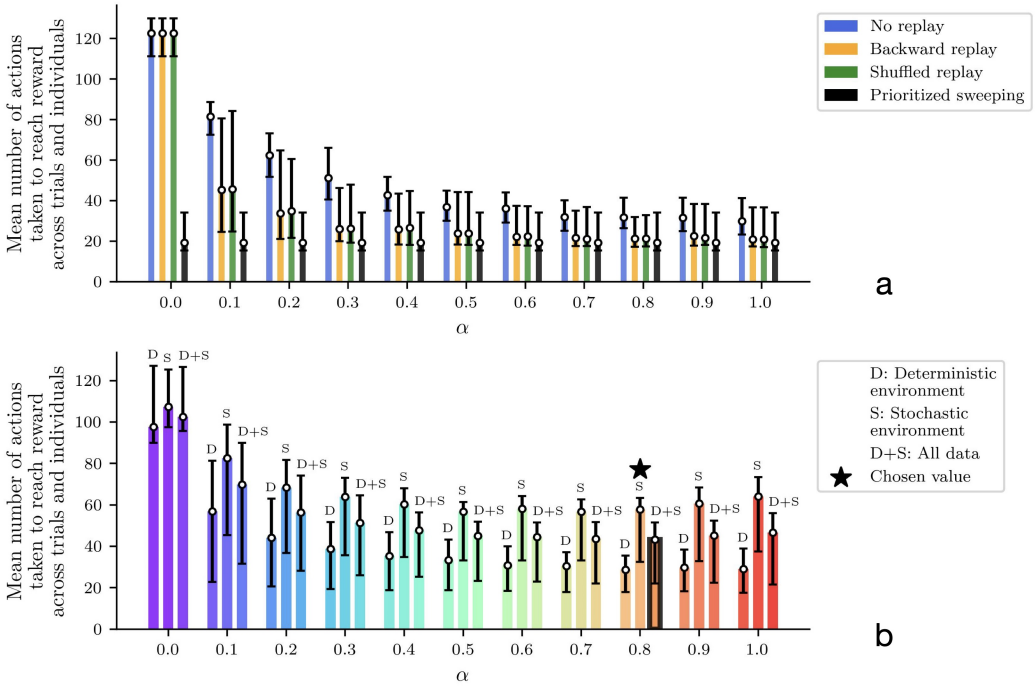


FIGURE 4.4: Performed analysis to find out the best learning rate α for all the replay strategies and the two environments (deterministic and stochastic). For different values of α, the figure shows the first, median, and third percentile of the number of actions to get to the reward, over 100 agents completing the simulated experiment over 50 trials. The average minimum number of model iterations to get to the reward is found for α equal to 0.8, and it was used for all the presented experiments (Tab. 4.2). A) Performances of the tested algorithms across the α values in the deterministic version of the maze. B) Final selection of α considering the mean performances between the deterministic and the stochastic version of the maze.

When the task is deterministic (Fig. 4.5a), all the three RL algorithms with replay learn a short path to the reward significantly faster than the *MF-RL no replay* learner (Fig. 4.5a and b, Trials 1-5). The same situation occurs when the reward position is switched at trial #25, assessing RL replays' role in improving the learning speed after such a task change (Fig. 4.5a and b, Trials 26-30). When the environment is stochastic, the situation is similar and, in particular, the prioritized sweeping algorithm is learning significantly faster than the other replay strategies (Fig. 4.5b, Trials 26-30) reflecting the importance of an MB strategy (with MB replay) to faster adapt to dynamical tasks, when the transition model is not deterministic. This suggests that moving towards more complex robotic tasks, MB-RL models of replay may be preferred since the higher information processing regarding the model of the environment at the beginning of the task can save real experimental time when the robot would need to adapt later in the experiment.

Moreover, the logarithmic scale makes it easier to notice that the no replay agent, even if it is slower at the beginning of the task, can converge to paths that are significantly shorter than the one covered by the other strategies before the change in reward location (Fig. 4.5a and b, Trials 20-25). In the stochastic environment, in

	No replay	MF backward replay	MF shuffled replay	MB prioritized sweeping
α	0.8	0.8	0.8	0.8
γ	0.9	0.9	0.9	0.9
β	15	15	15	15
ϵ	-	0.001	0.001	0.001
N	-	90	90	90

TABLE 4.2: Algorithm parameters used to generate the results in this section. α is the learning rate, optimized as shown in Fig. 4.4 and Eq. 4.6 and γ is the discount factor. β is the inverse temperature in the softmax function for decision-making (Equation 4.2), and its values were found by optimizing both the convergence time and the performance of the tested algorithms. N is the maximal length of the episodic memory buffer. This value was selected to replay the entire real experience during the first trials of the experiment and to replay experiences from several past trials later in the simulation. Finally, ϵ is the convergence threshold as for Sect. 4.1.2 and Cazé et al., 2018.

particular, the *MB-RL prioritized sweeping algorithm* reinforces the experience of a sub-optimal path, resulting in performance significantly different from the ones obtained from the other two replay strategies (Fig. 4.5b, Trials 20-25). This shows that, even if the stochastic environment leads the MF-RL replay strategy to explore the maze more, the *MB-RL prioritized sweeping algorithm*, that can learn the transition model from the beginning of the task, is not subjected to this "push" towards exploration and keeps reinforcing the shortest path previously found.

Instead, in the second convergence phase (Trials 45-50), we highlight the fact that the no replay agent is not showing anymore statistically better performances than all the replay algorithms (Fig. 4.5a and b, Trials 45-50). In the deterministic case, it still reaches the shortest path to the reward. However, the prioritized sweeping agent is also significantly better than the *MF-RL shuffled replay* strategy (Fig. 4.5a Trials 45-50). On the other hand, in the stochastic case, the *MB-RL prioritized sweeping*'s knowledge of the environment makes it attain performances that are compatible with the ones from the no replay strategy. In this case, we can notice that the replay strategies perform differently, with the shuffled replay, which performs worse than the other two replay strategies. This re-adaptation phase gives the agents the opportunity for more exploration, particularly the replay agents, which have strongly reinforced their previously experienced trajectory to maximize the reward and propagate this knowledge throughout the environment. As already happened in the second learning phase (Fig. 4.5b, Trials 26-30), the *MB-RL prioritized sweeping algorithm* significantly exceeds the performance of the other replay algorithms and converges to a shorter path to the reward. This gives insights into the need for a more consolidated knowledge of the environment (and so of the agent's interaction with it) for adaptive tasks. Consequently, we can predict that animals would need to retrieve knowledge about their experienced and learned model of the world to adapt more efficiently to dynamic circumstances.

Following the results shown in Fig. 4.5, we have further investigated the learning and replay dynamics of the proposed strategies. In Fig. 4.6, the level of propagation of the Q-values (Eq. 4.1) over the environment is shown for the different tested RL

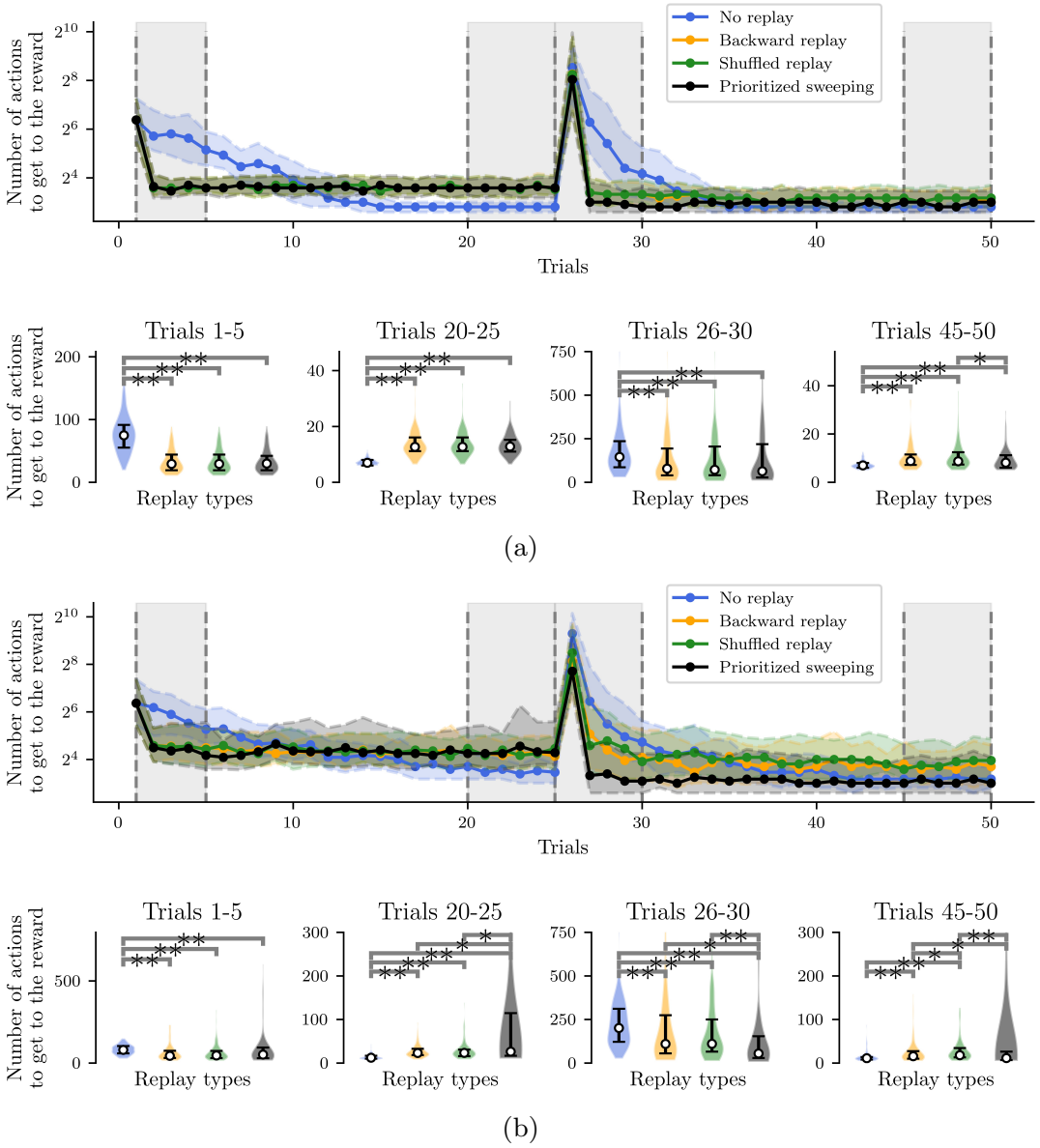


FIGURE 4.5: Performances of the simulated robot, learning the non-stationary task, and a post hoc Wilcoxon-Mann-Whitney pairwise comparison test on the relevant trial intervals among the different curves. The post hoc test has been performed following a Kruskal-Wallis H-test (Kruskal and Wallis, 1952) to reject the null hypothesis that the population median of all of the algorithms' average performances was equal. ** stands for p-value lower than 0.001 and * for p-value lower than 0.05. A) Deterministic environment. B) Stochastic environment.

algorithms and for both the deterministic (Fig. 4.6a) and the stochastic (Fig. 4.6b) environments. The shown learning dynamics are representative of the different strategies since they show the individual's behavior, which is the closest to the median performances of all the 100 individuals for each strategy.

In both cases (Fig. 4.6a and 4.6b, Trials 1,2 and 25), the presence of replay provides a drastically more extensive propagation of the Q-values, starting from the first reward state (22). This explains the significantly faster learning performances

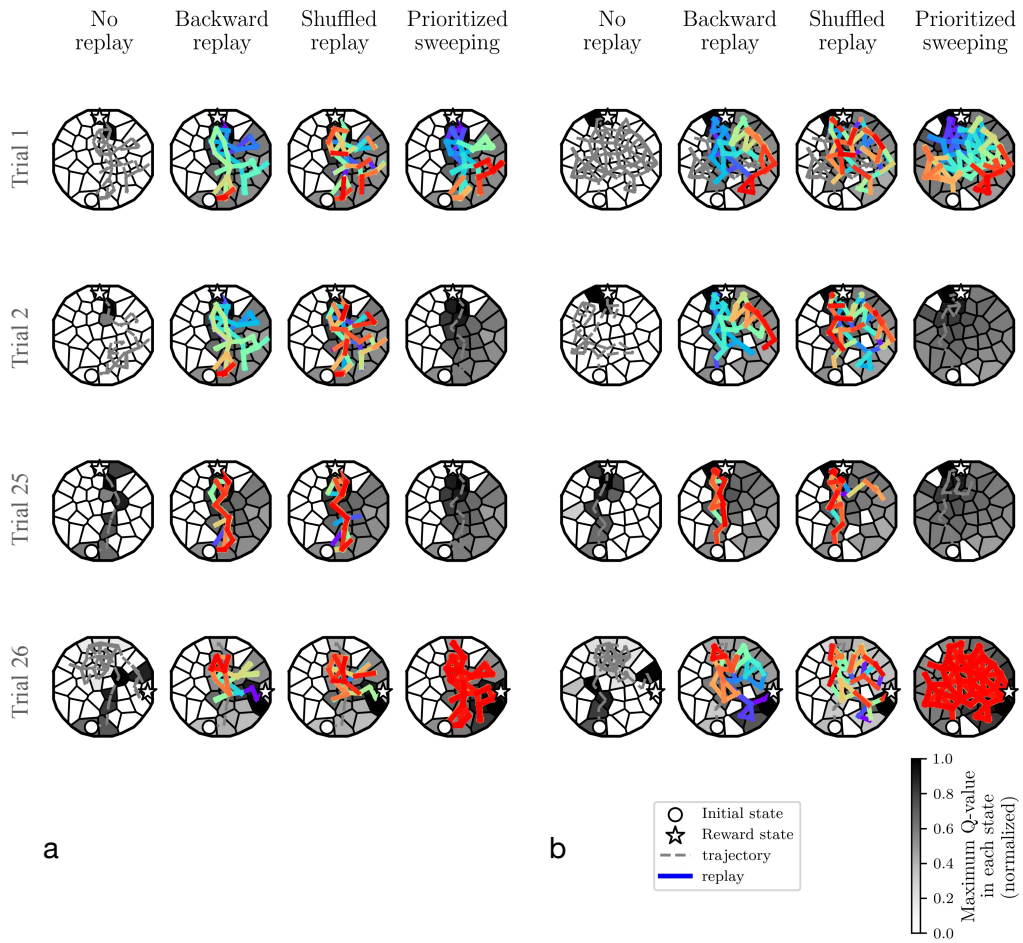


FIGURE 4.6: Learning dynamics of the most representative individual: covered trajectory and replay at some critical trials. Also, for each state s, the $maxQ(s, a_i)$, among all the a_i, with i from 1 to 8 (Fig. 4.8a, top right), is represented. The initial state and the reward state are also represented in the figure. A) Experiments in the deterministic MDP. B) Experiments in the stochastic MDP.

observed in the algorithms with replay compared to the *MF-RL no replay method*. In both environments, the no replay method is slower to learn. However, it explores more in the first trials (Trial 1 and 2), and that leads it to generally find a shorter path to the reward location in the end (Trial 25) compared to the other learning strategies (as shown in Fig. 4.5a and, Trials 20-25).

Comparing the two types of environments, we can understand that the MDP's stochasticity level leads to a larger exploration of the environment for all the strategies (Fig. 4.6b, looking at the explored trajectories and the replayed transitions). This results in a more extensive propagation of the maze Q-values, particularly in the prioritized sweeping algorithm. As in the deterministic case, the *MB-RL prioritized sweeping* is replaying a broader range of transitions after the first trial compared to the other strategies. With this MB-RL replay strategy, the replay activity is led by the surprise of the experienced events, resulting in longer replay phases, happening just at specific moments in the task (some of them are well visible in Fig. 4.6, Trial 1 and

26). This happens in both environments also thanks to the implementation of the algorithm, which also examine the predecessor of the surprising state (Sect. 4.1.3) and to the acquired knowledge of the environment (in particular in Trials 26, when the reward position changes). In both environments, as expected from the previously analyzed learning performance in Fig. 4.5, there is no practical difference in terms of Q-value propagation between *MF-RL backward replay* and *MF-RL shuffled replay*. Furthermore, the explored trajectories and the replay are also very similar, resulting in not significantly different performances (Fig. 4.5).

These results, which simulate a spatial learning experiment for rodents (Morris, 1981) in a robotic framework, suggest some first advantageous properties of using replay-inspired strategies in neurorobotics. Our results imply that MF-RL replays could be sufficient to speed up learning and adaptation to non-stationarity (Fig. 4.5, Trials 1-5 and 26-30), but MB-RL replay strategies could improve the adaptability of the system even more, with a higher level of stochasticity which often characterizes real robotic scenarios (Fig. 4.5, Trials 26-30). The proposed models and experiments contribute to a deeper understanding of the advantages and limitations of the existing RL replay models in such robotic tasks. This experimental comparison, examining either a deterministic or stochastic version of the same environment (which implies a significantly different level of explored trajectories in the maze, see Fig. 4.3) was helpful to observe that RL replay gives an important contribution to a robotic spatial learning task, even if the model of the interaction robot-environment is stochastic. Nevertheless, a good compromise between the exploration capability of MF replay strategies and the adaptability of MB ones has not yet been found within these experiments.

The following section will illustrate the performances of RL replay strategies in spatial learning when they are tested in combination in an MF-MB RL hybrid learning architecture in a more complex environment with obstacles, higher stochasticity, and non-stationarity.

4.1.4 Combining model-based and model-free replay in a changing environment

Hippocampal replay has not only been interpreted as a memory consolidation process from past experience (Foster and M. A. Wilson, 2006; Girardeau et al., 2009), putatively model-free but also as a possible model-based planning process that enables the mental simulation of hypothetical actions (Gupta et al., 2010; Ólafsdóttir, Bush, and Barry, 2018; Khamassi and Girard, 2020). Along these lines, it has been argued that model sampling can not only be used for planning but also to update action values (Seijen and R. Sutton, 2015; Cazé et al., 2018; Mattar and Daw, 2018). Moreover, some sequences of reactivated hippocampal neurons cannot be accounted for as a simple model-free reactivation of past experience and rather seem to represent creative combinations of past and experienced trajectories, which can only be accounted for by a model-based process (Gupta et al., 2010).

This suggests that both model-free *Memory Reactivations (MemR)* and model-based *Simulation Reactivations (SimR)* are required to account for the diversity of hippocampal replays. Importantly, state-of-the-art models of reinforcement learning processes in the mammalian brain assume a co-existence of model-based and model-free processes (Daw, Niv, and Dayan, 2005; Dollé et al., 2010; Keramati, Dezfouli, and Piray, 2011; Khamassi and Humphries, 2012; Pezzulo, Rigoli, and Chersi, 2013; Dollé et al., 2018; A. G. Collins and Cockburn, 2020). Hence, neurorobotics

constitutes a promising research area to study replay in robot control architectures that combine MB and MF reinforcement learning processes.

The experiments presented in the previous sections have analyzed the complementary properties and performances of MF replay and MB replay. In our presented tasks, RL agents with MB replays tended to be slower to converge to an optimal solution but eventually reached a faster path to the reward location. On the other hand, the same agent with MF replay learned faster but converged to a suboptimal solution. In this section, in addition to pushing robot simulations towards more complex environments with stochasticity and non-stationarity, we want to examine the benefits of combining Simulation Reactivations (SimR, similar to Pezzulo, Rigoli, and Chersi (2013) and Keramati et al. (2016), but unordered) and Memory Reactivations (MemR) in a robot control architecture which includes both MB and MF RL [3]. We thus investigate the effects of including replay in the algorithm proposed in Dromnelle, Renaudo, et al. (2020), which coordinates a Model-based and a Model-free RL expert within the decision layer of a robot control architecture. Interestingly, this algorithm had been previously tested in a navigation environment that includes open areas, corridors, dead-ends, a non-stationary task with changes in reward location, and a stochastic transition function between states of the task. In these conditions, previous results showed that the combination of MB- and MF-RL enables the robot to (1) adapt faster to task changes thanks to the MB expert and (2) avoid the high computational cost of planning when the MF expert has been sufficiently trained by observation of MB decisions (Dromnelle, Renaudo, et al., 2020). Nevertheless, replay processes have not been included in this architecture yet, and the present paper is the opportunity to do it.

The results that we are going to illustrate and discuss in the following subsections present the combination of SimR and MemR as a critical resource to optimize the trade-off between the increase in performance and the reduction of computational cost in a hybrid MB-MF RL architecture when solving a more complex non-stationary navigation task than the two previous sections.

Materials and Methods

The robot control architecture proposed in Dromnelle, Renaudo, et al., 2020, and also successfully applied to a simulated human-robot interaction task in Dromnelle, Girard, et al., 2020, takes inspiration from the mammalian brain's ability to coordinate multiple neural learning systems. Such ability is indeed considered to be vital in making animals able to show flexible behavior in various situations, to adapt to changes in the environment while simultaneously minimizing computational cost and physical energy (Renaudo et al., 2014). The proposed architecture in Fig. 4.7 is composed of a decision layer where a model-free (MF) expert and a model-based (MB) expert compete to determine the next action of the system. Both experts pass through three phases: learning, inference, and decision. Finally, a meta-controller (MC) determines which proposed decision will be executed, following an arbitration criterion described below.

Model-based expert

The MB algorithm is implemented to learn a transition model T and a reward model R of the specific task. Thanks to these two learned models, it can predict the consequences of a given action several steps ahead and adapt faster to non-stationary

[3] The code for these simulations is available at https://github.com/elimas9/combining_MB_MF_replay

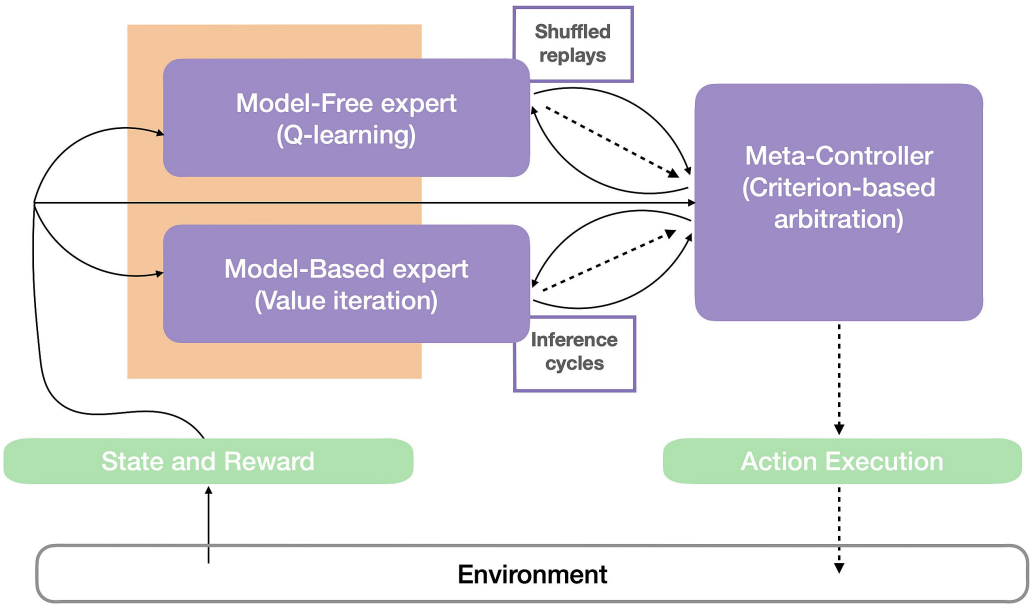


FIGURE 4.7: Robot control architecture. The agent-environment interaction can be described by (1) the state and the reward as perceptual information (continuous arrows) from the environment and (2) by the action (dashed lines) that the agent operates in the environment. The perceptual information is used by the Model-Free, the Model-Based expert, and the Meta-Controller (in purple). Based on this information and memory of their previous performances, the Meta-Controller estimates the entropy and computational cost of the experts, consistently with the criterion in Eq. 4.13, and thus chooses the expert that will be allowed to infer the probability distribution of the next agent's actions. This distribution, and the times consumed to compute it (dashed arrows), are then sent to the Meta-Controller. Differently from Dromnelle, Renaudo, et al. (2020), both experts here have a 'replay' (reactivation) budget (limited or until convergence) that will affect both their performance and computation time and thus impact the Meta-Controller's arbitration. Here, shuffled Memory Reactivations (MemR) are integrated with the Q-learning algorithm of the MF expert, while Simulation Reactivations (SimR) constitute the offline MB inference iterations in the Value Iteration algorithm of the MB expert.

environments. Yet these computations are very costly (*i.e.*, 1000 times higher than the computations of the MF expert in Dromnelle, Renaudo, et al. (2020)).

During the *learning process*, the transition model and the reward model are updated at each timestep after observing the departure state s of the robot, the action a that it has performed, the arrival state s', and the scalar reward r that this transition may have yielded. The transition model is updated by estimating $T(s,a,s')$, the probability of arriving in s' from (s,a), considering the past T_{tw} actions (Tab. 4.3). This probability is computed as already shown in Eq. 4.3. Besides, the reward model $R(s,a,s')$ is updated by considering the most recent reward r_t associated with the transition (s,a,s'), multiplied by the probability of the transition itself in Eq. 4.3.

The *inference process* estimates the action-value function via the Value Iteration algorithm (R. Sutton and A. Barto, 1998), and it operates as an offline planning phase that is continuously called every decision step, just before a decision is made by the

agent about which action to perform. The maximal duration of this planning process can be determined either by setting a finite budget for the number of transitions over which the agent will evaluate its decision or by employing a convergence criterion based on the sum of the absolute action-value function estimation errors. More precisely, the planning terminates at iteration c if:

$$\sum_{s,a} \left|\delta^c_{s,a}\right| < \epsilon_{MB} \quad \text{where} \tag{4.7}$$

$$\delta^c_{s,a} = \sum_{s'} p(s'|s,a)[R^c_{s,a} + \gamma V(s')^c] - Q(s,a)^c \tag{4.8}$$

Here $R^c_{s,a}$ is the reward function of performing action a from state s at the offline reactivation c and $V(s')$ is the value function of the arriving state s' at reactivation c, from state s and action a. γ is the discount factor (Tab. 4.3).

Finally, the *decision process* chooses the next action to be performed by the robot by converting the action-value function into a probability distribution using a softmax function (see Eq. 4.2), with an exploration/exploitation trade-off parameter β given in Tab. 4.3.

Model-free expert

The MF algorithm does not learn any transition or reward model of the task, in contrast to the MB expert. Rather, it locally updates the current action-value function $Q(s,a)$ at each timestep. This property of the MF expert saves computational cost, compared to the MB expert, at the expense of slow adaptability to task changes, given the expert's lack of topological knowledge of the environment.

The *inference process* consists of reading from the Q-table the line corresponding to s, which is then used by the *decision process*. The latter chooses the next action from the Q-values, also converted to a probability distribution with a β trade-off parameter in Tab. 4.3.

For the MF-RL expert, the *learning process* is defined as a tabular Q-learning algorithm in which the action-value function $Q(s,a)$ is updated according to Eq. 4.1. Following the online learning phase, *shuffled replay* is performed, using the (s,a,s',r) tuples experienced by the agent in a given time-window of past transitions R_{tw} (Tab. 4.3). As for the MB expert, these offline updates stop when either the maximal predefined budget is exhausted or when the Q-values have converged. Since the MF expert does not know the transition probabilities of the task, a convergence test is computed for every offline learning iteration c as in Eq. 4.9, where $act^c_{s,a} = \tau \cdot act^{c-1}_{s,a}$, with $act^{\tilde{c}_{s,a}}_{s,a} = RB$ during the first time $\tilde{c}_{s,a}$ when that specific transition is selected for replay and with $act^0_{s,a} = 0$. act is an activation function defined for each couple (s,a), and it is 0 if (s,a) has not been replayed before or otherwise it decays from RB (Tab. 4.3) along the replay iterations c with a time constant τ (Eq. 4.11).

$$\sum_{s,a} \delta^c_{s,a} act^c_{s,a} < \epsilon_{MF} \quad \text{where} \tag{4.9}$$

$$\delta^c_{s,a} = |Q(s,a)^c - Q(s,a)^{c-1}| \tag{4.10}$$

The principle behind the design of this convergence criterion is that the importance of each $\delta_{s,a}$ (Eq. 4.10) starts as RB and decreases over the offline learning iterations c, following the decay constant τ (Eq. 4.11). This strategy does not constrain the number of needed replay iterations because the agent would still perform replays due to

high $\sum_{s,a} \delta^c_{s,a} act^c_{s,a}$. Nevertheless, this value will slowly decrease the need for more replay iterations along with the offline learning phase. RB is a value representing one of the possible replay budgets needed to obtain performances that are comparable to the maximum amount of reward that the expert can collect, thus not inhibiting the offline learning phase when needed. Finally, the convergence threshold ϵ_{MF} is an order of magnitude larger than ϵ_{MB} (Tab. 4.3. which is the same used in Dromnelle, Renaudo, et al. (2020)). The MF expert does know the probabilities contained in the transitions model in Eq. 4.3. For this reason, its convergence criterion is based on the actual update of the action-value function $Q(s,a)$. This means that, in the MF case, the $\delta^c_{s,a}$ are not multiplied by any probability derived from the world model. Thus their values will usually be an order of magnitude larger than the $\delta^c_{s,a}$ of the MB case, multiplied instead by the probability of a given (s,a,s',r) tuple.

$$\tau = \sqrt[RB]{\frac{\epsilon_{MF}}{RB}} \tag{4.11}$$

Meta-controller

The MC selects which expert will take control of the next action by following a specific criterion that is a trade-off between the learning performances and the computational cost of the inference process of the two agents, and it is called *Entropy and Cost (EC)* (Dromnelle, Renaudo, et al., 2020).

On the one hand, the quality of learning is computed by Eq. 4.12 where $f(P(a|s,E,t)$ is a low-pass filtered action probability distribution with a time constant $\tau = 0.67$, previously used as an indicator of the learning quality in humans (Viejo et al., 2015).

$$H_{exp}(s,E,t) = -\sum_{a=0}^{|\mathcal{A}|} f(P(a|s,E,t)) \cdot \log_2 \left(f(P(a|s,E,t))\right) \tag{4.12}$$

On the other hand, the cost of the process $C(s,E,t)$ is the computation time needed to perform the inference phase for the expert E, at time t, and it is also filtered as the action probability distribution above.

Eventually, the MC chooses which expert will take control of the next decision by following the equation below (Dromnelle, Renaudo, et al., 2020):

$$EX(s,E,t) = -(H_{exp}(s,E,t) + \kappa C(s,E,t)) \tag{4.13}$$

$EX(s,E,t)$ is the expertise value of the expert E, which is then converted into a distribution of probabilities using a softmax function. κ weights the impact of time in the criterion by assigning greater importance to the computation time when the entropy component $H_{exp}(s,E,t)$ of the MF experts is low.

After applying Eq. 4.13, the MC draws the winning expert from the softmax of the distribution of their expertise $EX(s,E,t)$ (with a trade-off coefficient β shown in Tab. 4.3 and inhibits the inference process of the expert that is not selected.

The experimental set-up and implementation

This new hybrid MB-MF RL architecture with replay is tested in a dynamic navigation task where the robot has to learn how to reach a unitary rewarding state. The task remains stationary during the first 1600 over 4000 iterations, and then the reward is moved to another state (from state 18 to state 34, Fig. 4.8). In this experiment, an extra element of non-stationarity is represented by the starting state of the robot being uniformly selected with the same probability between state 0 and state 32 at

the beginning of each trial (Fig. 4.10). Differently from Dromnelle, Renaudo, et al., 2020, experiments where the reward is fixed or where a new obstacle is introduced have not been performed for this work.

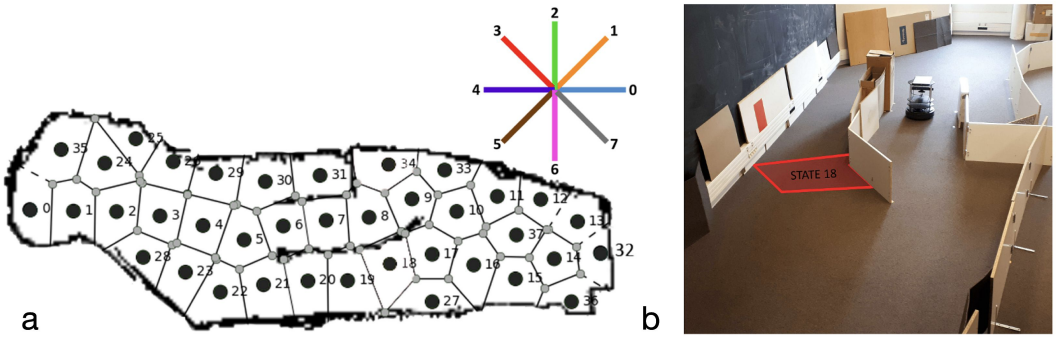


FIGURE 4.8: Description of the experimental set-up. A) Map of the discrete states of the maze. The eight-pointed star indicates the cardinal directions in which the robot can move. These directions are the same used for the experiment in Sect. 4.1.3. B) Photo of the real Turtlebot approaching the initial rewarding state 18, highlighted in the figure. Adapted from Dromnelle, Renaudo, et al. (2020)

First, the real Turtlebot autonomously navigates within the environment using a SLAM Gmapping algorithm (Fig. 4.8) and creates a discrete map of the maze (38 Markovian states are identified and shown in Fig. 4.8a). This autonomous state decomposition process is identical to the one used in the previous experiment described in Sect. 4.1.3. The robot-environment ratio is very similar to the previous experiment in Sect. 4.1.3: the state radius is 35 cm, in this case, and the robot size is 35,4 x 35,4 x 42 (L x W x H, cm).

Then, during a second free exploration phase, the robot learns the transition model of the environment, that is, the probability that the robot starts its move in one state s performs an action a, and arrives in another state s'. This second phase of the creation of the transition model is also conducted as in Sect. 4.1.3, but with the real robot.

After these exploration phases, the subsequent experiments involving a reward were performed in simulation to test the impact of different parameters of the algorithm and study the effect of replay on total performance and computation cost. During these simulations, the agent experienced the MDP based on the transition map that was empirically acquired with the real robot (as was done in (Dromnelle, Renaudo, et al., 2020)).

Fig. 4.10b shows the maximum level of uncertainty for each of the 38 states of the environment. This uncertainty is computed in the same way as for the other experiment in Eq. 4.5, and the transitions map is used to guide the robotic exploration in the simulation environment.

The action space is also discrete and consists of 8 possible cardinal directions equally distributed around the agent. Given the discrete and probabilistic nature of the state and action spaces, the transition model $T(s, a, s')$ (Eq. 4.3) and the reward model $R(s, a)$ of the MB expert are probability distributions.

Results

We tested several algorithms to evaluate the contribution of combining model-based and model-free replay in terms of performance and computational cost. First, we are interested in simulating the two baseline cases, pure MF and pure MB algorithms, and how they perform with the respective MemR and SimR and limited budgets. Finally, we want to test the combination of the two strategies by using the criterion proposed in Dromnelle, Renaudo, et al., 2020, with either an infinite or a limited reactivations budget. Here are the relevant combinations of the same controller that we tested in this task:

- MF only agent, no replay
- MF only agent with MF replay (infinite replay budget)
- MF only agent with MF replay (budget: 200 replay iterations)
- MB only agent with MB replay (infinite inference budget)
- MB only agent with MB replay (budget: 200 inference iterations)
- MB+MF agent with MB replay (infinite inference budget)
- MB+MF agent with MB budget (budget: 200 inference iterations)
- MB+MF agent with MF replay (budget: 100 replay iterations) and MB replay (budget: 100 inference iterations) (a fair comparison with the previous cases because here the reactivation buffer is split in a maximum of 100 iterations per expert)

All the MB+MF agents use the Entropy and Cost (EC) coordination criterion described in Section 4.1.4. This criterion was taken from Dromnelle, Renaudo, et al. (2020), who showed that it allows for advantageous coordination between MB and MF experts and significantly reduces the computational cost of the inference phase without relevantly impacting the amount of gained reward. Table 4.3 shows the values of the parameters we used for these experiments.

The learning speed of all the above-listed agents was impacted when the reward's position changed at iteration #1600 (Fig. 4.9a). It is interesting to notice that the *MB - inference budget 100 + MF - replay budget 100* agent, which exploited the Entropy and Cost criterion with a limited budget for the two experts, shows a faster increase in the cumulative reward compared to all the other agents, from around actions #2500. As observed in the previous experiment (Sect. 4.1.3), replay contributes to increasing the speed of learning, and by combining the action of both MF and MB replay, it is possible to better account for both adaptability and generalization, drastically leading to a steeper accumulated reward over time slope of the proposed strategy, without having the same growth on the computational cost side (Fig. 4.9a and b). Concerning the cumulative cost, Fig. 4.9b shows that it rapidly increases for the *MB - inference budget inf* agent when the environment changed. Eventually, by action #4000, its cumulative cost has doubled the ones of the other agents.

Thus, considering the final overview of the performances and computational costs in Fig. 4.9c, deeper analyses and comparisons of the tested algorithms can be presented. The results are represented in terms of first, median, and third percentiles over 50 experiments. The cumulative reward is the amount of reward each agent has accumulated over the entire experiment, which is composed of 4000 iterations of the

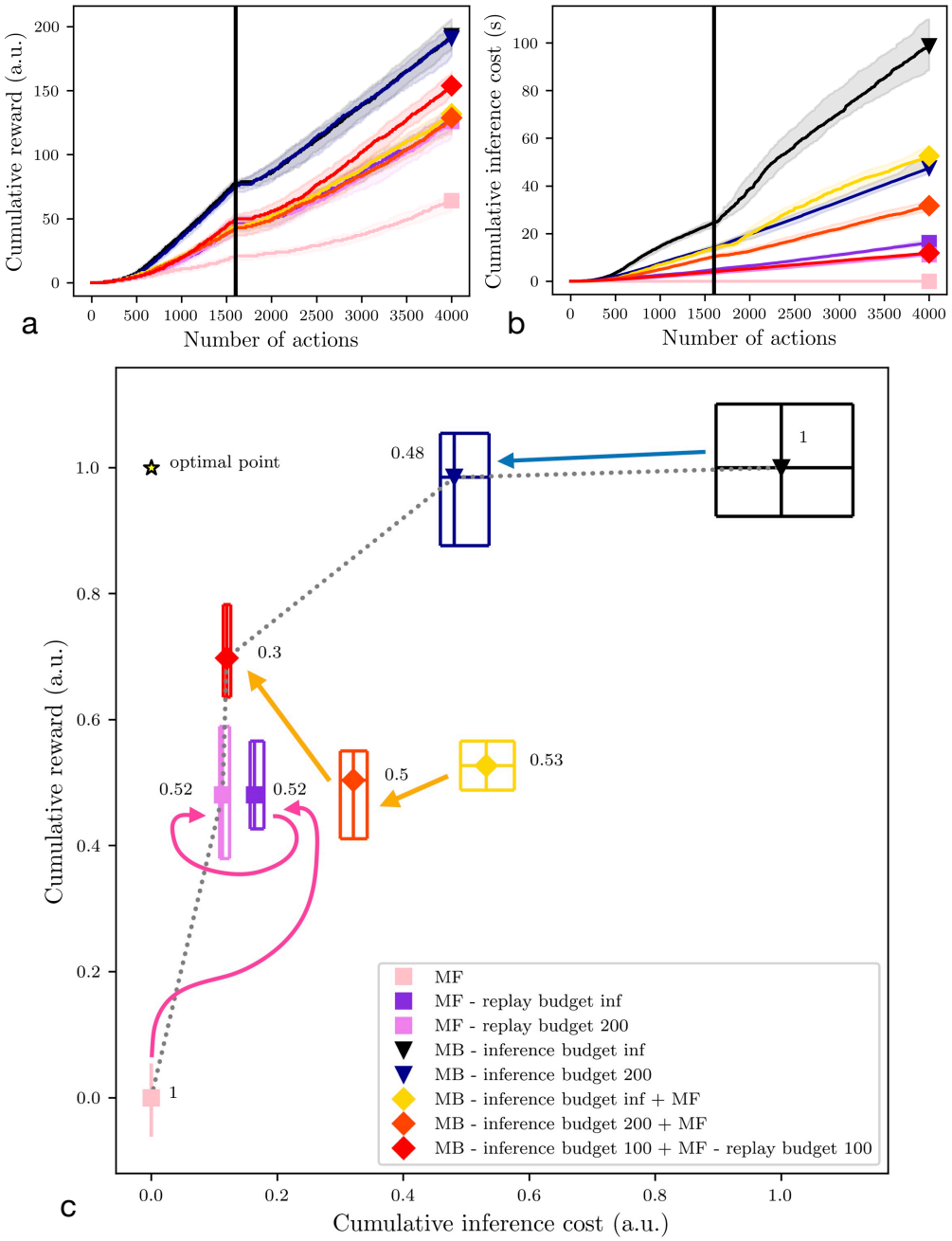


FIGURE 4.9: Overall performances of the different agents during their first 4000 actions in the environment. The vertical black line highlights the trial when the reward switch (1600). A) The dynamics of the reward's accumulation. B) The dynamics of the computational cost's accumulation. C) An overview of the algorithms' position within a normalized reward $\times$ cost space. The central polygons represent the median of the performance over 50 simulated experiments. Cumulative reward and costs have been normalized considering that the MF medians of the cumulative rewards and costs correspond to 0 and that the MB medians of cumulative rewards and cost correspond to 1.

	Model-based	Model-free	Meta-controller
α	-	0.6	-
γ	0.95	0.9	-
β	50	50	50
ϵ	0.01	0.1	-
RB	-	100	-
R_{tw}	-	100	-
T_{tw}	30	-	-

TABLE 4.3: Parameters used to generate the results in this section. They are taken from Dromnelle, Renaudo, et al. (2020) as a starting point for this work. α is the learning rate, γ is the discount factor and β is the exploration/exploitation trade-off parameter. For the MF expert, the converge threshold ϵ and replay constant RB have been introduced to design the convergence criterion, while ϵ for the MB expert is the same as in Dromnelle, Renaudo, et al. (2020). R_{tw} is the number of the last (s, a, s', r) tuples that the MF expert can replay. T_{tw} is the number of the last (s, a, s', r) tuples considered to built the transition model T for the MB expert.

learning, inference, and decision processes together (Sect. 4.1.4). The cumulative inference cost represents the time (in seconds) needed to perform the inference phase.

As expected, reward-wise, the best-performing agent is the pure MB, with an infinite inference budget (black triangle, on the top-right, in Fig. 4.9c). However, this agent is also the most costly in terms of computation during the inference phase. This issue can be partially fixed by reducing the MB replay budget to 200 iterations (blue triangle, in Fig. 4.9c). In this case, the inference phase will be stopped if the action-values have converged or if the number of inference iterations has reached the maximum budget (in this case, 200).

On the opposite side of the figure, the pure MF agent (pink square, on the bottom-left, in Fig. 4.9c) shows the minimum cost of the entire set of experiments but also the lowest cumulative reward. Adding replay to the MF expert, with an infinite replay budget (dark violet square in Fig. 4.9c) or a 200-iteration budget (light violet square in Fig. 4.9c) doubles the reward accumulation performance, with a limited increase in the computational cost (compared to the MB costs), in particular when adding the budget of 200 iterations budget.

From the results in Fig. 4.9c, we can deduce that for both the MF and the MB experts, most of the time, the number of needed reactivations is in the same order of magnitude as the proposed finite budget of 200 (since the cumulative costs are comparable). As already shown in Dromnelle, Renaudo, et al. (2020), with an MB expert with an infinite inference budget, the coordination of MB and MF experts via the EC criterion produces agents which are halfway between MB-only and MF-only experts regarding performances and costs (yellow diamond in Fig. 4.9c). Nevertheless, when limiting the MB inference budget to 100 and adding the contribution of 100 replay iterations for the MF expert (red diamond in Fig. 4.9), the cumulative reward increases, and the inference cost diminishes, moving the performance of the agent closer to the optimal point (star in Fig. 4.9). Moreover, the arrows highlight the progressions of the *MF-only* (pink), the *MB-only* (blue), and the *MB+MF* (orange) agents. Looking in more detail, the performance of the MF-only agents is improved by adding a budget of 200 MF replays. On the other hand, the performance of the

MB-only agents is slightly decreased by limiting the inference budget to 200 iterations, but the cumulative computational cost is significantly decreased. Starting from the performance obtained in Dromnelle, Renaudo, et al. (2020), in yellow in the figure, we obtain similar performances but decrease the computational cost when we limited the inference budget to 200 inference iterations for the MB expert, producing agents, which are halfway between MB-only and MF-only experts. After this analysis, we have tested the combination of the best strategies tried so far: the MB expert with a limited inference budget and the MF one with a limited replay budget. We have combined them through the EC criterion (Eq. 4.13). In this case, to have the same total reactivations budget as the other tested algorithm, we have shared the initial 200 reactivations budget to 100 SimR for the MB expert and 100 MemR for the MF one. With this combined replay effort, the overall performances reached an optimal compromise between performance and cost since the inference cost is substantially decreased while the cumulative reward was significantly raised, compared to the results obtained by Dromnelle, Renaudo, et al. (2020).

Given that the aim of each agent and its EC meta-controller is composed of two objectives: (1) maximizing the cumulative reward and (2) minimizing the cumulative inference cost, we compute the pareto front (black dotted line in Fig. 4.9c) which represents the solutions that approximate the set of all optimal trade-offs of the two given objectives. As expected, the pure MB and MF experts are pareto optimal solutions, very specialized in one of the two objectives. At the same time, by reducing and splitting their budgets, we can have agents that interestingly converge closer to the $OptimalPoint$ (star in Fig. 4.9c). To rank all the agents ag, the Chebyschev distance (Cantrell, 2000) from their median performance to the $OptimalPoint$ is computed as shown in the following equation.

$$Chebyschev\ distance\ (ag) = \max_{obj} \mid OptimalPoint_{obj} - median(ag_{obj}) \mid \tag{4.14}$$

where obj are the 2 normalized objectives of the solutions space (cumulative inference costs and cumulative reward). The computed Chebyschev distances are shown in Fig. 4.9c, on the side of each algorithm point, and show a clear picture concerning the proposed solutions; the agent sharing the reactivations budget between the MB and MF is the closest to the optimal point, followed by the MB expert with limited SimR budget. MF with MemR and MB + MF without MemR have very similar distances to the optimal points, meaning that the contribution of the MB expert is crucial in adapting to a dynamical environment. However, the cost of this computation can essentially decrease just when it cooperates with an MF agent with replay, which can also learn faster from the MB expert's Q-values update.

These results open new possibilities for the design of reinforcement learning control architectures in robotics. On the one hand, when dealing with probabilistic environments, MF replay might focus mainly on rare and not relevant transitions, leading to interesting exploration and computational economy, but misguiding the memory consolidation of relevant experience when changes happen in the task (as also seen in Sect. 4.1.3). On the other hand, when the transitions model is stochastic, combining the computationally competitive MF replay with the general knowledge of the environment acquired by MB replay can bring artificial agents and robots to better deal with a non-stationary reinforcement learning task.

4.1.5 Discussion

In this paper, our research question was whether reinforcement learning (RL) strategies using neuro-inspired replay methods, based on neuroscience knowledge about hippocampal reactivations, could improve the speed and the adaptability of robotic agents engaged in spatial navigation tasks. Model-free, model-based, and no replay RL techniques were compared in three simulated robotic experiments of increasing complexity and realism. Our results showed that in all levels of abstraction, the neurorobots learned the spatial task faster when replay was involved in the process and more efficiently when a Model-based replay method was used. Conversely, we show how a synergy between model-based and model-free replay strategies can be more effective in a more realistic and stochastic experimental set-up.

Applying reinforcement learning techniques to robotics requires coping with some specificities of operating in the real world (Kober, Bagnell, and Peters, 2013). First, making actual movements in the real world takes time, wares out the robotic platform, and can potentially damage it. Acquiring new data requires moving and is thus costly too: online learning processes must be as parsimonious on data use as possible. Second, making decisions also takes time, especially when using limited embedded computation systems, while operating in a dynamic world may require the ability to react exceptionally rapidly to avoid damage. Learning systems should thus be as computationally cheap as possible. Finally, both moving and computing consume the robot's energy, which is always available in limited amounts. This highlights the importance of developing robotic controllers that can (1) maximize their learning capabilities over experience and energy scarcity and (2) reduce the complexity of their algorithm to meet the computational limitations of embedded platforms.

Along with this paper, we have presented simulated experiments (sometimes based on data like transition maps first generated with a real robot) to investigate the possible advantages of equipping neurorobots with offline learning mechanisms inspired by hippocampal place cells' reactivations. These advantages are, first, to extract as much information as possible from the already gathered data and, by mixing the multiple types of learning processes with the multiple types of reactivations, to limit deliberation time and the costs mentioned above intrinsic to robotics. Starting with simpler and deterministic environments, as the double T-maze experiment presented in Sect. 4.1.2, this research illustrates that as the complexity of the states-actions transitions increases, model-based *Simulation reactivations* become more strategic for the learning capabilities of the agent (Sect. 4.1.3). Finally, in Sect. 4.1.4, the combination of model-free *Memory reactivations* and model-based *Simulation reactivations* is presented as an interesting proposal to merge the benefits of both techniques: prioritizing the model-based expert when the task requires more inference and generalization effectiveness to be solved (for example facing non-stationarity), while on the contrary giving priority to the model-free expert when an effective solution can be found relying only on recent experience.

When simulations increase in complexity, thus getting closer to a real robotic experiment, the challenges regarding the internal representation of the world (in particular, the states-actions space and the reward) increase. As presented in Fig. 4.10 where the environments of the two last experiments (presented in Sect. 4.1.3 and 4.1.4 respectively) are displayed in terms of maximum entropy per state, it is visible that the transition probability matrix created by the navigation of the real robot (Fig. 4.10b) results in a representation of the environment which is less homogeneous and more uncertain than the one learned with the simulated robot (Fig. 4.10a). In

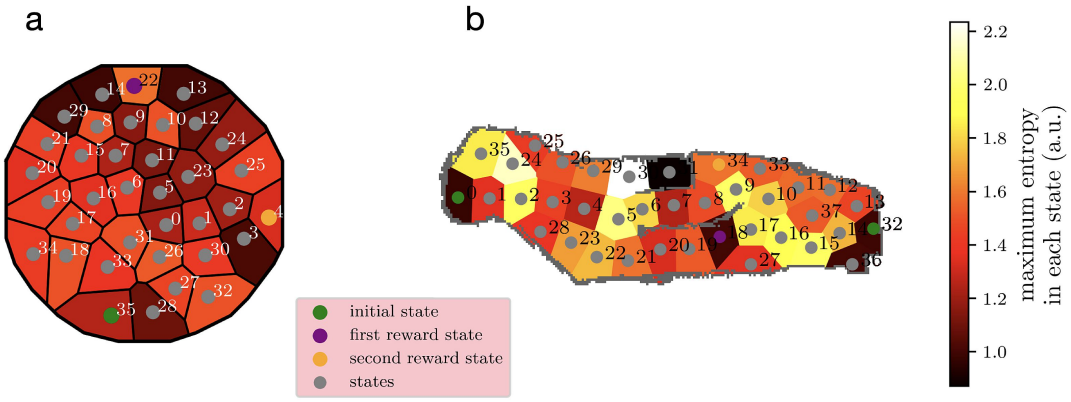


FIGURE 4.10: Representation of the navigation environments for the previous experiments (Sect. 4.1.3 and Sect. 4.1.4), organized in respectively 36 and 38 discrete Markovian states decomposed from the data acquired during the autonomous navigation of the robot, when no reward was present in the mazes. The initial and reward states for the tasks are also highlighted in the figure. In these heatmaps, the lighter the color of the state, the greater the maximal entropy of that specific state, according to Eq. 4.5. The represented scale of entropy values (0.87-2.23 a.u.) has been selected to cover the whole range of the computed entropies. Moreover, in both environments, the robots have navigated 5357 actions. A) In the case of the circular maze (Sect. 4.1.3), the navigation and the transition model are acquired after simulated navigation on ROS Gazebo. B) In the second experiment (Sect. 4.1.4), the navigation and the transition model are instead computed after the real robot navigation, which generated a wider range of maximal entropy values, sometimes also very low due to the presence of walls that categorically constrained certain states of the environment.

mobile robotics, localization may often depend on a few sensory information, as in the case of the mobile robots used in our experiments. Such limited information is fundamental for acquiring a solid representation of the environment. For these reasons, the entropy maps in Fig. 4.10 reflect the nature of the two mazes: the uncertainty is more homogeneous in the circular maze (Fig. 4.10a) since the environment is an open space which gives the agent an even chance to end up visiting the neighboring states. In contrast, the second environment (Fig. 4.10b) is more extended in one dimension and presents inner walls that result in a fuzzier level of uncertainty on the transitions model of the environment.

Future works in this research direction would include the comparison with the RL algorithms performing forward replays, which are of crucial importance in standard rodent navigation tasks, such as the multiple T-maze (Johnson and Redish, 2007). These forward-shifted spatial representations have been demonstrated to happen mainly at decision points to predict the consequences of the following actions. Their effect has already been successfully modeled in neurorobotics by Maffei et al. (2015), where they implemented the extraction of relevant policies by consulting memory. On the other hand, Seijen and R. Sutton (2015) argued that it is mathematically equivalent to update Q-values in a model-free way combined with replay and to update Q-values in a model-based way, given that the elements in the memory buffer, used for replay, are the same than those used to build the model. Moreover, RL-based replay strategies can also generate forward replay events (Khamassi and Girard, 2020) and enable RL-based models to still account for neurobiological

data (Mattar and Daw, 2018; Cazé et al., 2018).

In summary, this work presented new and crucial results concerning the advantages and the limitations of different RL-based replay techniques for robotics, gradually testing them in more complex and realistic circumstances. Additionally, this research paves the way for new studies on the role of replays in neurorobotics, in particular in spatial navigation tasks where generalization effectiveness and time efficiency are key.

Finally, the addition of RL techniques, inspired by hippocampal replays, shows an improvement in the performances in the presented navigation task, in particular concerning the exploitation of the past experience, knowledge propagation, and as a consequence, the speed of learning. In particular, model-based *Simulation Reactivations* significantly contributed in the case of non-stationarity. However, a fruitful coordination with model-free *Memory Reactivations* became crucial in terms of computational cost reduction. All these insights, found in robotic experiments, implemented with different levels of abstraction, can encourage new neuroscientific experimental protocols and shed light on a better understanding of the phenomenon of hippocampal replay.

4.2 TaVAR: a robotic demonstration for teaching reinforcement learning

The work described in this section has been accepted to the conference "Drôles d'objets - Un nouvel art de faire" as "B. Girard, L. Gaillot, L. Laval, L. Le Peutit, E. Massi, F. Sangaré, I. Tuzun (2023). TaVAR : Une Table lumineuse pour Vulgariser l'Apprentissage par Renforcement."

Explaining the topic of your thesis or some classic Reinforcement Learning (RL) algorithms can be easy when you are talking with your co-workers or other researchers in a similar domain. Nevertheless, explaining simply and clearly how RL can help an artificial agent or a robot to reach a desired goal by learning from its interactions with the environments is not straightforward. Showing the equations that make the robot learn an optimal behavior to optimize the cumulative reward is not the most effective way to divulge RL basic concepts to your audience, if they have a different background.

Coordinated by Benoît Girard, and with the help of the master students Lydia Gaillot, Léo Laval, Laurine Le Peutit, Ilke Tuzun, and Fousseyni Sangaré, we propose a robotic navigation set-up that, with the help of a transparent table and a projector, shows which information the robot gets from its interaction with the environment and how it exploits it to make decisions and generate a goal-directed intelligent behavior. The proposed robotic set-up is called *TaVAR (Table lumineuse pour Vulgariser l'Apprentissage par Renforcement)*. TaVAR is also used to explain model-free unordered replay, with the possibility of replay just one experience per time, to show the details and the effect of replay in learning the navigation task. This replay demonstration also gives the chance to introduce what hippocampal reactivations are and to talk about the transfer of their computational principle in RL and robotics.

4.2.1 Material and methods

TaVAR is composed of an alluminium frame that sustains the table's wooden boards. The table surface is built with a transparent 180cmx160cm plexiglass layer with a translucide film attached to it.

The main communication node is a desktop computer exchanging information with the robot, the projector and the person presenting the demonstration (Fig. 4.11).

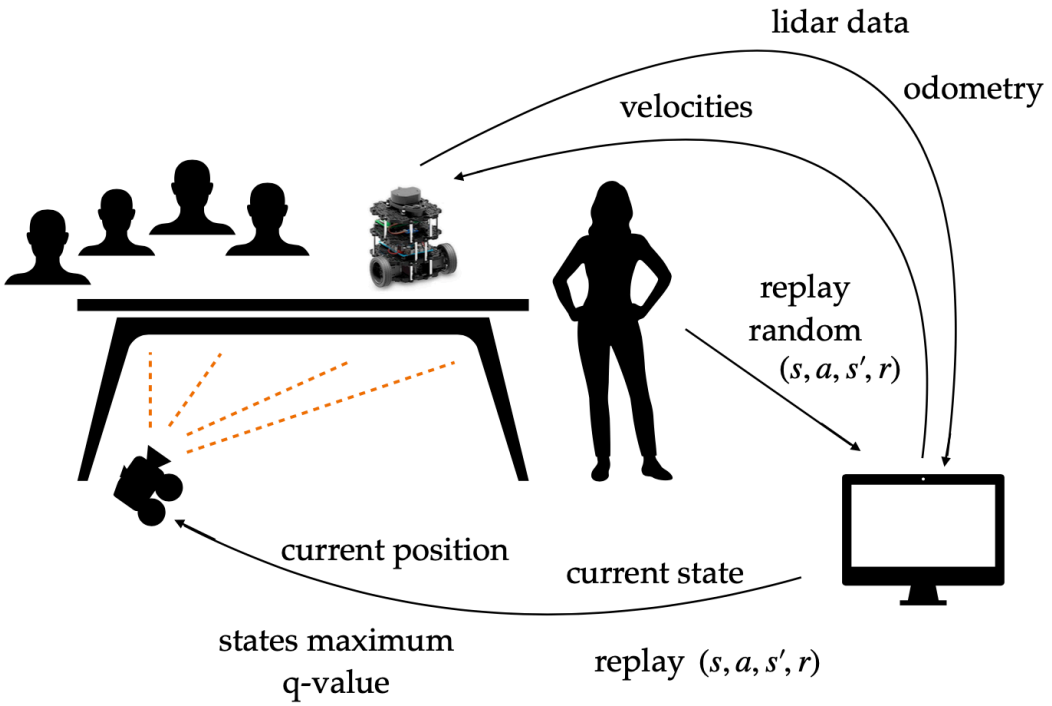


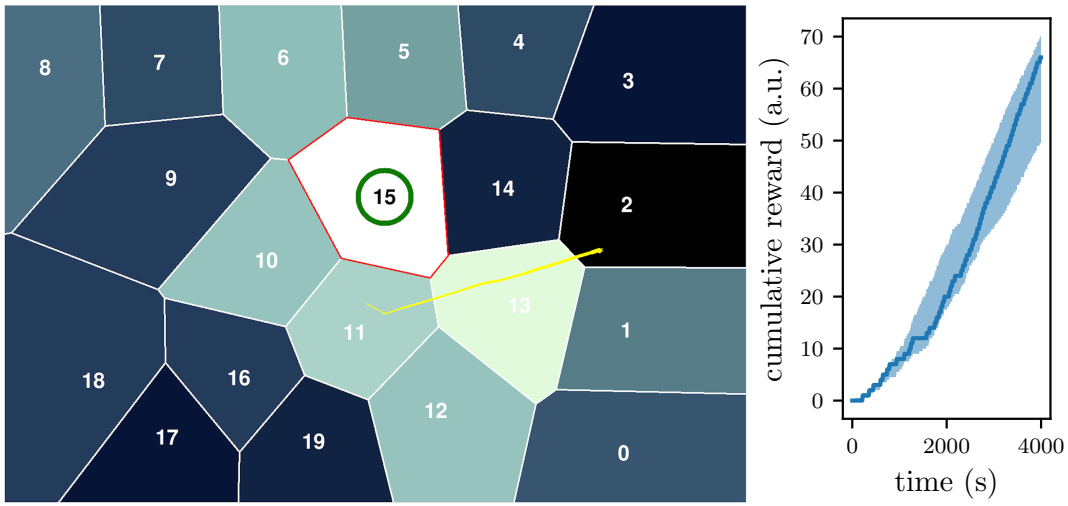
FIGURE 4.11: TaVAR communication set-up

The robotic platform is a Turtlebot3 burger [4] which can localize itself thanks to a classic SLAM (Simultaneous Localization And Mapping) algorithm, GMapping (Grisetti, Stachniss, and Burgard, 2007), running on the desktop computer. The robot is sending in real-time the data measured from its lidar and odometry sensors to the desktop computer which is then able to localize it with respect to the map of the environment (table) by using GMapping.

The table space was previously divided by a random autonomous robotic exploration into 20 states (Fig. 4.12a, always using GMapping) and from each state 8 actions are possible in the cardinal directions of the absolute reference frame of the environment (the same as for Sect. 4.1.4). In addition, there is one reward state, state 15, which gives an unitary reward to the robot that come back alternatively to two opposite starting states (state 0 and 8) after having reached the reward state (Fig. 4.12a). Thus, the real-time information about the robot position and orientation is also used to derive the current state of the robot.

The robot learns through experience by using a Q-learning algorithm (Watkins, 1989) running on the desktop computer with the following parameters: α is 0.2 and γ is 0.9. Fig. 4.12b shows that, after around 20/30 min, the algorithm makes the robot converge to an optimal behavior where it accumulates reward at a constant speed. In

[4] https://emanual.robotis.com/docs/en/platform/turtlebot3/overview/



(A) Final states' q-values visualization (B) Learning performances

FIGURE 4.12: Demonstration experiment's dynamics. A) Example of the visualization projected on the table close to the end of an around 1h experiment. The white state 15 with the green circle is the reward state and the initial states are alternatively 0 and 8. The blue gradient shows the normalized max $Q(s,a)$ for each s with lighter values been larger max $Q(s,a)$. The yellow line shows the most recent trajectory covered by the robot. B) Median, first and third percentile (in the blue shadow) over 3 experiments of around 1h of the cumulative reward obtained by the robot.

order to get a smoother demonstration with a more efficient coverage of the states-action possibilities during the first stage of the experiment, when there is an equal probability to choose one of the possible next actions, the decision-making process of the robot picked one of the possible next actions based on the softmax distribution (Eq. 4.2 with $occur(s,a)$ in the place of $Q(s,a)$) created from the negative exponential $-e^{occur(s,a)}$ of the number of the previous situation $occur(s,a)$ when that action a was chosen from s (always with the same $\beta = 20$).

The colors of the tiles projected on the table are a blue gradient representing the maximum q-values among the state-action couples starting from that specific state. The blue gradient scale is adapted to the relative maximum q-values and uses colors visible from people suffering from different types of color blindness.

To receive the necessary data and generate the visualization showed on the table, we used a Raspberry Pi connected to the projector where a ROS node controls the real-time image using the *tkinter* python library. By pressing a button on the Raspberry Pi keyboard, the person who is presenting the demonstration can choose at any time (either while the robot is navigating or when it is pausing in its starting position or wherever on the table) to randomly select one of the past robot transitions (s,a,s',r) and to replay it, to directly show the effect of an experience replayed transition (Lin, 1992). The replayed transitions will be visible on the table as a green arrow pointing from the centre of s to the centre of s'. In the meantime, the possible change in the maximum q-values for s will be showed on the table, since the maximum states' q-values are continuously updating on the table's visualization. In fact, the visualization of the states' blue gradient that the robot is using to learn is continuously projected from a short focal projector, which is fixed under the table, to the table surface in a way that the audience can see the projected real-time information

from above the table.

The desktop computer is the node sending the velocities to the robot's motors. This motor commands depend on the action decided by the navigation policy of the robot, which picks the next action from the softmax distribution (where β is 20) of the q-values associated to (s, a_i), with $i \in [0, 8]$ (Eq. 4.2) or by the speaker's instructions.

At each time, the speaker presenting the demonstration can choose to stop the robot's navigation and restart it as it better follows the flow of their explanation. To ease its explanation, by tapping different buttons on the keyboards, the speaker can also show and hide a trace of the most recent robot trajectory (in yellow in Fig. 4.12a) and color the current state where the robot (in gray). Finally, another keyboard button deletes the visualization of all the current states q-values.

4.2.2 Results and discussion

TaVAR has been used for the first time in October 2022 for the *Fête de la Science 2022* at the Institut des Systèmes Intelligents et de Robotique (ISIR). The demonstration has been run for 5 hours without significant problems. Fig. 4.13 shows a photo that was taken during the demonstration day.

FIGURE 4.13: TaVAR demonstration for the *Fête de la Science 2022*. The robot is navigating on the table during the demonstration. On the table, the states' discretization is visible with colder states representing larger q-values around the reward state (state 15, in white with the green circle.

After this first demonstration, the aim is to further improve the demonstration by:

- adding obstacles on the table to create narrow corridors in the task to make the exploration environment more similar to the mazes usually employed in studying spatial navigation in mice (like the double T-maze Gupta et al. (2010));

- showing also the learning contribution of more structured types of replay (*i.e.*, reverse-ordered replay);
- optimizing the code that deals with the visualization on the table; the data that are visualized on the table are slightly accumulating a delay with respect to the computations done by the robot because the new information is shown on top of the previous one. Thus, the idea is to test if refreshing the table's "canvas" in real-time could lighten the visualization process and solve this problem;
- for each state, indicating with an arrow the action which has the largest q-values, to better explain that the learning values are assigned to state-action pairs and just to states. This is important to explicit which action has been learned to be a good behavior from each state because it can happen the ambiguous situation where states which are next to the reward have low q-values because the robot has not discovered yet that the reward is just one step away);
- investigating the educational value of the demonstration by asking the audience/participants to fill out a brief survey on the main RL concepts they have just seen.

The robotic set-up designed for TaVAR is also being used for running experiments with the learning methods presented Sect. 4.1.4 to test the contribution of replay-inspired control strategies when all the phases of the experiment are performed on the real robot. The one factor that has been changed in the transfer to the real robot concerns the decrease in size of the memory buffer. In fact, the real robot adds an extra level of stochasticity concerning sensors' noise and localization errors. These factors, that were not present in the results described in Sect. 4.1, together with a smaller learning rate, require the robot to faster adapt to uncertain scenarios. By constraining the information that builds the robot's model of the world to the most recent ones, we can prevent long adaptation loops when the environment changes or sensors and localization measurements are inaccurate.

To the best of our knownledge, our results in studying the role and the combination of different types of RL-based replay in robotics are among the first ones that have been presented in the field. Moreover, the simplicity that is thoroughly employed in the design of a demonstration for a general audience is highly helpful to get a clear identifcation of the main features and criticalities of the different tested learning strategies.

To summarize the contributions of the machine learning and robotics chapter of this thesis:

- By testing different RL-based RL strategies in spatial learning experiments that range from pure simulation to real robots experiments, we have concluded that MB strategies are always preferrable when dealing with dynamic scenarios when the agent needs to be adaptive and, in particular, the MB prioritized sweeping algorithm is among the most performing strategies since it concentrates the replay activity just in the moments when the experiment changes;
- By adding MB- and MF-RL replay strategies on the meta-controller orchestrating between habitual and goal-directed behavior proposed by Dromnelle, Renaudo, et al. (2020), we have improved the learning performances of the agent and also minimizing its computational costs;
- We have implemented a robotic demonstration to explain RL and the role of hippocampal reactivations through a goal-directed navigation task. This new

robotic set-up is also used for new experiments on spatial learning and reactivations.

Chapter 5

Conclusions

In this chapter, a summary of this thesis's scientific contributions is presented (Sect. 5.1). Secondly, we discuss the possible future developments and perspectives from this thesis to address the limitations of the presented results in order to improve the future scientific contributions that will follow this thesis (Sect. 5.2).

5.1 Summary

We summarize how our scientific contributions answer the research questions we have presented in Sect. 1.2.

- Sect. 3.1 presents a new **data-driven value-based decision-making model for rodent free spatial exploration** that has been designed and optimized against three different datasets. In this contribution, we have proposed a general framework for modeling free exploration in rodents by:
  - a common pre-processing and sampling routine for exploratory rodent behavioral data;
  - **the identification and formalization of three relevant behavioral components for rodent free exploration: safety, biomechanical cost, and biomechanical persistence**;
  - **a Markov Decision Process (MDP) formalization for generating a decision-making process which produces a rodent-like free exploratory behavior.**
- Sect. 3.2 expands the contribution of the proposed free exploration model to account for externally conditioned situations. Our results predict that:
  - as already observed in the results by Bryzgalov (2021), it exists a negative correlation trend between the number of offline replay sessions and the differential occupancy of the negative stimulation areas between the post- and the pre-conditioning phases. This means that **the number of offline reactivations could predict how strong the animal would avoid the stimulation area**.
  - the number of offline replay sessions is slightly negatively correlated to the amount of negative stimulation, while in the positive case, no correlation exists. Even though the negative correlation that exists in negative conditioning is not significant, this can suggest that offline replaying emotionally relevant experiences could be increasingly important as the interaction with the negative stimulation becomes shorter.

  - **the contribution of offline sleep replay is more relevant when learning to avoid a negative stimulus than learning to approach a positive one**. The post-conditioning behavior reflects a biased occupancy of the maze either for the negative stimulation areas (avoided) or the positive ones (occupied), even though the negative stimulus has been very scarcely experienced (around one order of magnitude lower, Sect. 3.2.4 and Fig. 3.30) compared to the positive one.

- Sect. 4.1 tests different hippocampal replay-inspired reinforcement learning (RL) strategies for spatial navigation in neurorobotics. With a particular focus on the role of model-based (MB) and model-free (MF) RL replay strategies, our results imply that:

  - **the integration of replay-inspired RL techniques always improves the performances of the tested RL algorithms** (in terms of cumulative reward or time to get to the reward location).
  - while **moving towards real robotic scenarios**, where the stochasticity of the MDP describing the experiment increases or the task becomes non-stationary, MB replay (also called ***Simulation Reactivations*** in Massi et al. (2022)) **are desirable because they can faster adapt thanks to their acquired knowledge on the MDP**.
  - when transferring a simulated experiment on a real robot, new constraints arise due to the real-time computations and the short battery life. In this case, **a coordination between MB and MF learning systems with replay** (*i.e., Simulation Reactivations (SimR)* and *Memory Reactivations (MemR)* in Massi et al. (2022)) **is a proper trade-off which alternatively exploits the MB's learning velocity and adaptability and the MF's low computational cost. Adding SimR and MemR to the algorithm also improves the results obtained in Dromnelle, Renaudo, et al. (2020)**.

In conclusion, the main scientific contributions presented in this thesis assess the importance, while using the theory of RL, of integrating mechanisms inspired by hippocampal replay in the behavioral modeling of spatial learning in rodents and in the design of neuro-inspired controllers (*i.e.,* neuro-controllers) for rapid and adaptive robotic navigation.

5.2 Discussion and future perspectives

This thesis applies the theory of Reinforcement Learning (RL) either to model data-driven behaviors (Chapter 3) or to design neuro-controllers (Chapter 4) to foster the exchange and the integration between computational neuroscience and robotics. We have focused this philosophy on the topic of hippocampal reactivations by investigating how they can be studied, modeled, and adopted in both computational neuroscience and robotics.

The uniqueness of this work consists in a computational study of the role of hippocampal reactivations in spatial learning which covers biological behavior, in rodents, and artificial behavior, in robots. From one side, our neuroscientific results suggest that two parallel learning strategies are adopted in appetitive and aversive learning. The sampling bias that occurs when animals experience a very few times a negative stimulus compared to an equivalent positive one could be computationally overcome by associating higher relevance to negative events and by replaying

them more often than the corresponding positive ones. This would encourage the implementation of two different controllers, and possibly two different RL-replay strategies, when opposite valence conditioning affects the spatial navigation of an artificial agent or a robot. On the other side, the robotic results of the thesis assess a prominent role of model-based knowledge in efficiently solving dynamical spatial learning tasks and, in particular, different stages of the tasks, namely the beginning, the period after policy convergences, and the task change, privilege respectively MB, MF, and MB learning algorithms and replay strategies. As it has been already studied and described in Sect. 2.1.2, the brain learning strategies that orchestrate behavior are multiple and can be mainly traced back to MB to tackle goal-directed behavior and MF to tackle the habitual one. Our results suggest that this aribitration, based on accuracy and computational cost, between MB and MF learning, could also guide the generation of hippocampal reactivation into more planning-oriented replay or experince-driven ones.

A more detailed discussion over the main contributions, challenges, and future perspectives of the thesis is given in the next sections.

5.2.1 Neuroscience

We suggest that general behavioral tendencies exist for rodents freely exploring a new environment for the first time. To our knowledge, most of these behavioral trends have already been studied in the literature (Fonio, Benjamini, and Golani (2009), among the others reviewed in Sect. 2.2.3). These previous studies have disclosed the existence of behavioral patterns under specific constraints, such as the presence of homecages in the environment. However, their permanence in different mazes' morphologies or for different time intervals experiments has not been explored yet.

Compared to the other models proposed in the literature (Gordon and Ahissar, 2012; Gordon, Fonio, and Ahissar, 2014a), our novelty is to propose a formalization of the problem as an MDP. The proposed model aims to describe and reproduce the main general behavioral characteristics of free spatial exploration in rodents in terms of attraction for safer locations and of exploratory dynamics (static or active, stronger or weaker directional preference). The parameters optimization process that is executed for each animal intends to generate an artificial agent whose decision-making process, in the proposed MDP, generates a behavior comparable to the one from that specific rodent in terms of the behavioral metrics we have identified from the data.

One of the limitations of this contribution is that its first design was proposed based only on the first u-maze dataset we had. Thus, it does not consider the other two datasets made available to us later in the thesis (square open-maze and grid-maze). In the process of generalizing the free exploration model to the two new datasets, we have detected possible improvements for the model even though, in most cases, our results are significantly better than random exploration without further adaptation or modifications of the model. In particular, the selection of the relative directional bins of the definition for the biomechanical cost metric should adapt to the possible next actions in the MDP's formalization to properly fit the simulated agent's decision-making process, which aims at replicating the animal's behavior.

Moreover, even though the parameters selection process is minimizing the three objective metrics (safety, biomechanical cost, and biomechanical persistence objective) at the same time, the action of one behavioral component influences not only

its own behavioral objective but also the others. For example, a strong safety component could influence the biomechanical persistence metric by leading to a very static behavior with long exploratory pauses in the corners of the maze. That is why to assess the contribution of each behavioral component independently, it would be interesting to study the behavioral metrics' response when the free exploration model is composed just by one of the three behavioral components and investigate deeper if just one or two behavioral components could suffice to capture the three behavioral metrics alone. If this is the case, this could lead to a simplified version of the model, able to describe the same behavioral complexity as the proposed one.

Nevertheless, the proposed model, which reproduces the main behavioral features of rodent free exploration through a decision-making agent in an MDP framework, is relevant for future RL developments or extensions of this model and also for bridging these results to neurorobotics navigation tasks. In fact, even though we have not used this strategy during the TaVAR demonstration, when we were trying different exploratory solutions, we easily integrated the biomechanical cost component, identified in the rodent data we have analyzed, on the behavioral controller for the robotic demonstration, generating a similar directional bias as we saw in the data (even though in a different timescale).

Moreover, this value-based framework allows for an easy integration of the state-of-the-art RL replay strategies that are usually tested in modeling the contribution of hippocampal replay in computational neuroscience (Pezzulo, Kemere, and Van Der Meer, 2017; Cazé et al., 2018; Mattar and Daw, 2018; Khamassi and Girard, 2020) and in neurorobotics goal-directed navigation (Whelan, Prescott, and Vasilaki, 2020; Massi et al., 2022).

In particular, recent studies suggest that Reinforcement Learning (RL) can be considered a proper framework to model value-based spatial learning (Glimcher, 2011; A. Collins and Khamassi, 2021) and eventually the contribution of hippocampal replay in this process (Pezzulo, Kemere, and Van Der Meer, 2017; Cazé et al., 2018; Mattar and Daw, 2018; Khamassi and Girard, 2020). Thus, we include to the free exploration model an extra component, which learns a spatial representation of the values of an exogenous emotional conditioning. After a learning phase, based on the reconstruction of the conditioning experimental sessions for each animal, we optimize the number of unordered offline replay sessions and the learning parameters of the adopted RL algorithm based on the post-conditioning occupancy of the mice in the u-maze. Thus, the exploratory behavior of the agent is the result of its free exploratory behavior (previously optimized based on its exploration during the first time it enters the maze), together with the contribution of a conditioning component based on the learned and replayed states' values.

Our results assess that, based on the theory of RL and not involving any complexifying assumption on the type of hippocampal reactivations happening in the sleep phase after the conditioning sessions, a recall and re-elaboration of the past experience is computationally convenient to replicate the post-conditioning behavior of the animals.

One of the limitations of the extension of the free exploration model with the conditioning component concerns the fact that after new stimulations trigger an emotional response in the animal, the main characteristics of its free exploratory behavior could also change. This means that after a negative stimulation, rodents could, for example, prefer safer places and remain more static than during free exploration, like when freezing reactions happen (D. C. Blanchard, Griebel, and R. J. Blanchard, 2001). Our other idea for the parametrization of the conditioning exploration model is to re-optimize also the parameters that describe the relevance of the behavioral

components ($p1$, κ, bp for the safety, biomechanical cost, and biomechanical persistence component, respectively) to capture the change of contributions of these components in the post-conditioning case by optimizing the original behavioral metrics of free exploration plus the conditioning one. In a first attempt at optimizing the parameters, we had actually employed a four objectives optimization. However, we realized the four objectives were not properly being optimized because this would require many more generations than the case with a single objective, to allow for a deeper model parameters search. Exploring the use of multi-objective optimization, also in this conditioning exploration case, would be another interesting research approach, but that requires more time or computational power.

The integration of RL replay mechanisms in our model allows for new possible considerations and predictions. For example, the fact that opposite valence stimulation could differently activate the hippocampus has already been observed by Segal, Disterhoft, and Olds (1972) who noticed a more prominent hippocampal global activity following a positive stimulation than an aversive one. Nevertheless, experimental studies and computational models addressing the activity of hippocampal reactivations after positive or negative conditioning have only recently started to be studied. Also, this type of research entails some criticisms due to the ethical responsibilities of imposing negative stimulation on animals.

Thanks to this new behavioral data and the work by Bryzgalov (2021), we could optimize our exploration model proposal for the same set of mice experiencing opposite valence stimulations. Interestingly, our model was able to replicate the correlation between the amount of sleep reactivations and the differential avoidance of the shock areas. Further, based on the assumption that negative events are situations animals and humans do not want to experience very often, but they equally want to learn from them efficiently, we look into the existence of other potential relationships between replay sessions and behavior in our optimized results. Based on the prediction of our model, negative experience computationally requires longer unordered offline replay sessions to succeed in avoiding the shock locations as the corresponding animal does. Coherently with this interpretation, our results suggest that, as the amount of negative stimulation received increases, the replay sessions needed decrease.

Naturally, experimental studies and new validations of the presented results against more data are needed to assess this hypothesis. We hope that these data-driven computational predictions inspire other researchers in the field to investigate more deeply the relationship between the emotional valence of the stimulation and the following asleep (but also awake) hippocampal reactivations. In particular, the proposed exploration model could be helpful in predicting rodent exploratory behavior and the length or the relevance of the replay sessions in new spatial learning experimental protocols.

In the future, once more information concerning the observed hippocampal reactivations will be available (*e.g.*, if they are unordered, sequential, or biased towards emotionally relevant locations, for example), computational models, as the one we propose, could be improved by adopting more specific computational strategies (*e.g.*, reverse replay, trajectory sampling, and prioritized sweeping) for describing and reproducing the effect of hippocampal reactivations and reveal a more robust predictive power.

5.2.2 Machine learning and robotics

The integration of RL mechanisms, inspired by hippocampal reactivations, is not new in machine learning. However, the impact of diverse replay strategies has just recently started to be investigated, particularly in robotics. The transfer of these results to real robotic experiments is crucial and scarcely explored. With our machine learning and robotic contribution (Chapter 4), we start to address this concern by testing different replay strategies in different simulated and robotic scenarios, always in the framework of the theory of RL. Even though more research studies in this direction are encouraged, our contribution suggests that the trade-off between learning performance and computational cost, identified to be crucial for real robotic (Kober, Bagnell, and Peters, 2013; Dromnelle et al., 2022), could also be successfully appliedon the study and the introduction of hippocampal reactivations in robotic spatial learning tasks. We show that the MB/MF coordination strategy for instrumental goal-directed navigation proposed by Dromnelle et al. (2022) can be improved if limited budgets of RL model-based (MB) and model-free (MF) replay strategies are included in the two learning systems, respectively.

An interesting future development for coordinating the MB and MF learning systems will be to dynamically allocate the replay budget between Memory (MemR) and Simulation Reactivations (MemR), instead of evenly sharing it between the two experts, as we proposed in our work. A crucial point for this improvement will be to compute the budget allocation with a light computational strategy to not significantly deteriorate the computing efficiency gained with the fixed assignment of the replay budgets. By taking inspiration from the observations and models of hippocampal reactivations in neuroscience (Cazé et al., 2018; Mattar and Daw, 2018; Khamassi and Girard, 2020), on the one hand, we can suggest favoring MemR when an unexpected reward or absence of reward (high reward prediction error in the Eq. 2.4 formulation) is found, to consolidate the memory of this new significant event. On the other hand, it would be suitable to favor SimR at crucial decision points (when the decision entropy in Eq. 4.12 is high) to use the MB knowledge about the transitions model to decrease the decision uncertainty before acting. This information (reward prediction error and entropy) is already computed by the MB and MF learning experts and, thus, is already available to allocate the replay budget without further expensive computations.

Further efforts in this research direction are required to integrate such coordination between multiple RL controllers and replay strategies to a bio-inspired self-localizing and navigating system, such as the ones proposed by Dollé et al. (2010), Whelan, Prescott, and Vasilaki (2020), and Souza Muñoz et al. (2022). Also, the definition of "fuzzy" (Touretzky, Wan, and Redish, 1994) and multi-scale (Llofriu et al., 2015) discrete spatial areas, inspired by the spontaneous activations of hippocampal place cells, have been observed to reproduce the biological place cell circuitry dynamics better and improve the learning speed in goal-directed robotic navigation tasks.

Another critical issue that we have started to address in this thesis and that would need more attention in the future is the study of the relationship between the stochasticity level of an MDP and the subsequent performances of the RL replay strategies. In particular, a standard definition for defining the level of stochasticity of an MDP, for example, based on the entropy of the states, as we proposed in Sect. 4.1.5, would be desirable. Also, real robotics makes even more evident that both MB and MF replay crucially speed up the learning in a navigation task but sometimes reinforce the learning of sub-optimal strategies, which can be, for instance, the

first ones identified by the robot. This is why it would be interesting to look systematically into the interplay among different levels of stochasticity of the MDP, the learning parameters of the RL algorithms, and the types, duration, and termination criteria of the replay methodologies used.

Finally, we propose TaVAR, a new robotic framework to explain RL and experience replay with the help of an intuitive real-time visualization of the spatial information used by the robot to make decisions and achieved the desired goal. The main improvements to the presented version of TaVAR concern a quantitative evaluation of the dissemination efficacy of the demonstration and a more detailed visualization of the robot's decision-making process.

Appendix A

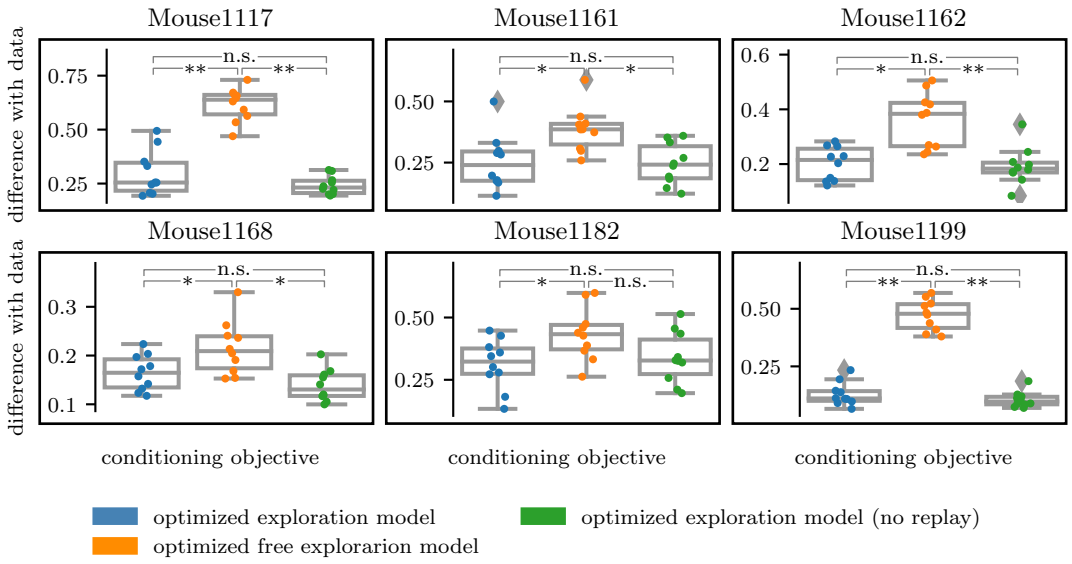
Supplementary material

A.1 Results

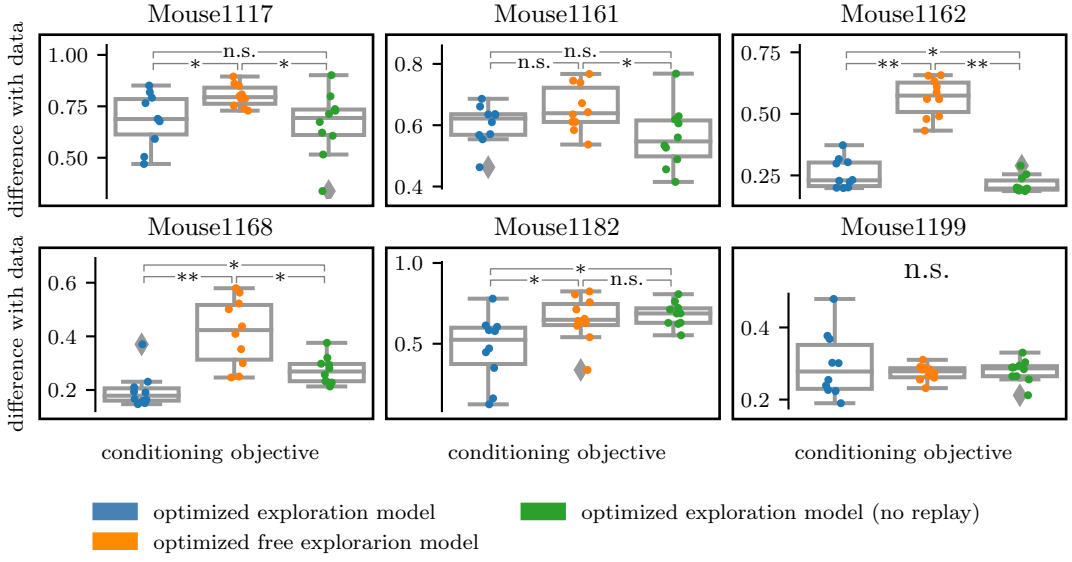
To further test what the possibility of performing offline RL replay mechanisms brings to the model and the subsequent behavior, we have tried to optimize a version of the exploration model where the replay buffer size #*rs* was not a parameter, and it was fixed to zero. In Fig. A.1, we can see that in most of the cases (9/12 between in total considering Fig. A.1a-A.1b) the optimization of the two models (with replay and without replay) obtains comparable performances in term of difference with the data for the occupancy of the sub-areas.

In the 2/3 remaining cases (Mouse1168 and Mouse1182, in negative conditioning), the optimized model with replay (in blue) is significantly better than the one without the replay (in green) and than to the free exploration one (in orange). In the case of positive conditioning (Fig. A.2a) where the results between replay and not replay model are always comparable (Fig. A.1a), to compensate for the absence of replay, the optimized α values of the models with no possibility to perform offline backups are usually higher than the ones in the model with replay (in 4/6 cases).

Also, as in the case of the optimized exploration model with replay (Fig. 3.29), in the optimized no-replay model in 4/6 mice, α and Wr have higher values for the negative than positive conditioning case, showing the need for a more decisive contribution of this conditioning component in the exploration model compared to the positive case.

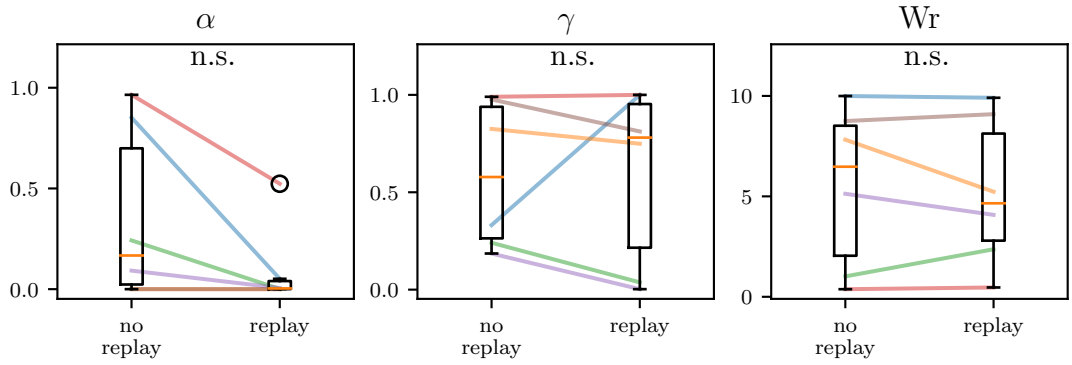


(A) Positive conditioning

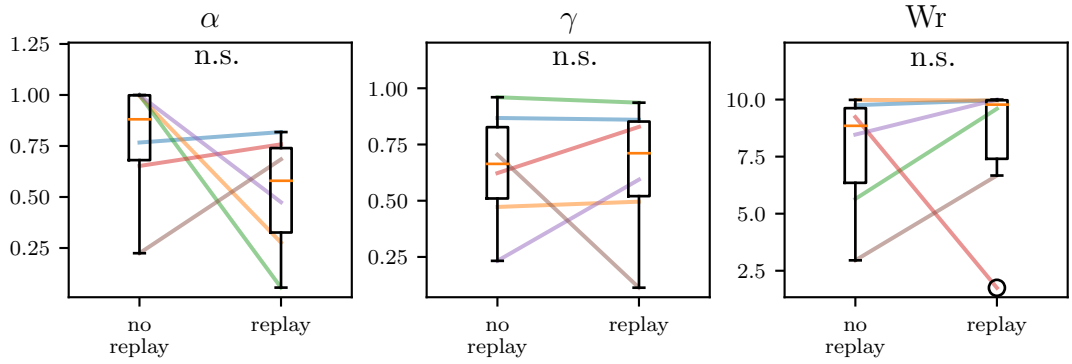


(B) Negative conditioning

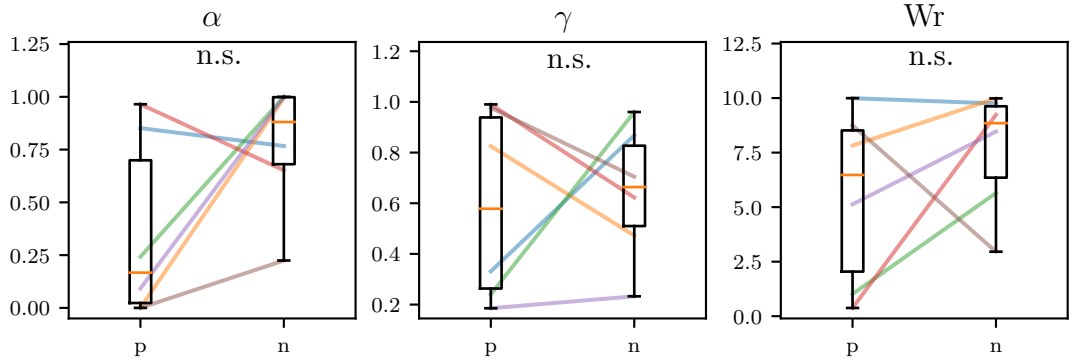
FIGURE A.1: Comparative statistical analysis on the conditioning objectives for the selected optimized models with replay, without replay, and the corresponding free exploration model. Each sub-figure represents the results for each mouse individual in terms of behavioral difference with the data. This difference is expressed as the conditioning objective (Eq. 3.25). ** indicates that the p-value resulting from the post hoc Wilcoxon-Mann-Whitney pairwise comparison test is lower than 0.001, * that it is lower than 0.05 and non-significant (n.s.) otherwise. The post hoc test has been performed following a Kruskal-Wallis H-test (Kruskal and Wallis, 1952) to reject the null hypothesis that the population median of all of the models' difference with the data was equal (this happens just for Mouse1199, in negative conditioning).



(A) Positive conditioning



(B) Negative conditioning



(C) No replay

FIGURE A.2: Statistical analysis on the exploration model parameters for the best individuals found by CMA-ES with and without replay, in case of the positive stimulation data (p) and negative stimulation ones (n). *#rs* indicates the number of replay sessions, α the learning rate, γ the discount factor, and finally *Wr* the weigth for the conditioning component. * means that the p-value resulting from the Wilcoxon-Mann-Whitney comparison test between the distributions of the model parameters in positive and negative stimulation is lower than 0.05 otherwise it is non-significant (n.s.).

A.2 Figures

This section contains additional figures for Sec. 3.1-3.2 .

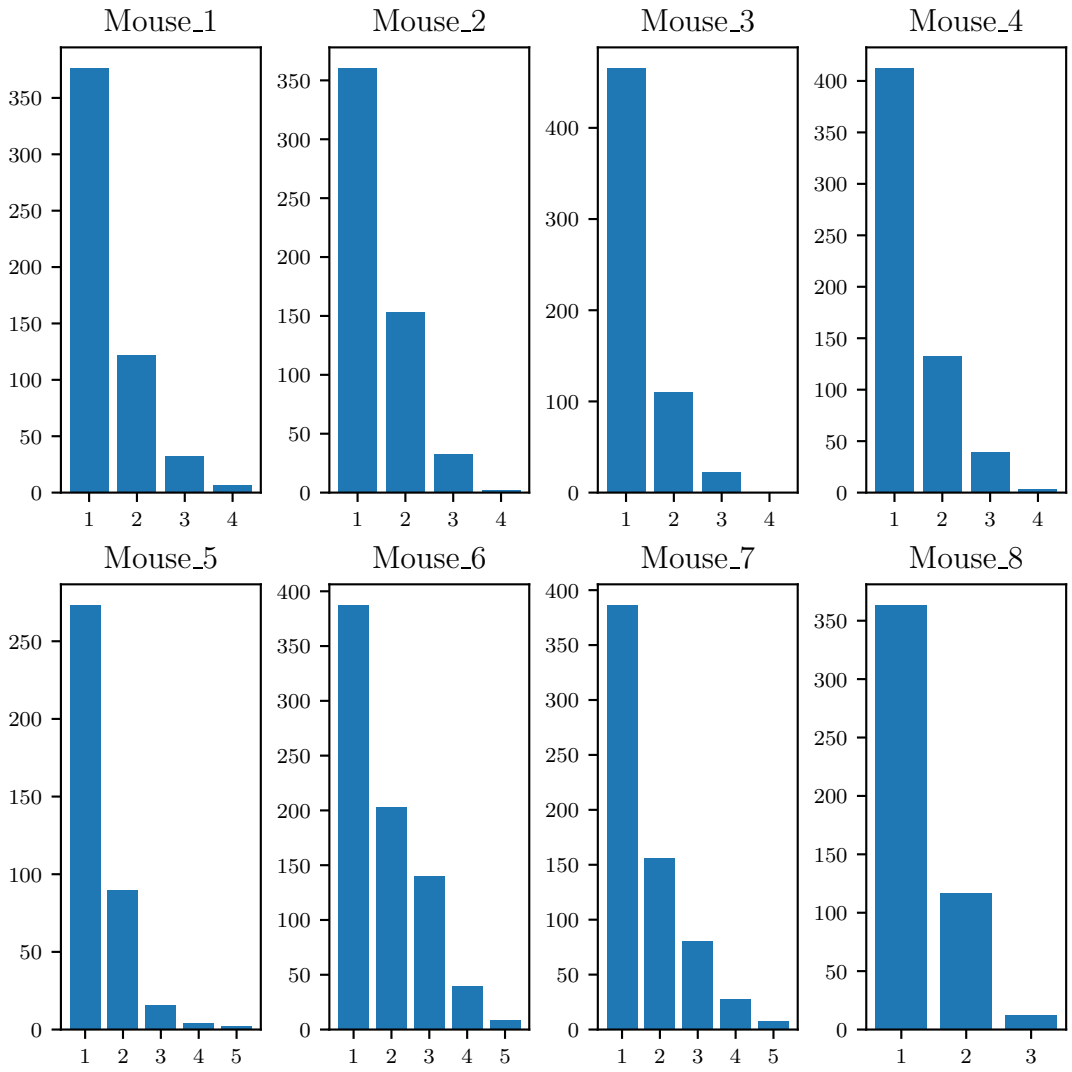


FIGURE A.3: Discrete movements for the mice in the u-maze (not counting static time steps).

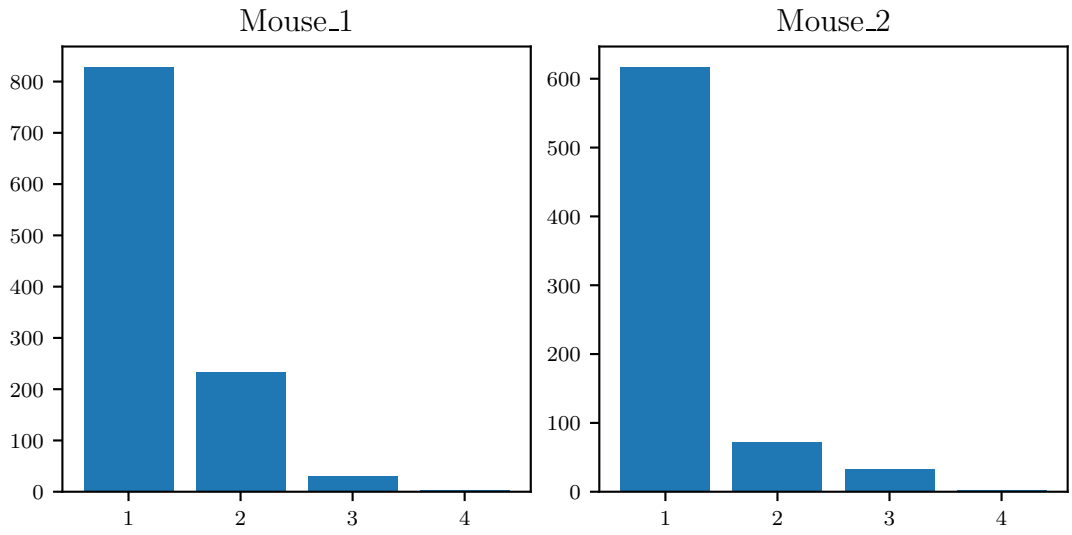


FIGURE A.4: Discrete movements for the mice in the squared open-maze (not counting static time steps).

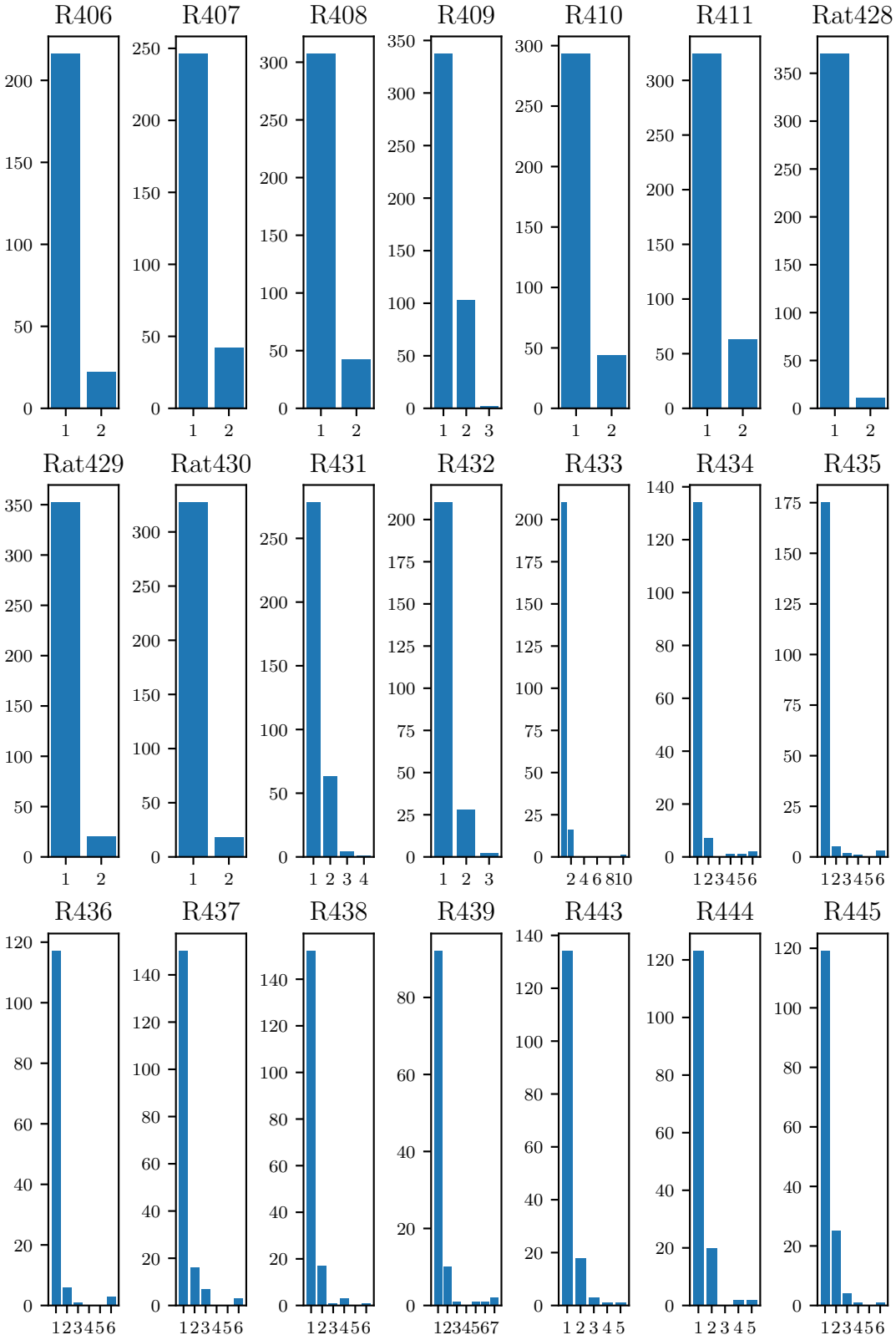


FIGURE A.5: Discrete movements for the rat in the grid-maze (not counting static time steps).

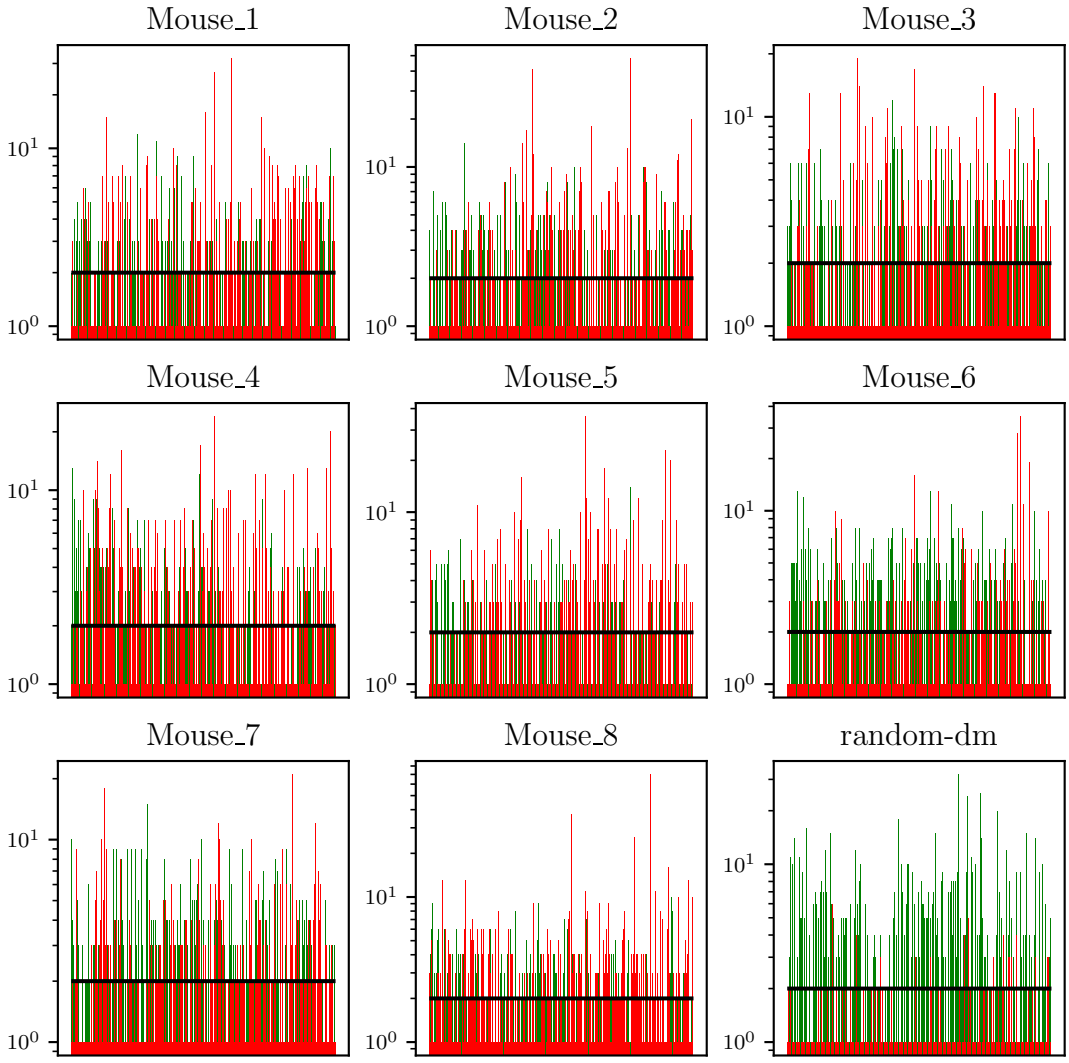


FIGURE A.6: Duration of the moving (green) and static (red) bouts for all the mice in the u-maze. The black line indicates the median value of all the bouts' heights.

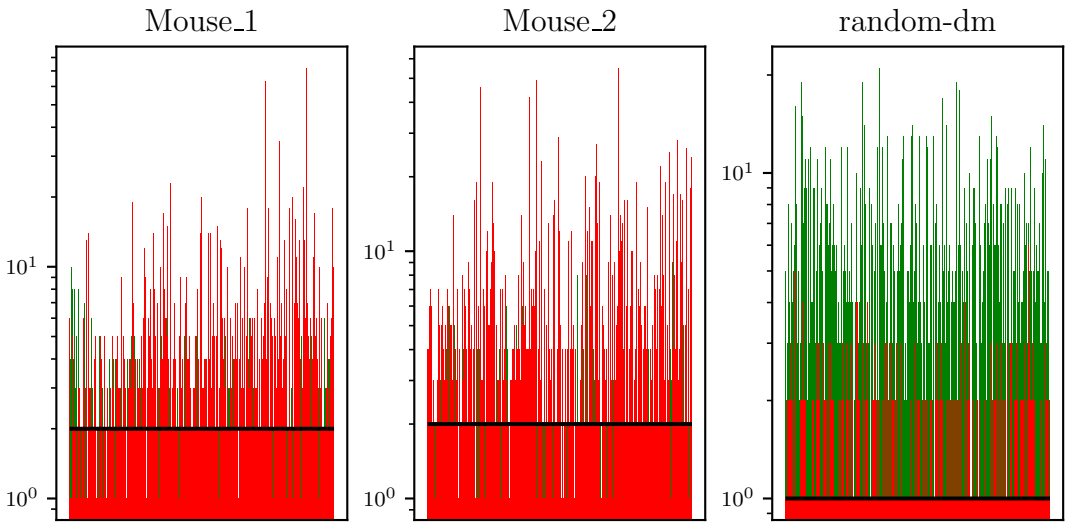


FIGURE A.7: Duration of the moving (green) and static (red) bouts for all the mice in the squared open-maze. The black line indicates the median value of all the bouts' heights.

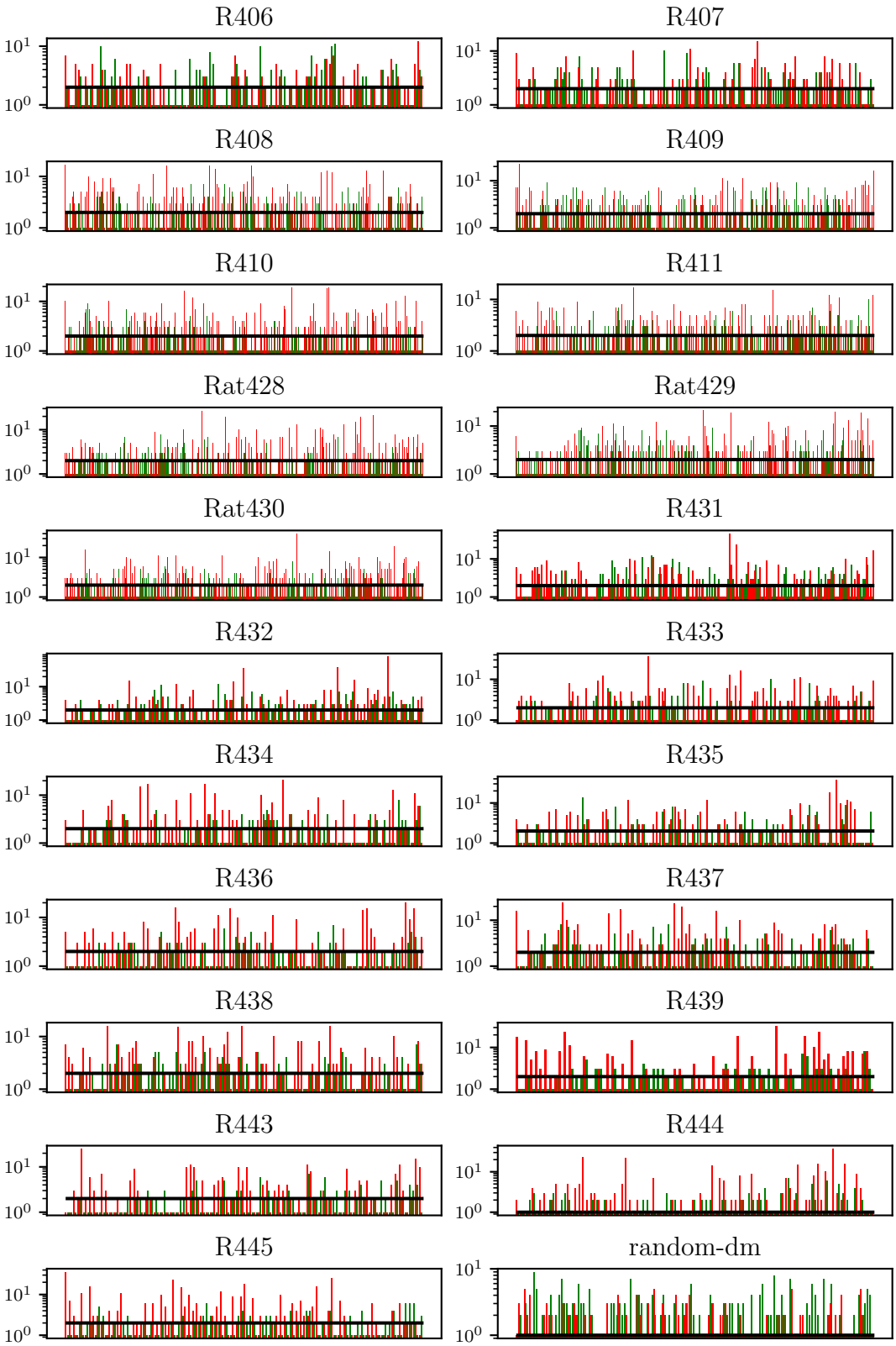
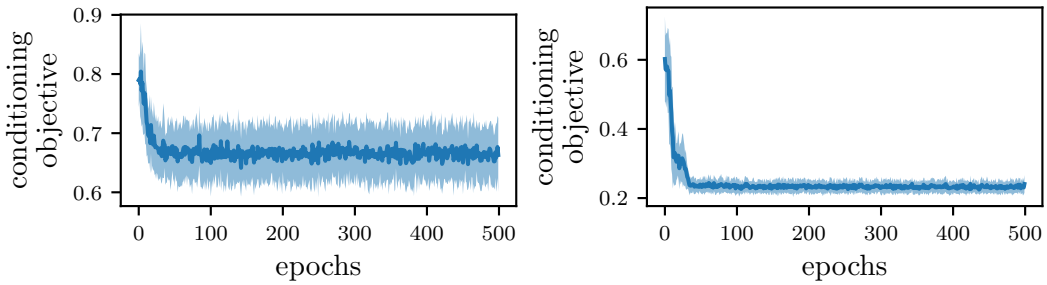
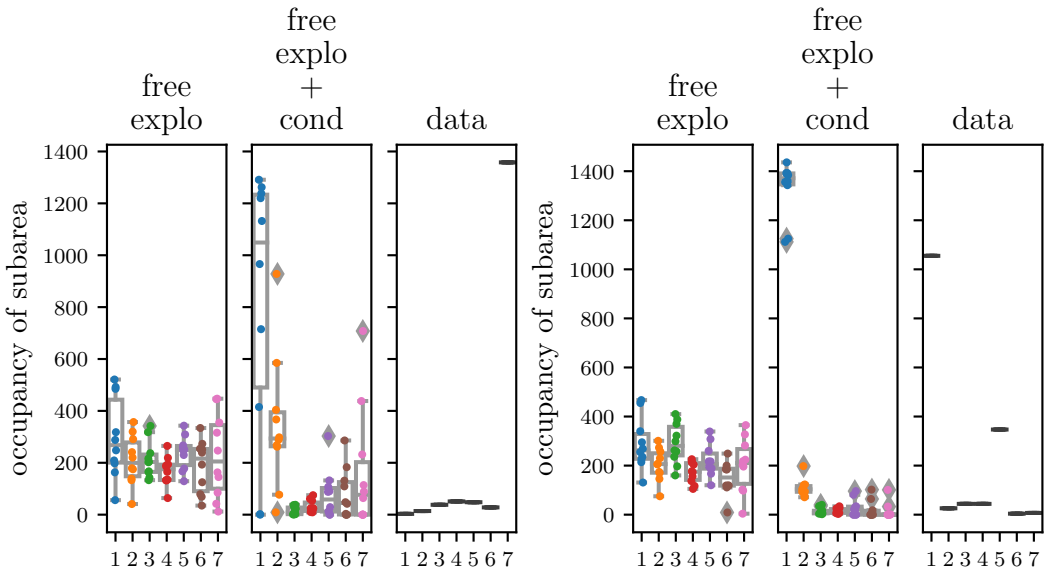


FIGURE A.8: Duration of the moving (green) and static (red) bouts for all the rats in the grid-maze. The black line indicates the median value of all the bouts' heights.



(A) Aversive stimulation

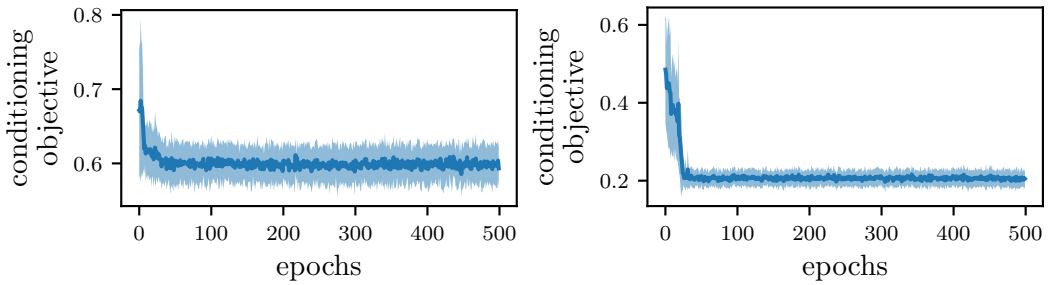
(B) Positive stimulation



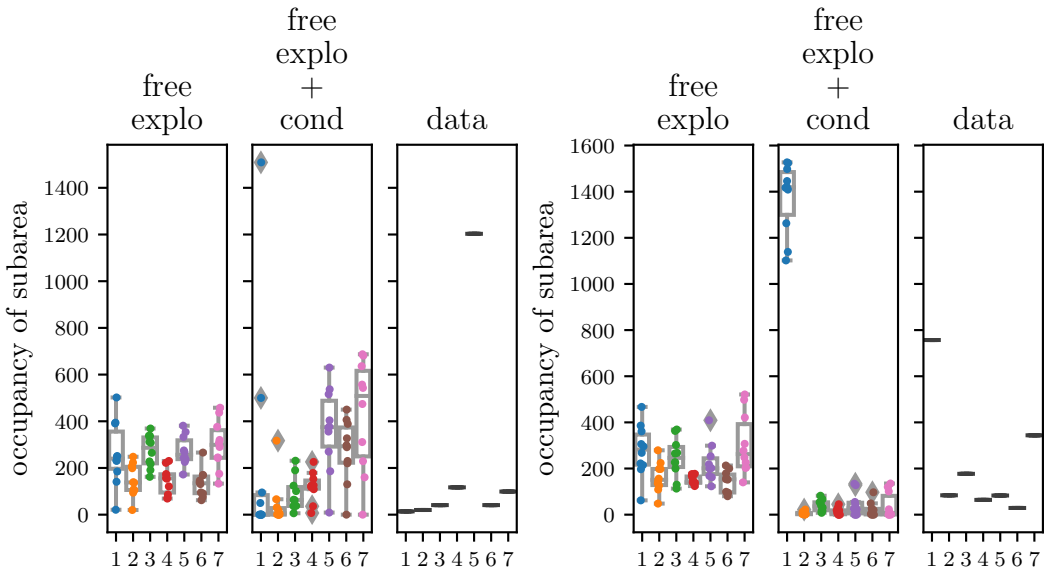
(C) Negative stimulation

(D) Positive stimulation

FIGURE A.9: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1117.

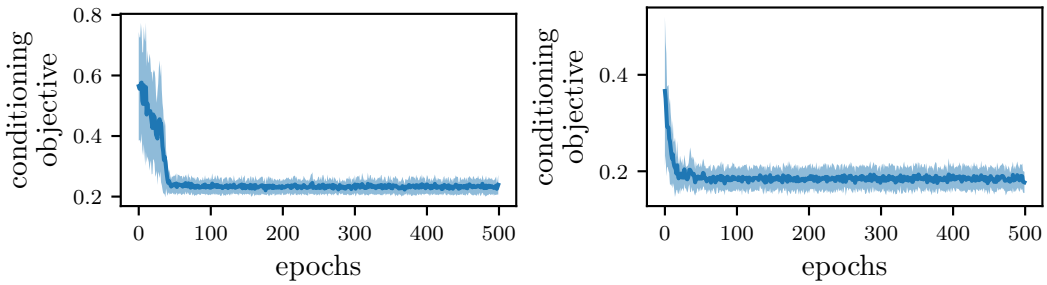


(A) Aversive stimulation (B) Positive stimulation



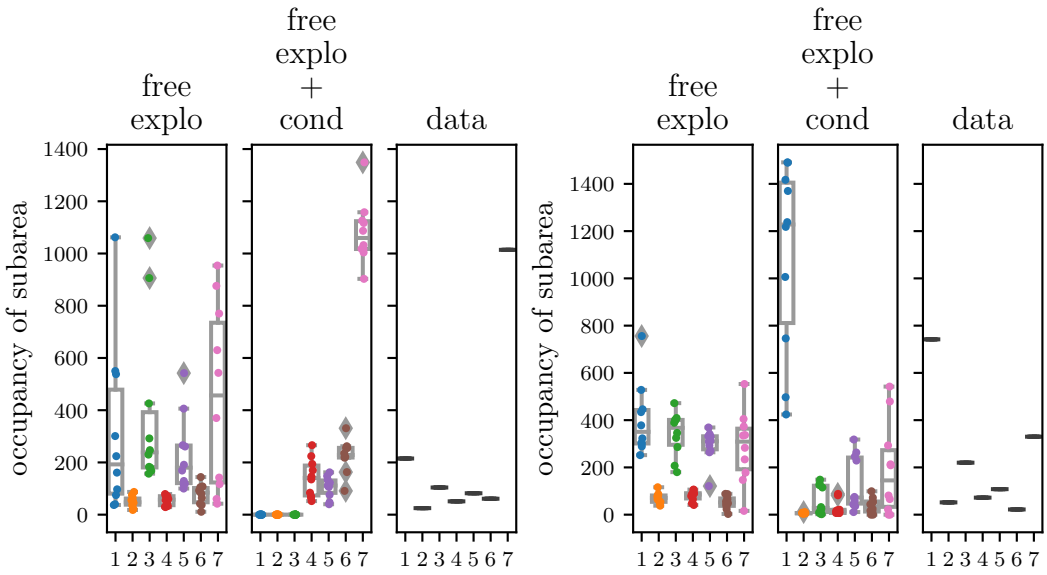
(C) Negative stimulation (D) Positive stimulation

FIGURE A.10: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1161.



(A) Aversive stimulation

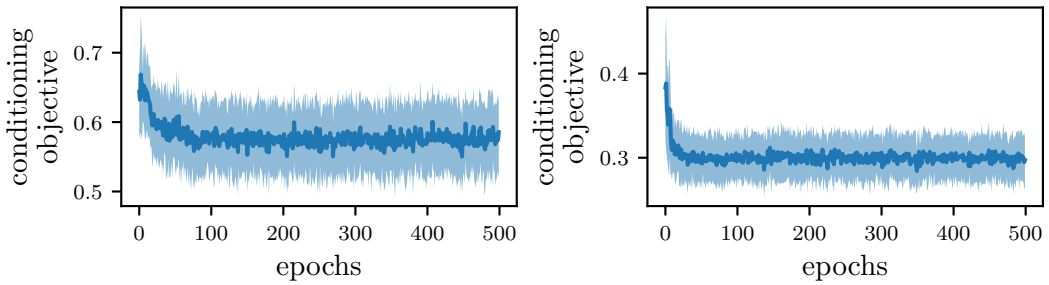
(B) Positive stimulation



(C) Negative stimulation

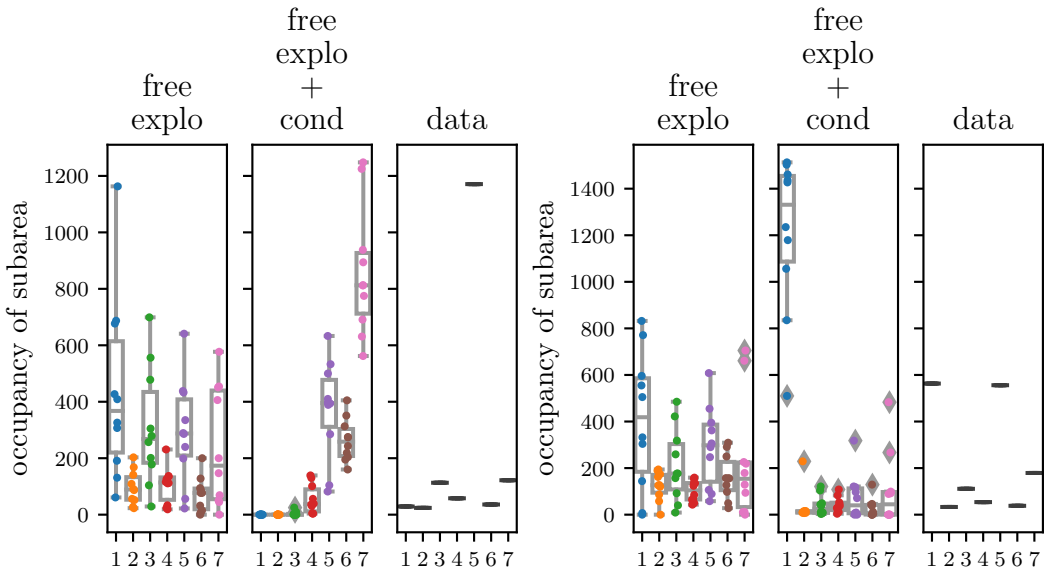
(D) Positive stimulation

FIGURE A.11: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1162.



(A) Aversive stimulation

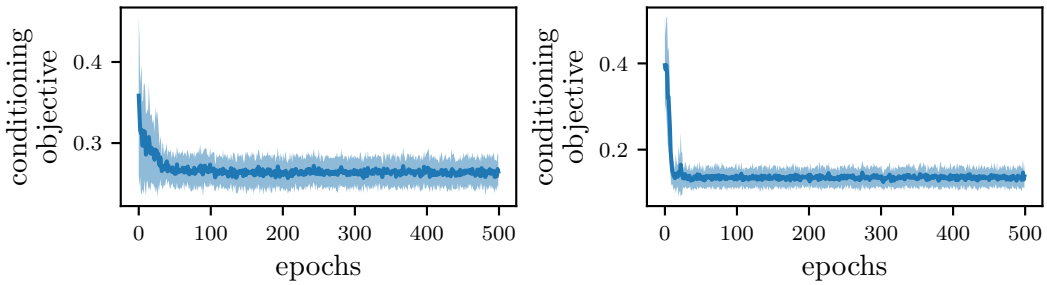
(B) Positive stimulation



(C) Negative stimulation

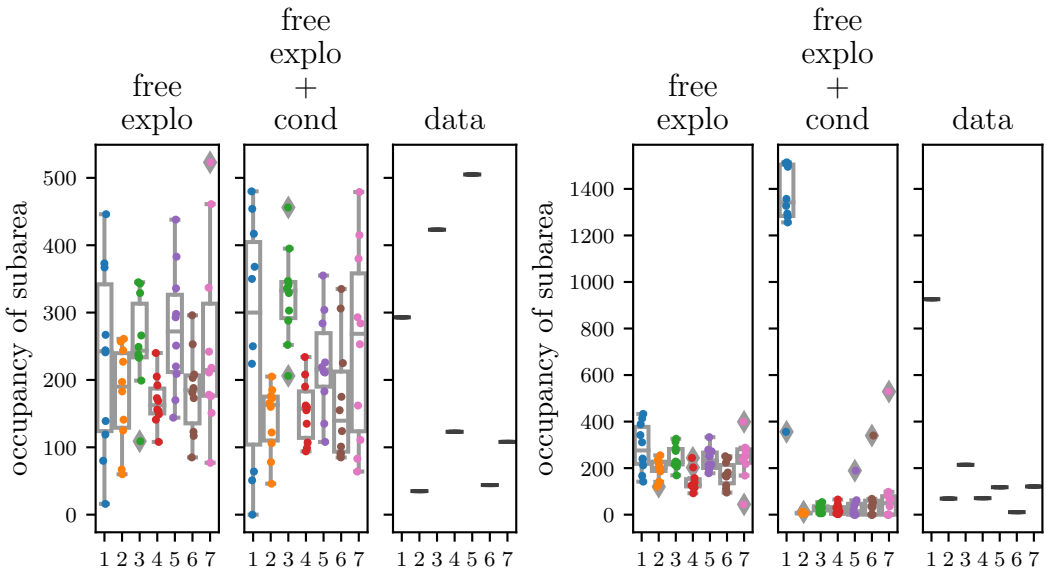
(D) Positive stimulation

FIGURE A.12: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1182.



(A) Aversive stimulation

(B) Positive stimulation



(C) Negative stimulation

(D) Positive stimulation

FIGURE A.13: Optimized exploration model (free explo + cond) in comparison to the previously optimized free exploration model (free explo), and to the data; example for Mouse1199.

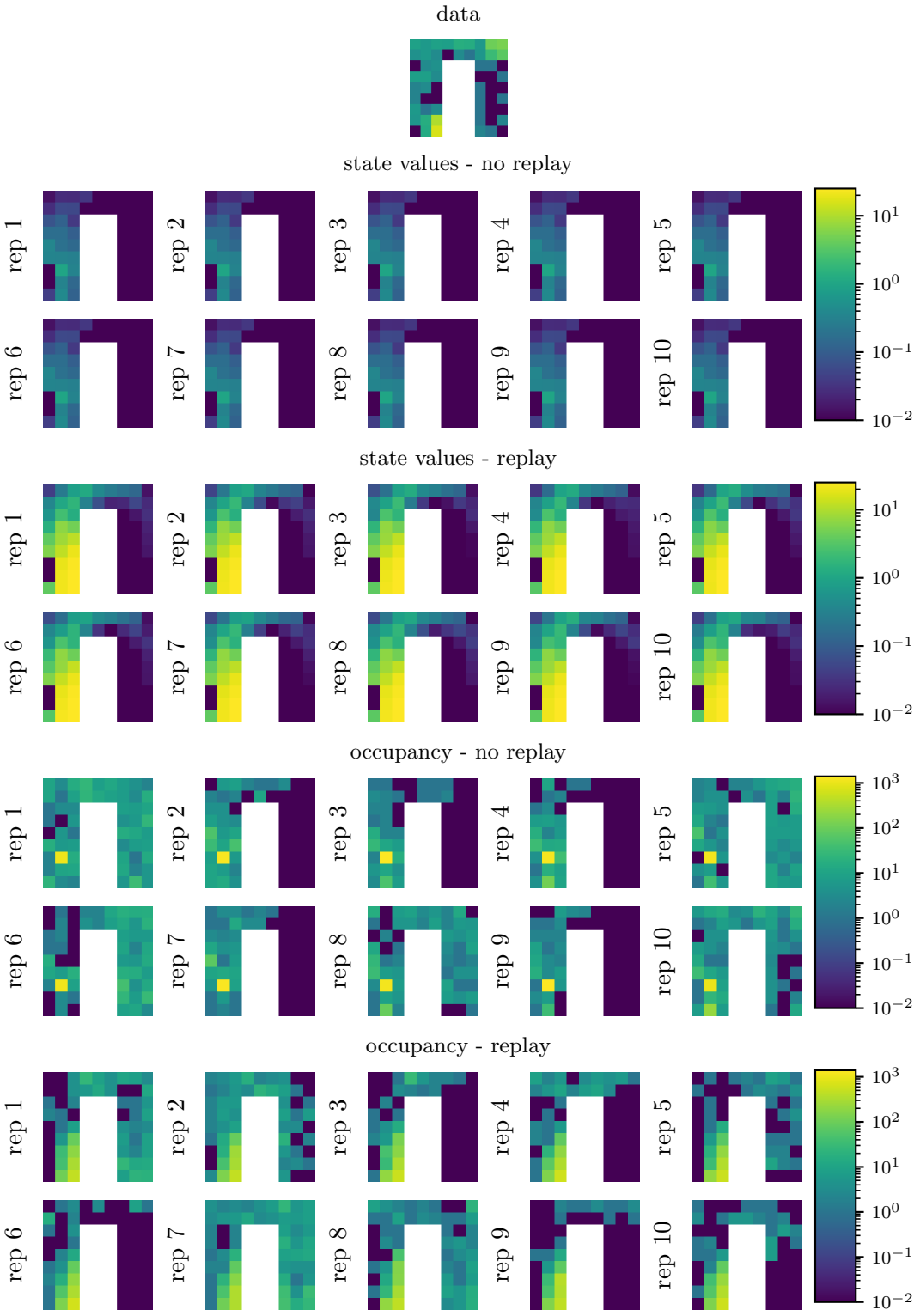


FIGURE A.14: Mouse1117 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

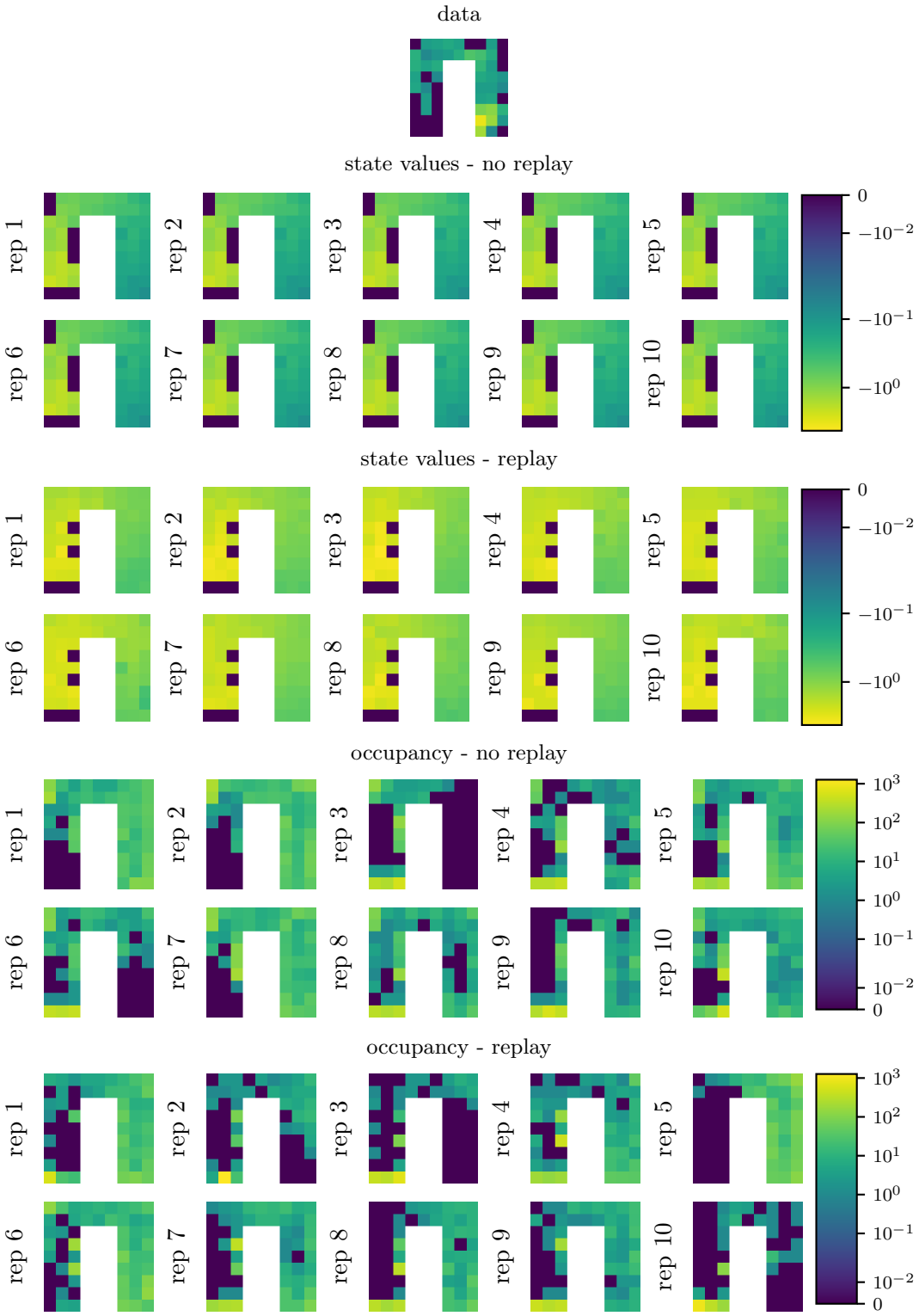


FIGURE A.15: Mouse1117 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

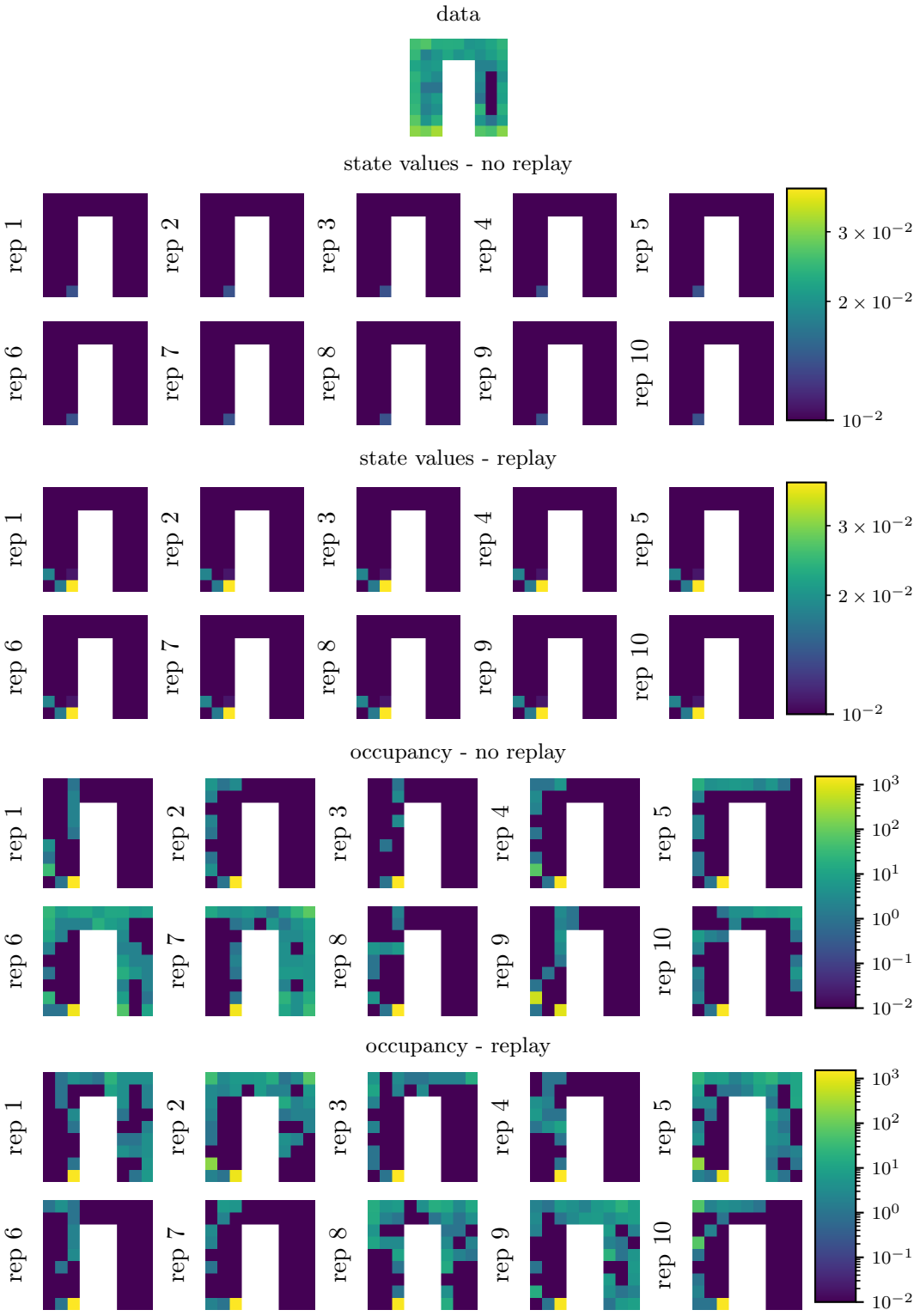


FIGURE A.16: Mouse1161 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

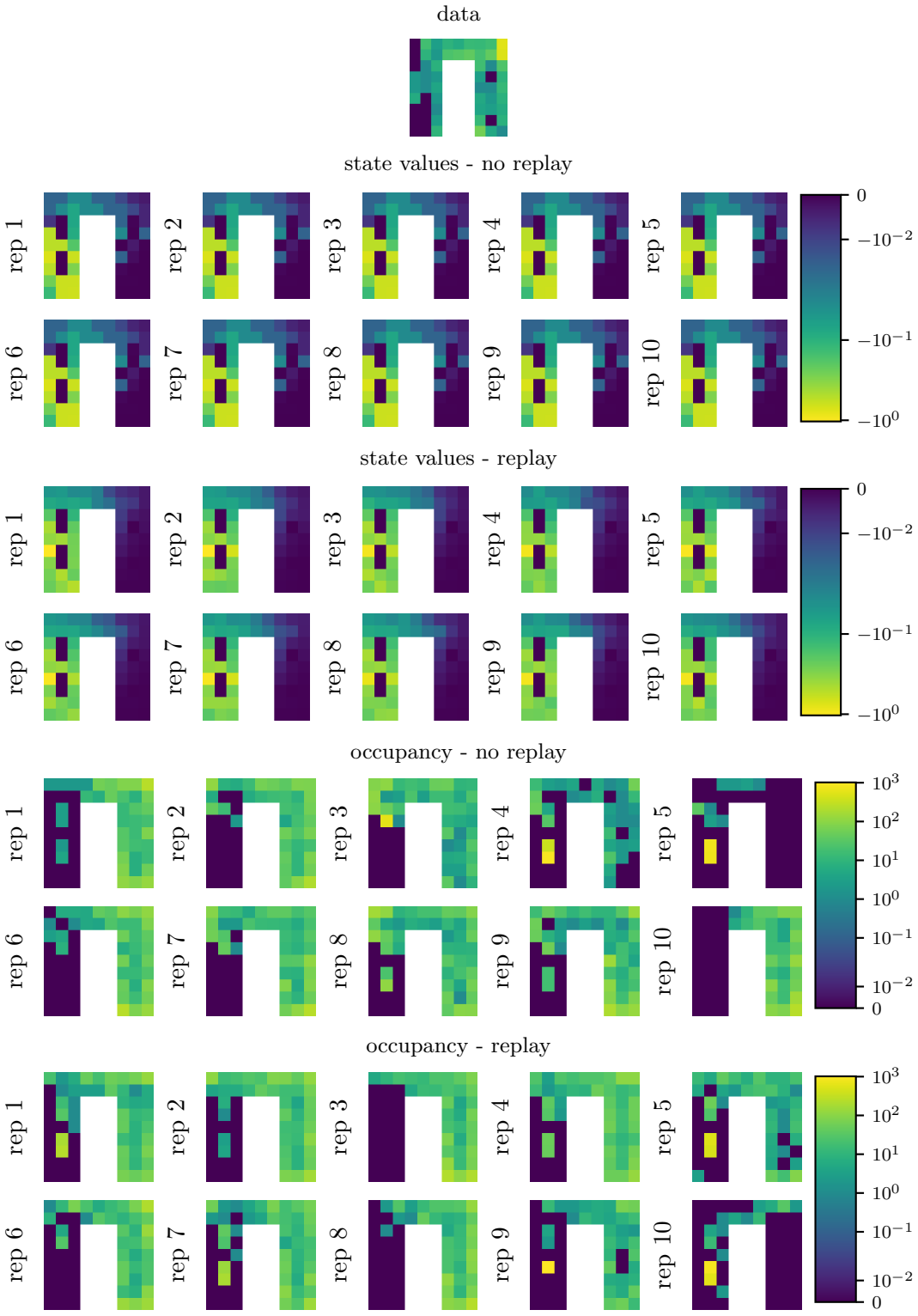


FIGURE A.17: Mouse1161 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

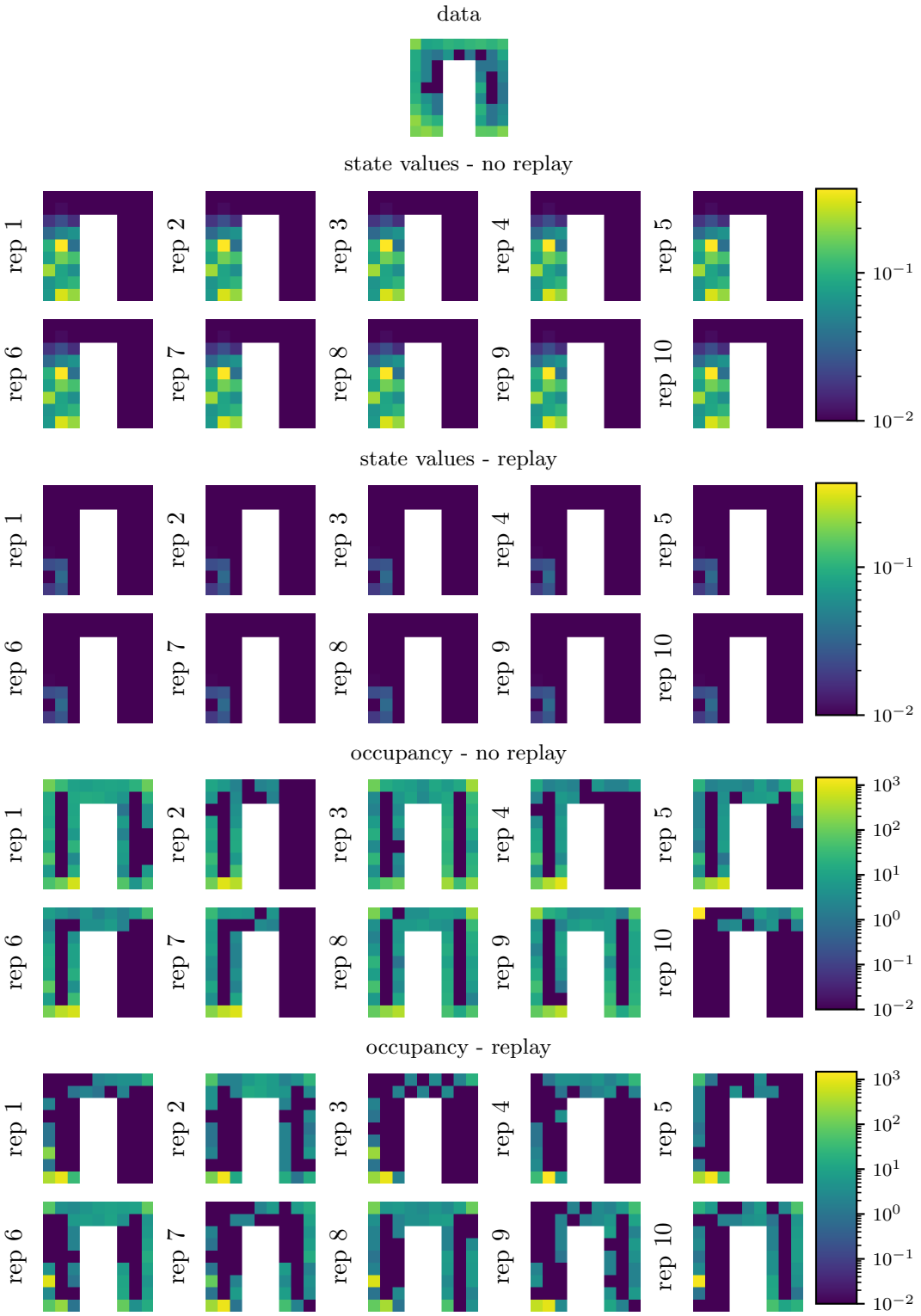


FIGURE A.18: Mouse1162 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

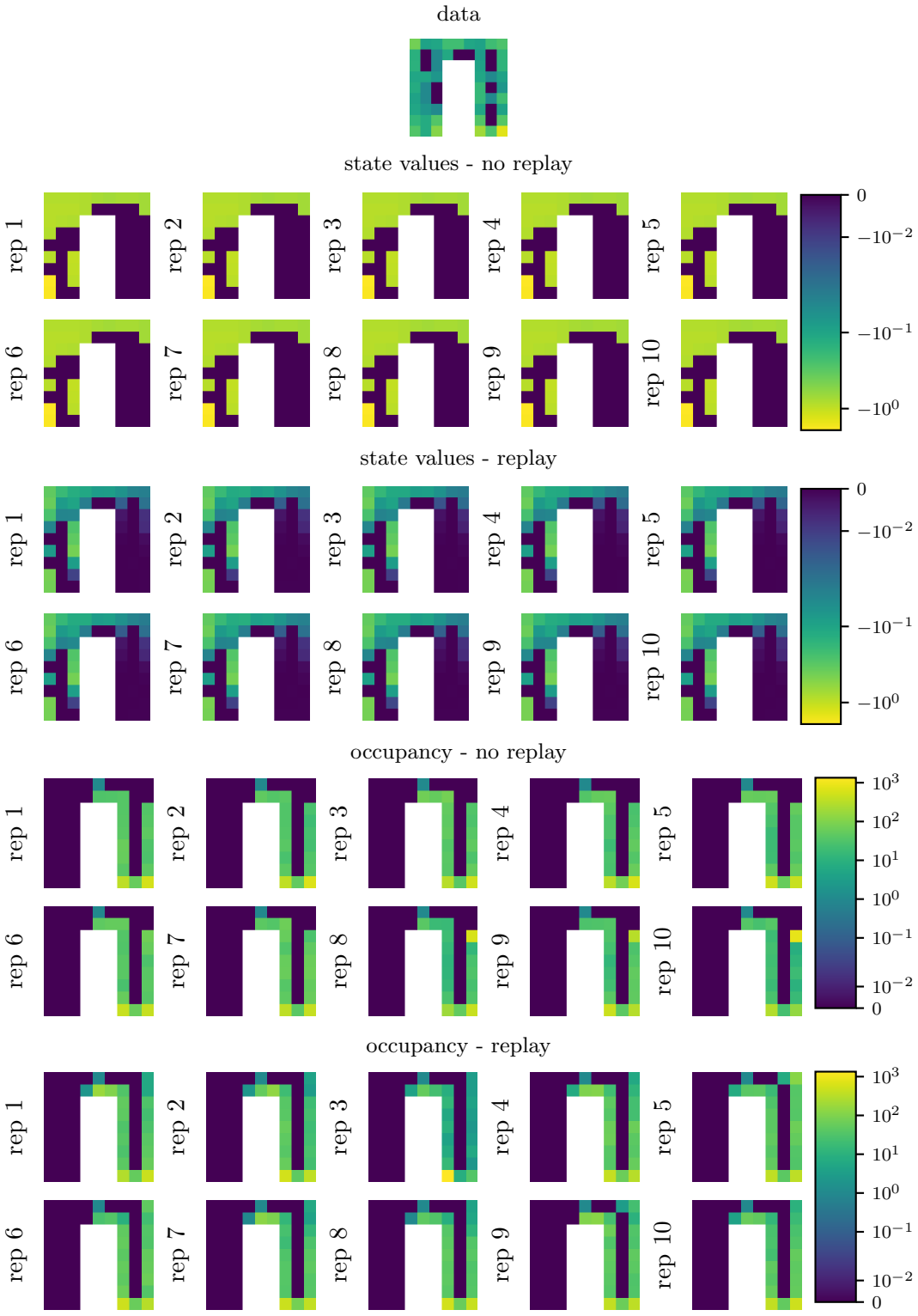


FIGURE A.19: Mouse1162 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

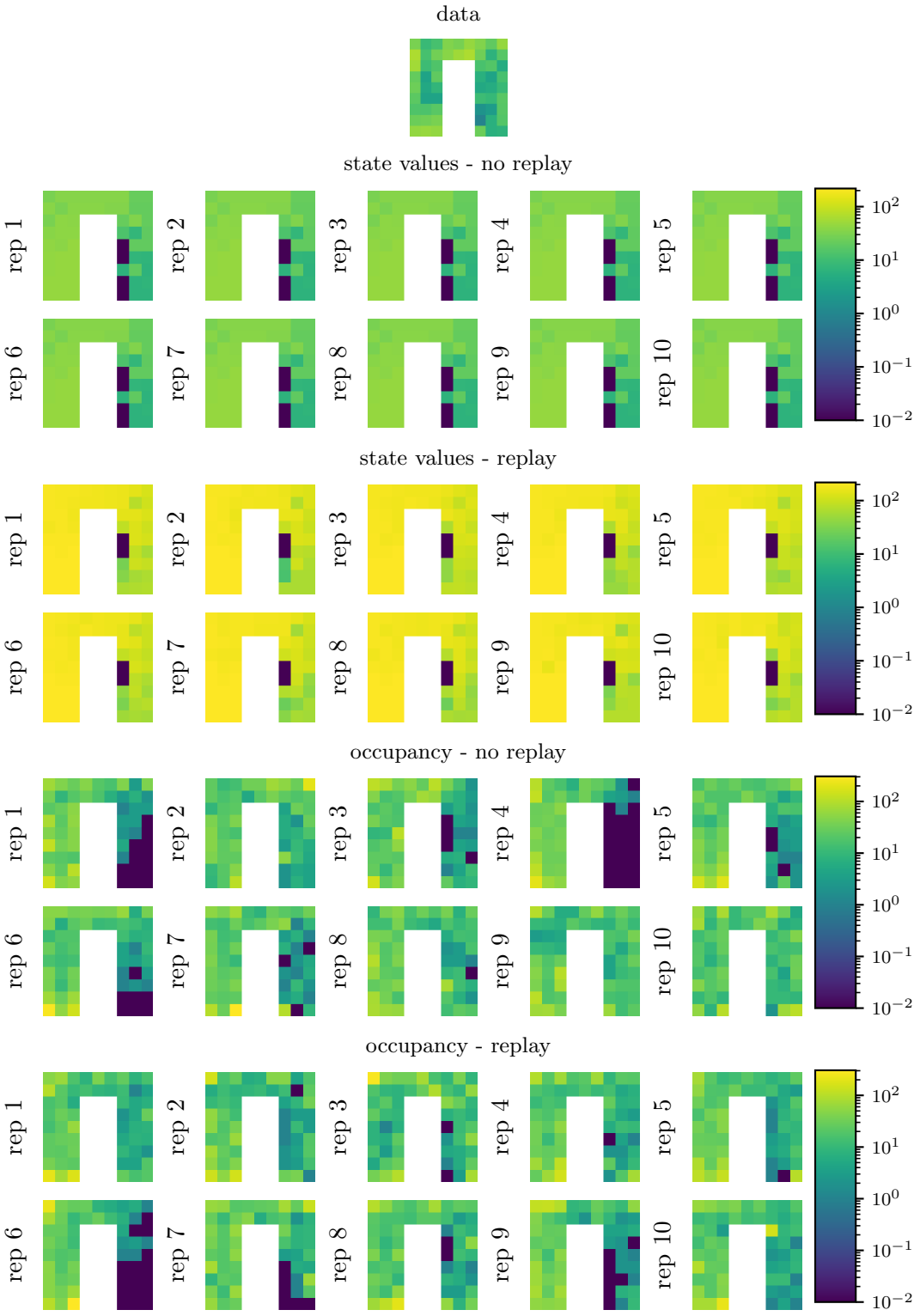


FIGURE A.20: Mouse1168 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

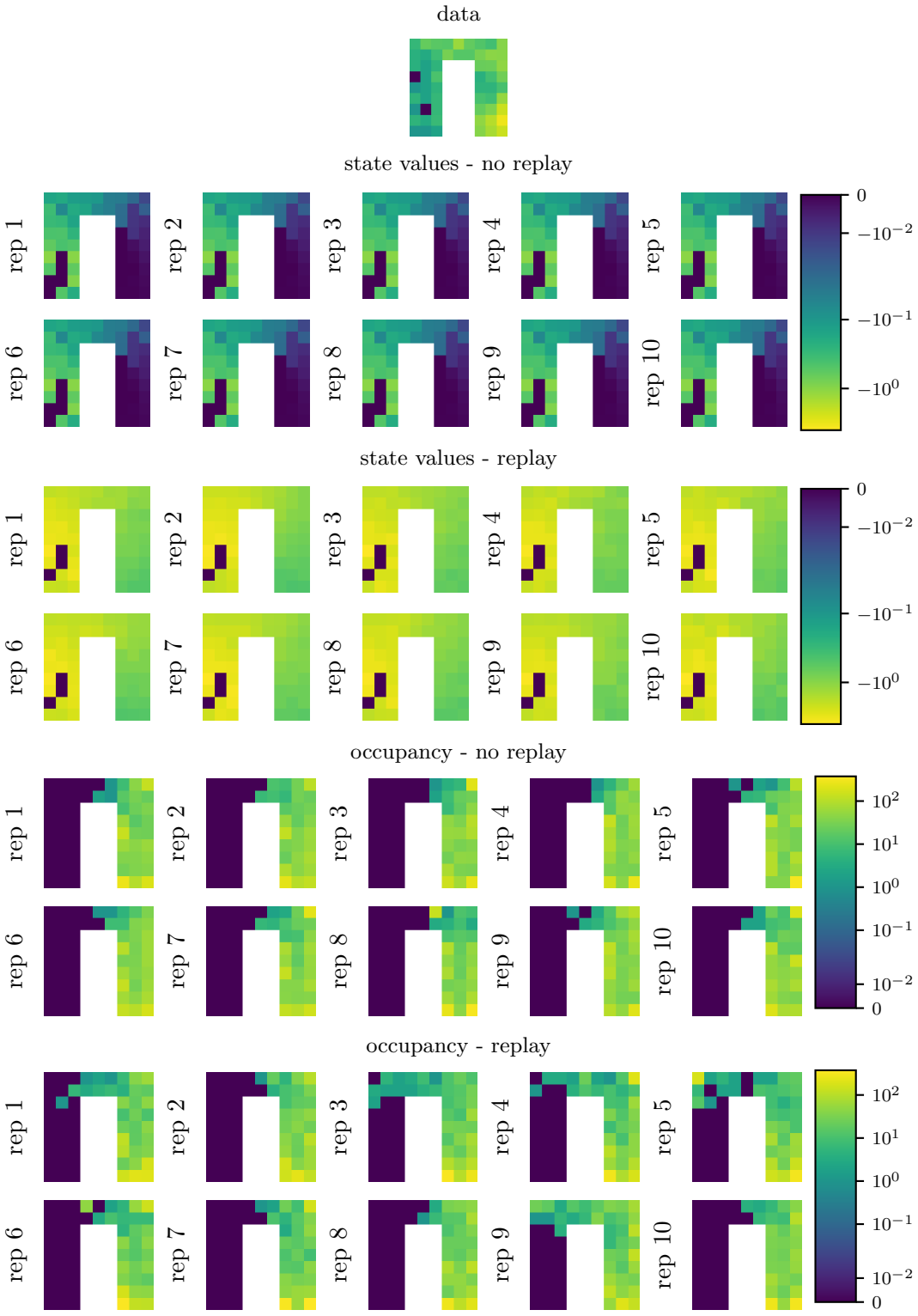


FIGURE A.21: Mouse1168 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

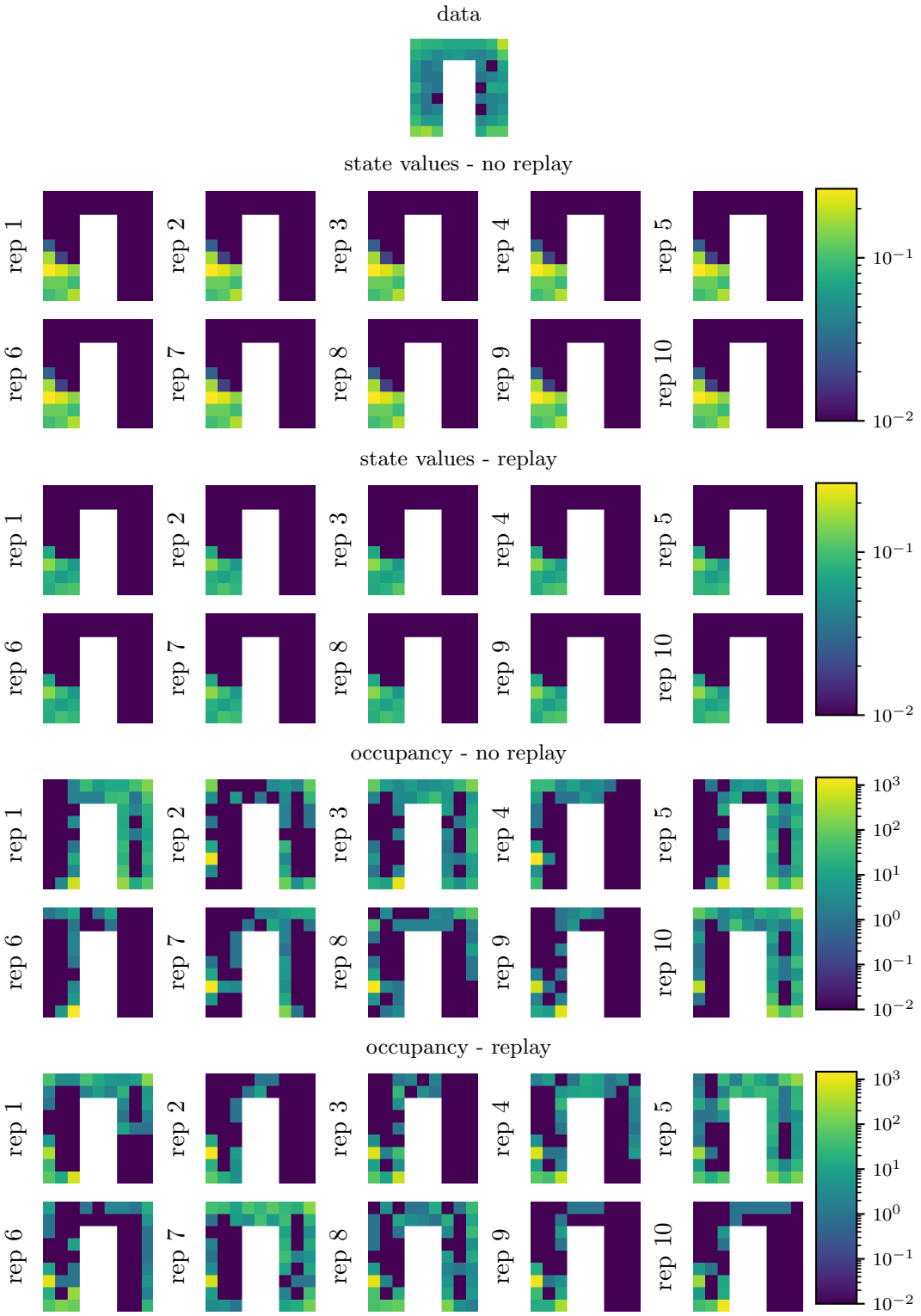


FIGURE A.22: Mouse1182 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

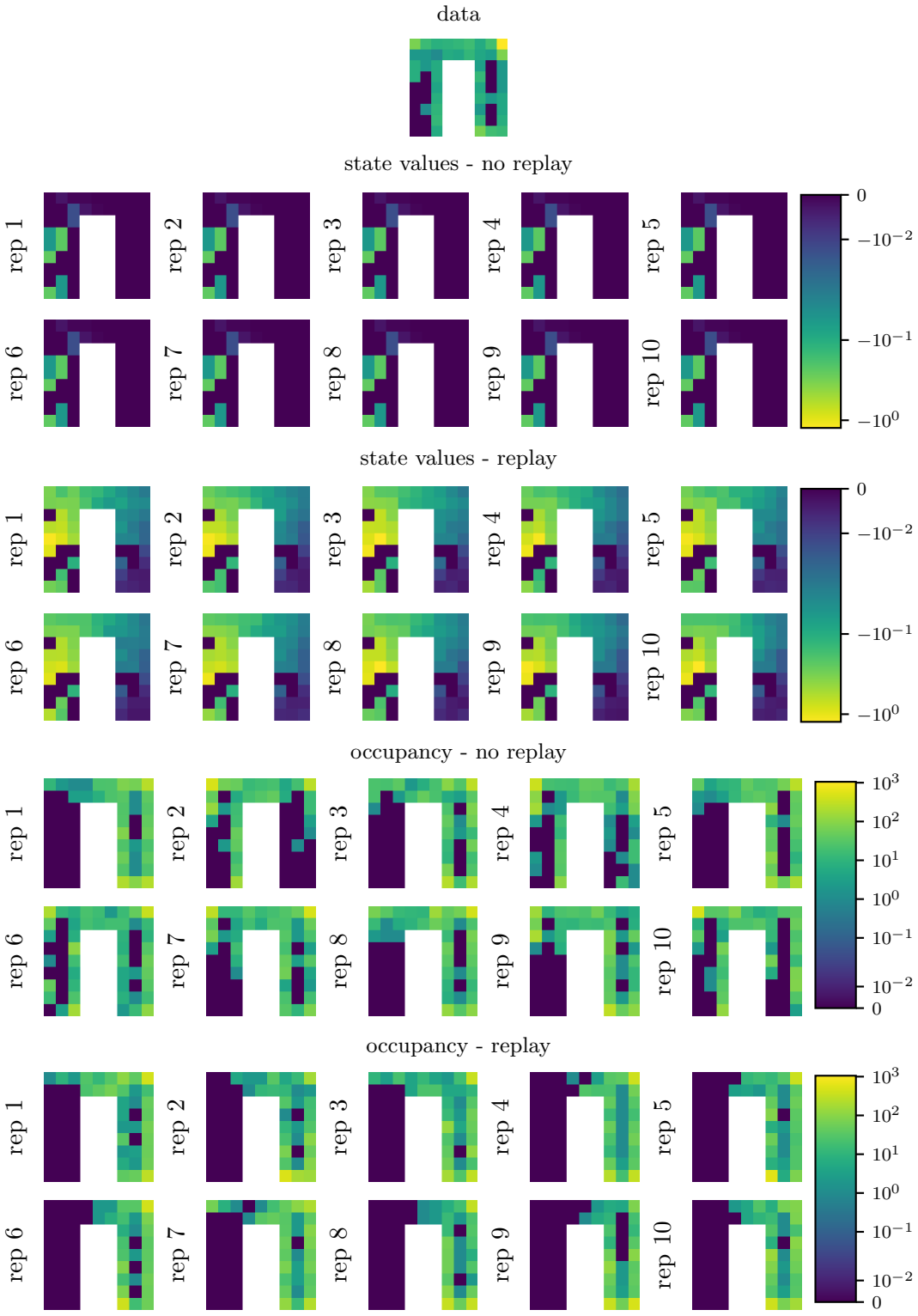


FIGURE A.23: Mouse1182 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the states' values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

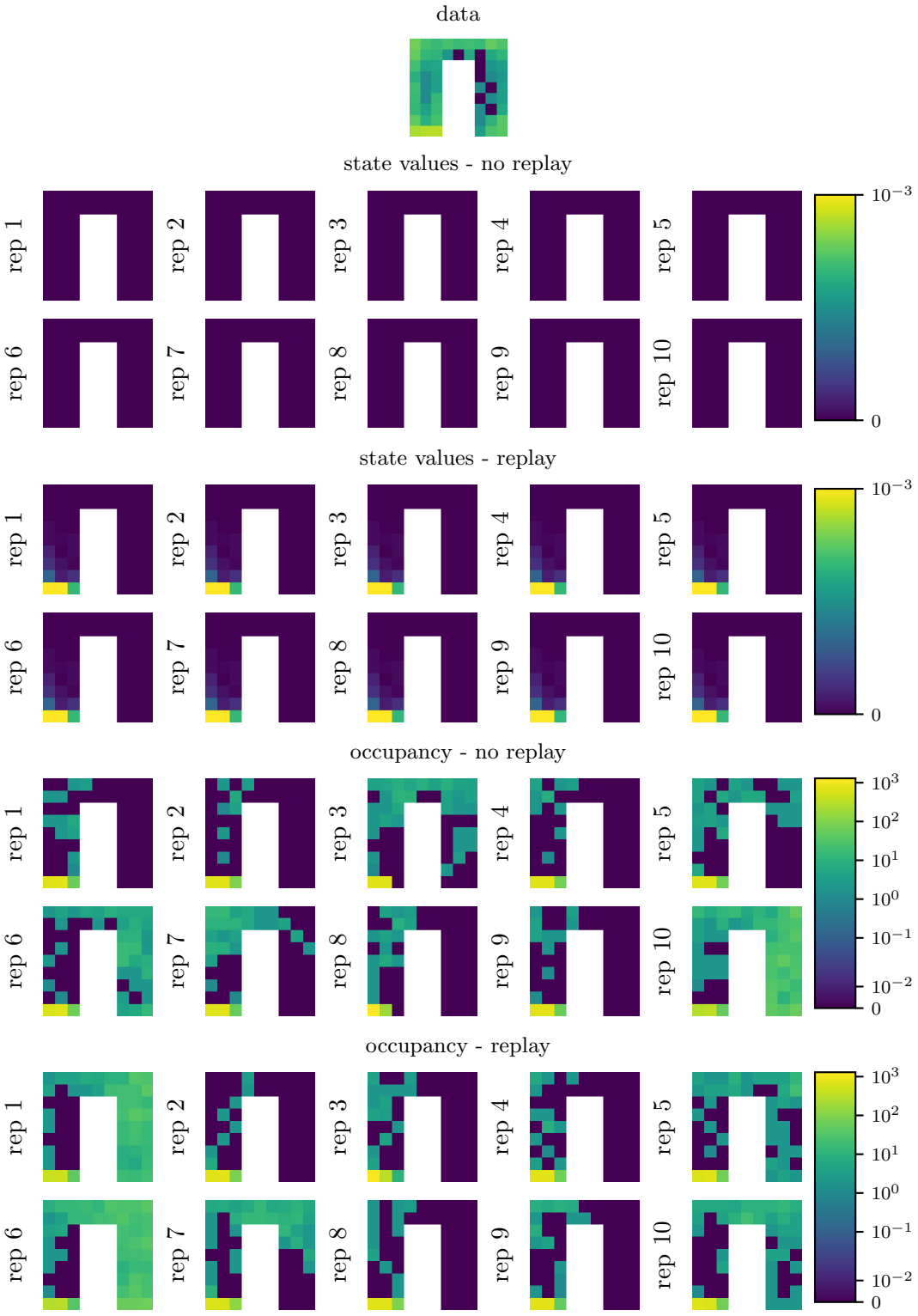


FIGURE A.24: Mouse1199 - positive conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

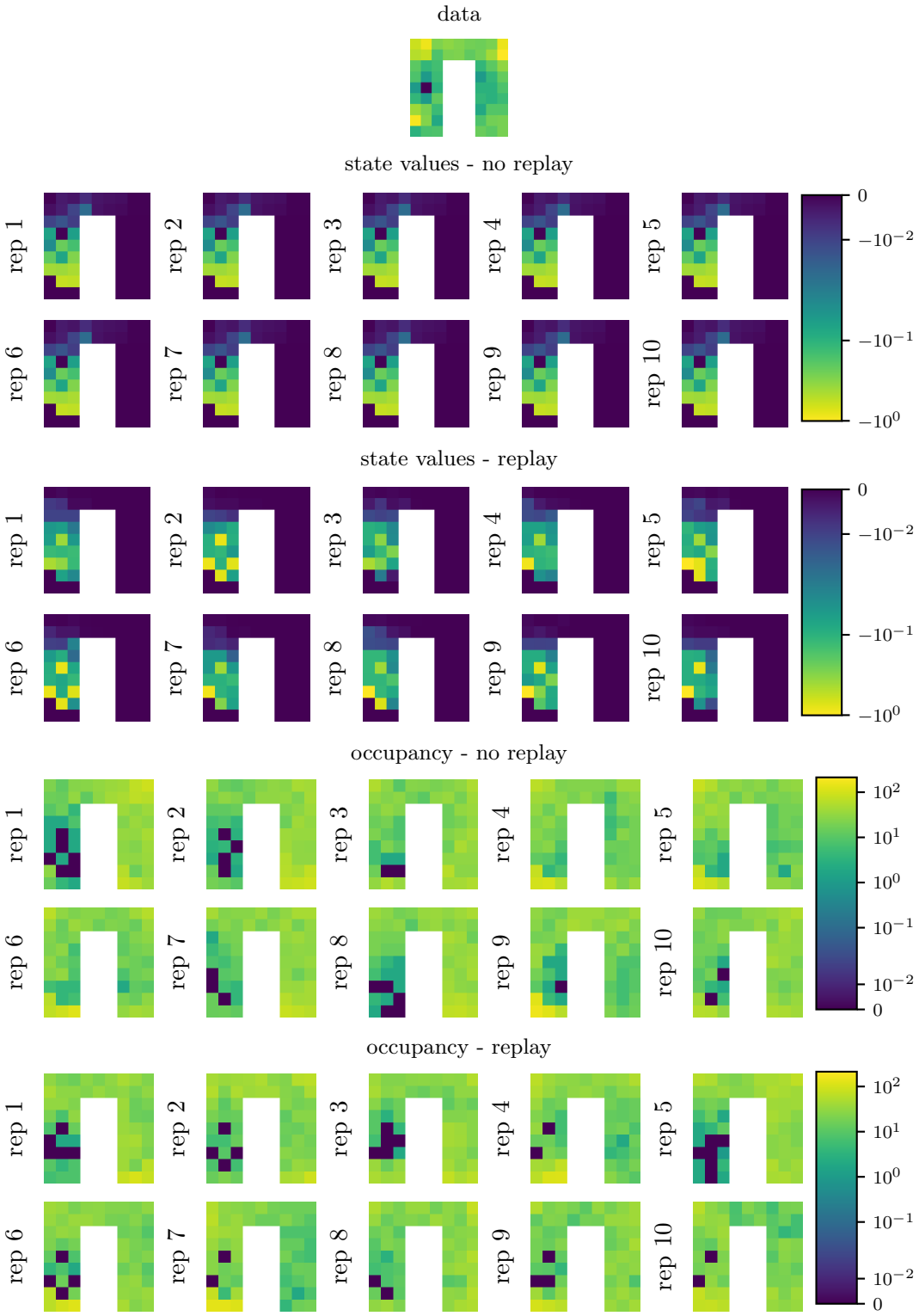


FIGURE A.25: Mouse1199 - negative conditioning. Top-down, the figure shows in logarithmic scale the post-conditioning occupancy of the data, the state's values for the optimized exploration model (without and with replay), and the post-conditioning occupancy of the maze for the optimized exploration model (without and with replay).

A.3 Tables

This section contains additional tables for Sec. 3.2.2.

		positive	negative
occupancy	1	0.041	0.015
	2	0.59	0.025
	3	0.94	0.31
	4	0.87	0.94
	5	0.75	1.0
	6	0.093	0.24
	7	0.026	0.94

TABLE A.1: Comparative statistical analysis between the occupancy of the seven sub-areas of the maze (Fig. 3.21a) in the pre- and post-conditioning data. For each dataset (u-maze positive and u-maze positive) and for each distribution, corresponding to the bins of the occupation of the seven sub-areas, a Wilcoxon-Mann-Whitney comparison test is performed. Here, we report the p-values for each comparison, and the blue gradient decreasingly shows non-significant statistical difference (dark blue) and statistical difference; p-values < 0.05 (medium blue) and p-values < 0.001 (light blue).

Bibliography

Abraham, A. and L. Jain (2005). "Evolutionary multiobjective optimization". In: *Evolutionary multiobjective optimization*. Springer, pp. 1–6.

Adams, C. D. (1982). "Variations in the sensitivity of instrumental responding to reinforcer devaluation". In: *The Quarterly Journal of Experimental Psychology Section B* 34.2b, pp. 77–98.

Alkon, D. L., D. G. Amaral, M. F. Bear, J. Black, T. J. Carew, N. J. Cohen, J. F. Disterhoft, H. Eichenbaum, S. Golski, L. K. Gorman, et al. (1991). "Learning and memory". In: *Brain research reviews* 16.2, pp. 193–220.

Amorapanth, P., K. Nader, and J. E. LeDoux (1999). "Lesions of periaqueductal gray dissociate-conditioned freezing from conditioned suppression behavior in rats". In: *Learning & Memory* 6.5, pp. 491–499.

Arleo, A. and W. Gerstner (2000). "Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity". In: *Biological cybernetics* 83.3, pp. 287–299.

Atherton, L. A., D. Dupret, and J. R. Mellor (2015). "Memory trace replay: the shaping of memory consolidation by neuromodulation". In: *Trends in neurosciences* 38.9, pp. 560–570.

Aubin, L., M. Khamassi, and B. Girard (2018). "Prioritized sweeping neural DynaQ with multiple predecessors, and hippocampal replays". In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 16–27.

Auger, A. and N. Hansen (2005). "A restart CMA evolution strategy with increasing population size". In: *2005 IEEE congress on evolutionary computation*. Vol. 2. IEEE, pp. 1769–1776.

Aulinas, J., Y. Petillot, J. Salvi, and X. Lladó (2008). "The SLAM problem: a survey". In: *Artificial Intelligence Research and Development*, pp. 363–371.

Bachem, O., M. Lucic, S. H. Hassani, and A. Krause (2016). "Fast and Provably Good Seedings for k-Means." In: *Nips*, pp. 55–63.

Bäck, T. and H.-P. Schwefel (1993). "An overview of evolutionary algorithms for parameter optimization". In: *Evolutionary computation* 1.1, pp. 1–23.

Balcombe, J. P. (2006). "Laboratory environments and rodents' behavioural needs: a review". In: *Laboratory animals* 40.3, pp. 217–235.

Balleine, B. W., M. R. Delgado, and O. Hikosaka (2007). "The role of the dorsal striatum in reward and decision-making". In: *Journal of Neuroscience* 27.31, pp. 8161–8165.

Bard, Y. (1974). *Nonlinear parameter estimation*. Tech. rep.

Barnes, C. (1988). "Aging and the physiology of spatial memory". In: *Neurobiology of aging* 9, pp. 563–568.

Barto, A. G., S. J. Bradtke, and S. P. Singh (1995). "Learning to act using real-time dynamic programming". In: *Artificial intelligence* 72.1-2, pp. 81–138.

Baxter, M. G. and E. A. Murray (2002). "The amygdala and reward". In: *Nature reviews neuroscience* 3.7, pp. 563–573.

Beck, J. V. and K. J. Arnold (1977). *Parameter estimation in engineering and science*. James Beck.

Behrens, T. E., T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson (2018). "What is a cognitive map? Organizing knowledge for flexible behavior". In: *Neuron* 100.2, pp. 490–509.

Bellman, R. (1957). "A Markovian decision process". In: *Journal of mathematics and mechanics*, pp. 679–684.

Belzung, C. (1999). "Chapter 4.11 Measuring rodent exploratory behavior". In: *Handbook of Molecular-Genetic Techniques for Brain and Behavior Research*. Ed. by W. Crusio and R. Gerlai. Vol. 13. Techniques in the Behavioral and Neural Sciences. Elsevier, pp. 738–749.

Benchenane, K., A. Peyrache, M. Khamassi, P. L. Tierney, Y. Gioanni, F. P. Battaglia, and S. I. Wiener (2010). "Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning". In: *Neuron* 66.6, pp. 921–936.

Benjamini, Y., E. Fonio, T. Galili, G. Z. Havkin, and I. Golani (2011). "Quantifying the buildup in extent and complexity of free exploration in mice". In: *Proceedings of the National Academy of Sciences* 108.supplement_3, pp. 15580–15587.

Berlyne, D. E. (1950). "Novelty and curiosity as determinants of exploratory behaviour". In: *British journal of psychology* 41.1, p. 68.

Berners-Lee, A., T. Feng, D. Silva, X. Wu, E. R. Ambrose, B. E. Pfeiffer, and D. J. Foster (2022). "Hippocampal replays appear after a single experience and incorporate greater detail with more experience". In: *Neuron* 110.11, pp. 1829–1842.

Berthoz, A. (1991). "Reference frames for the perception and control of movement." In.

Bhatnagar, S., M. Ghavamzadeh, M. Lee, and R. S. Sutton (2007). "Incremental natural actor-critic algorithms". In: *Advances in neural information processing systems* 20.

Blanchard, D. C., G. Griebel, and R. J. Blanchard (2001). "Mouse defensive behaviors: pharmacological and behavioral assays for anxiety and panic". In: *Neuroscience & Biobehavioral Reviews* 25.3, pp. 205–218.

Brown, G. R. and C. Nemes (2008). "The exploratory behaviour of rats in the hole-board apparatus: is head-dipping a valid measure of neophilia?" In: *Behavioural processes* 78.3, pp. 442–448.

Brown, M. A. and P. E. Sharp (1995). "Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens". In: *Hippocampus* 5.3, pp. 171–188.

Bryzgalov, D. (Sept. 2021). "Hippocampal reactivations after aversive or rewarding experience : classical and deep learning approaches". Theses. Université Paris sciences et lettres.

Burgard, W., D. Fox, H. Jans, C. Matenar, and S. Thrun (1999). "Sonar-based mapping with mobile robots using EM". In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. MORGAN KAUFMANN PUBLISHERS, INC., pp. 67–76.

Buzsáki, G. (1989). "Two-stage model of memory trace formation: a role for "noisy" brain states". In: *Neuroscience* 31.3, pp. 551–570.

Buzsáki, G. and E. I. Moser (2013). "Memory, navigation and theta rhythm in the hippocampal-entorhinal system". In: *Nature neuroscience* 16.2, pp. 130–138.

Caluwaerts, K., M. Staffa, S. N'Guyen, C. Grand, L. Dollé, A. Favre-Félix, B. Girard, and M. Khamassi (2012). "A biologically inspired meta-control navigation system for the psikharpax rat robot". In: *Bioinspiration & biomimetics* 7.2, p. 025009.

Caluwaerts, K., A. Favre-Félix, M. Staffa, S. N'Guyen, C. Grand, B. Girard, and M. Khamassi (2012). "Neuro-inspired navigation strategies shifting for robots: Integration of a multiple landmark taxon strategy". In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 62–73.

Canteras, N. S. and M. Goto (1999). "Fos-like immunoreactivity in the periaqueductal gray of rats exposed to a natural predator". In: *Neuroreport* 10.2, pp. 413–418.

Cantrell, C. D. (2000). *Modern mathematical methods for physicists and engineers*. Cambridge University Press.

Carrera, J. and S. P. Neuman (1986). "Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information". In: *Water Resources Research* 22.2, pp. 199–210.

Cazé, R., M. Khamassi, L. Aubin, and B. Girard (2018). "Hippocampal replays under the scrutiny of reinforcement learning models". In: *Journal of neurophysiology* 120.6, pp. 2877–2896.

Chatila, R. and J.-P. Laumond (1985). "Position referencing and consistent world modeling for mobile robots". In: *Proceedings. 1985 IEEE International Conference on Robotics and Automation*. Vol. 2. IEEE, pp. 138–145.

Chaudhuri, R., B. Gerçek, B. Pandey, A. Peyrache, and I. Fiete (2019). "The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep". In: *Nature neuroscience* 22.9, pp. 1512–1520.

Chen, L. L., L.-H. Lin, E. J. Green, C. A. Barnes, and B. L. McNaughton (1994). "Head-direction cells in the rat posterior cortex". In: *Experimental brain research* 101.1, pp. 8–23.

Cheng, S. and L. M. Frank (2008). "New experiences enhance coordinated neural activity in the hippocampus". In: *Neuron* 57.2, pp. 303–313.

Coello, C. A. C., G. B. Lamont, D. A. Van Veldhuizen, et al. (2007). *Evolutionary algorithms for solving multi-objective problems*. Vol. 5. Springer.

Collins, A. and M. Khamassi (2021). *Initiation à la modélisation computationnelle*.

Collins, A. G. and J. Cockburn (2020). "Beyond dichotomies in reinforcement learning". In: *Nature Reviews Neuroscience* 21.10, pp. 576–586.

Cos, I., L. Canamero, G. M. Hayes, and A. Gillies (2013). "Hedonic value: Enhancing adaptation for motivated agents". In: *Adaptive Behavior* 21.6, pp. 465–483.

D'Ardenne, K., S. M. McClure, L. E. Nystrom, and J. D. Cohen (2008). "BOLD responses reflecting dopaminergic signals in the human ventral tegmental area". In: *Science* 319.5867, pp. 1264–1267.

Dan, Y. and M.-m. Poo (2004). "Spike timing-dependent plasticity of neural circuits". In: *Neuron* 44.1, pp. 23–30.

Daniel, R. and S. Pollmann (2014). "A universal role of the ventral striatum in reward-based learning: evidence from human studies". In: *Neurobiology of learning and memory* 114, pp. 90–100.

Daunizeau, J., V. Adam, and L. Rigoux (2014). "VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data". In: *PLoS computational biology* 10.1, e1003441.

Dave, A. S. and D. Margoliash (2000). "Song replay during sleep and computational rules for sensorimotor vocal learning". In: *Science* 290.5492, pp. 812–816.

Davison, A. J. and D. W. Murray (2002). "Simultaneous localization and map-building using active vision". In: *IEEE transactions on pattern analysis and machine intelligence* 24.7, pp. 865–880.

Daw, N. D. et al. (2011a). "Trial-by-trial data analysis using computational models". In: *Decision making, affect, and learning: Attention and performance XXIII* 23.1.

Daw, N. D., S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan (2011b). "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6, pp. 1204–1215.

Daw, N. D., Y. Niv, and P. Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12, pp. 1704–1711.

Daw, N. D., J. P. O'doherty, P. Dayan, B. Seymour, and R. J. Dolan (2006). "Cortical substrates for exploratory decisions in humans". In: *Nature* 441.7095, pp. 876–879.

De Lavilléon, G., M. M. Lacroix, L. Rondi-Reig, and K. Benchenane (2015). "Explicit memory creation during sleep demonstrates a causal role of place cells in navigation". In: *Nature neuroscience* 18.4, pp. 493–495.

Deb, K. (2011). "Multi-objective optimisation using evolutionary algorithms: an introduction". In: *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, pp. 3–34.

Deb, K., S. Agrawal, A. Pratap, and T. Meyarivan (2000). "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II". In: *International conference on parallel problem solving from nature*. Springer, pp. 849–858.

Deb, K. and H. Jain (2013). "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints". In: *IEEE transactions on evolutionary computation* 18.4, pp. 577–601.

Delgado, M. R. (2007). "Reward-related responses in the human striatum". In: *Annals of the New York Academy of Sciences* 1104.1, pp. 70–88.

Devan, B. D. and N. M. White (1999). "Parallel information processing in the dorsal striatum: relation to hippocampal function". In: *Journal of neuroscience* 19.7, pp. 2789–2798.

Diba, K. and G. Buzsáki (2007). "Forward and reverse hippocampal place-cell sequences during ripples". In: *Nature neuroscience* 10.10, pp. 1241–1242.

Dickinson, A. and B. Balleine (2002). *The role of learning in motivation In Gallistel CR (Ed.), Stevens' handbook of experimental psychology (Vol. 3, pp. 497–533)*.

Dickinson, A. (1985). "Actions and habits: the development of behavioural autonomy". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308.1135, pp. 67–78.

Dickinson, A. and B. Balleine (1994). "Motivational control of goal-directed action". In: *Animal Learning & Behavior* 22.1, pp. 1–18.

Diekmann, N. and S. Cheng (2022). "A Model of Hippocampal Replay Driven by Experience and Environmental Structure Facilitates Spatial Learning". In: *bioRxiv*.

Dolan, R. J. and P. Dayan (2013). "Goals and habits in the brain". In: *Neuron* 80.2, pp. 312–325.

Dollé, L., R. Chavarriaga, A. Guillot, and M. Khamassi (2018). "Interactions of spatial strategies producing generalization gradient and blocking: A computational approach". In: *PLoS computational biology* 14.4, e1006092.

Dollé, L., M. Khamassi, B. Girard, A. Guillot, and R. Chavarriaga (2008). "Analyzing interactions between navigation strategies using a computational model of action selection". In: *International Conference on Spatial Cognition*. Springer, pp. 71–86.

Dollé, L., D. Sheynikhovich, B. Girard, R. Chavarriaga, and A. Guillot (2010). "Path planning versus cue responding: a bio-inspired model of switching between navigation strategies". In: *Biological cybernetics* 103.4, pp. 299–317.

Doya, K., K. Samejima, K.-i. Katagiri, and M. Kawato (2002). "Multiple model-based reinforcement learning". In: *Neural computation* 14.6, pp. 1347–1369.

Drai, D., N. Kafkafi, Y. Benjamini, G. Elmer, and I. Golani (2001). "Rats and mice share common ethologically relevant parameters of exploratory behavior". In: *Behavioural brain research* 125.1-2, pp. 133–140.

Dromnelle, R., B. Girard, E. Renaudo, R. Chatila, and M. Khamassi (2020). "Coping with the variability in humans reward during simulated human-robot interactions through the coordination of multiple learning strategies". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 612–617.

Dromnelle, R., E. Renaudo, M. Chetouani, P. Maragos, R. Chatila, B. Girard, and M. Khamassi (2022). "Reducing Computational Cost During Robot Navigation and Human–Robot Interaction with a Human-Inspired Reinforcement Learning Architecture". In: *International Journal of Social Robotics*, pp. 1–27.

Dromnelle, R., E. Renaudo, G. Pourcel, R. Chatila, B. Girard, and M. Khamassi (2020). "How to reduce computation time while sparing performance during robot navigation? A neuro-inspired architecture for autonomous shifting between model-based and model-free learning". In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 68–79.

Ego-Stengel, V. and M. A. Wilson (2010). "Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat". In: *Hippocampus* 20.1, pp. 1–10.

Euston, D. R., M. Tatsuno, and B. L. McNaughton (2007). "Fast-forward playback of recent memory sequences in prefrontal cortex during sleep". In: *science* 318.5853, pp. 1147–1150.

Fedus, W., P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, and W. Dabney (2020). "Revisiting fundamentals of experience replay". In: *International Conference on Machine Learning*. PMLR, pp. 3061–3071.

Fleischer, J. G., J. A. Gally, G. M. Edelman, and J. L. Krichmar (2007). "Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device". In: *Proceedings of the National Academy of Sciences* 104.9, pp. 3556–3561.

Fogel, D. B. (1997). "The Advantages of Evolutionary Computation." In: *Bcec*, pp. 1–11.

Fonio, E., Y. Benjamini, and I. Golani (2009). "Freedom of movement and the stability of its unfolding in free exploration of mice". In: *Proceedings of the National Academy of Sciences* 106.50, pp. 21335–21340.

Foster, D. J. (2017). "Replay comes of age". In: *Annu. Rev. Neurosci* 40.581-602, p. 9.

Foster, D. J. and M. A. Wilson (2006). "Reverse replay of behavioural sequences in hippocampal place cells during the awake state". In: *Nature* 440.7084, pp. 680–683.

Fragaszy, D. and Q. Liu (2012). "Instrumental Behavior, Problem-Solving, and Tool Use in Nonhuman Animals". In: *Encyclopedia of the Sciences of Learning*. Ed. by N. M. Seel. Boston, MA: Springer US, pp. 1579–1582.

Gaussier, P., J. Banquet, F. Sargolini, C. Giovannangeli, E. Save, and B. Poucet (2007). "A model of grid cells involving extra hippocampal path integration, and the hippocampal loop". In: *Journal of integrative neuroscience* 6.03, pp. 447–476.

Gaussier, P., A. Revel, J.-P. Banquet, and V. Babeau (2002). "From view cells and place cells to cognitive map learning: processing stages of the hippocampal system". In: *Biological cybernetics* 86.1, pp. 15–28.

Girard, B. (2021). *Replays of a Sleeping Mouse*.

Girardeau, G., K. Benchenane, S. I. Wiener, G. Buzsáki, and M. B. Zugaro (2009). "Selective suppression of hippocampal ripples impairs spatial memory". In: *Nature neuroscience* 12.10, pp. 1222–1223.

Girardeau, G., I. Inema, and G. Buzsáki (2017). "Reactivations of emotional memory in the hippocampus–amygdala system during sleep". In: *Nature neuroscience* 20.11, pp. 1634–1642.

Gläscher, J., N. Daw, P. Dayan, and J. P. O'Doherty (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Glimcher, P. W. (2011). "Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis". In: *Proceedings of the National Academy of Sciences* 108.supplement_3, pp. 15647–15654.

Golani, I., Y. Benjamini, and D. Eilam (1993). "Stopping behavior: constraints on exploration in rats (Rattus norvegicus)". In: *Behavioural brain research* 53.1-2, pp. 21–33.

Goldberg, D. E. (1989). "Genetic algorithms in search, optimization, and machine learning. Addison". In: *Reading*.

Gordon, G. and E. Ahissar (2011). "Reinforcement active learning hierarchical loops". In: *The 2011 International Joint Conference on Neural Networks*. IEEE, pp. 3008–3015.

— (2012). "Hierarchical curiosity loops and active sensing". In: *Neural Networks* 32, pp. 119–129.

Gordon, G., E. Fonio, and E. Ahissar (2014a). "Emergent exploration via novelty management". In: *Journal of Neuroscience* 34.38, pp. 12646–12661.

— (2014b). "Learning and control of exploration primitives". In: *Journal of computational neuroscience* 37.2, pp. 259–280.

Grisetti, G., C. Stachniss, and W. Burgard (2007). "Improved techniques for grid mapping with rao-blackwellized particle filters". In: *IEEE transactions on Robotics* 23.1, pp. 34–46.

Guazzelli, A., M. Bota, F. J. Corbacho, and M. A. Arbib (1998). "Affordances. motivations, and the world graph theory". In: *Adaptive Behavior* 6.3-4, pp. 435–471.

Gupta, A. S., M. A. Van Der Meer, D. S. Touretzky, and A. D. Redish (2010). "Hippocampal replay is not a simple function of experience". In: *Neuron* 65.5, pp. 695–705.

Hafting, T., M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser (2005). "Microstructure of a spatial map in the entorhinal cortex". In: *Nature* 436.7052, pp. 801–806.

Hammond, L. J. (1980). "The effect of contingency upon the appetitive conditioning of free-operant behavior". In: *Journal of the experimental analysis of behavior* 34.3, pp. 297–304.

Hansen, N. (2006). "The CMA evolution strategy: a comparing review". In: *Towards a new evolutionary computation*, pp. 75–102.

— (2016). "The CMA evolution strategy: A tutorial". In: *arXiv preprint arXiv:1604.00772*.

Hao, G.-S., M.-H. Lim, Y.-S. Ong, H. Huang, and G.-G. Wang (2019). "Domination landscape in evolutionary algorithms and its applications". In: *Soft Computing* 23.11, pp. 3563–3570.

Haruno, M. and M. Kawato (2006). "Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning". In: *Journal of neurophysiology* 95.2, pp. 948–959.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hikosaka, O., H. Nakahara, M. K. Rand, K. Sakai, X. Lu, K. Nakamura, S. Miyachi, and K. Doya (1999). "Parallel neural networks for learning sequential procedures". In: *Trends in neurosciences* 22.10, pp. 464–471.

Hok, V., E. Save, P. Lenck-Santini, and B. Poucet (2005). "Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex". In: *Proceedings of the National Academy of Sciences* 102.12, pp. 4602–4607.

Houk, J. C., J. L. Davis, and D. G. Beiser (1995). *Models of information processing in the basal ganglia*. MIT press.

Insafutdinov, E., L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele (2016). "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: *European conference on computer vision*. Springer, pp. 34–50.

Jadhav, S. P., C. Kemere, P. W. German, and L. M. Frank (2012). "Awake hippocampal sharp-wave ripples support spatial memory". In: *Science* 336.6087, pp. 1454–1458.

Jarboui, F. and A. Akakzia (2022). "Delayed Geometric Discounts: An Alternative Criterion for Reinforcement Learning". In: *arXiv preprint arXiv:2209.12483*.

Jauffret, A., N. Cuperlier, and P. Gaussier (2015). "From grid cells and visual place cells to multimodal place cell: a new robotic architecture". In: *Frontiers in neurorobotics* 9, p. 1.

Ji, D. and M. A. Wilson (2007). "Coordinated memory replay in the visual cortex and hippocampus during sleep". In: *Nature neuroscience* 10.1, pp. 100–107.

Jiang, R., S. Ci, D. Liu, X. Cheng, and Z. Pan (2021). "A hybrid multi-objective optimization method based on NSGA-II algorithm and entropy weighted TOPSIS for lightweight design of dump truck carriage". In: *Machines* 9.8, p. 156.

Joel, D., Y. Niv, and E. Ruppin (2002). "Actor–critic models of the basal ganglia: New anatomical and computational perspectives". In: *Neural networks* 15.4-6, pp. 535–547.

Johnson, A. and A. D. Redish (2007). "Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point". In: *Journal of Neuroscience* 27.45, pp. 12176–12189.

Karlsson, M. P. and L. M. Frank (2009). "Awake replay of remote experiences in the hippocampus". In: *Nature neuroscience* 12.7, pp. 913–918.

Kawato, M. (1999). "Internal models for motor control and trajectory planning". In: *Current opinion in neurobiology* 9.6, pp. 718–727.

Keramati, M., A. Dezfouli, and P. Piray (2011). "Speed/accuracy trade-off between the habitual and the goal-directed processes". In: *PLoS computational biology* 7.5, e1002055.

Keramati, M., P. Smittenaar, R. J. Dolan, and P. Dayan (2016). "Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum". In: *Proceedings of the National Academy of Sciences* 113.45, pp. 12868–12873.

Khamassi, M. (2007). "Complementary roles of the rat prefrontal cortex and striatum in reward-based learning and shifting navigation strategies". PhD thesis. Université Pierre et Marie Curie-Paris VI.

Khamassi, M. and B. Girard (2020). "Modeling awake hippocampal reactivations with model-based bidirectional search". In: *Biological Cybernetics* 114.2, pp. 231–248.

Khamassi, M. and M. D. Humphries (2012). "Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies". In: *Frontiers in behavioral neuroscience* 6, p. 79.

Khamassi, M., L. Lachèze, B. Girard, A. Berthoz, and A. Guillot (2005). "Actor–Critic models of reinforcement learning in the basal ganglia: from natural to artificial rats". In: *Adaptive Behavior* 13.2, pp. 131–148.

Khamassi, M., S. Lallée, P. Enel, E. Procyk, and P. F. Dominey (2011). "Robot cognitive control with a neurophysiologically inspired reinforcement learning model". In: *Frontiers in neurorobotics* 5, p. 1.

Klatzky, R. L. (1998). "Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections". In: *Spatial cognition*. Springer, pp. 1–17.

Kober, J., J. A. Bagnell, and J. Peters (2013). "Reinforcement learning in robotics: A survey". In: *The International Journal of Robotics Research* 32.11, pp. 1238–1274.

Koechlin, E. and C. Summerfield (2007). "An information theoretical approach to prefrontal executive function". In: *Trends in cognitive sciences* 11.6, pp. 229–235.

Konda, V. and J. Tsitsiklis (1999). "Actor-critic algorithms". In: *Advances in neural information processing systems* 12.

Konorski, J. (1967). "Integrative activity of the brain; an interdisciplinary approach". In.

Kruskal, W. H. and W. A. Wallis (1952). "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260, pp. 583–621.

Kudrimoti, H. S., C. A. Barnes, and B. L. McNaughton (1999). "Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics". In: *Journal of Neuroscience* 19.10, pp. 4090–4101.

Lansink, C. S., P. M. Goltstein, J. V. Lankelma, B. L. McNaughton, and C. M. Pennartz (2009). "Hippocampus leads ventral striatum in replay of place-reward information". In: *PLoS biology* 7.8, e1000173.

Laventure, S. and K. Benchenane (2020). "Validating the theoretical bases of sleep reactivation during sharp-wave ripples and their association with emotional valence". In: *Hippocampus* 30.1, pp. 19–27.

Le Merre, P., V. Esmaeili, E. Charrière, K. Galan, P.-A. Salin, C. C. Petersen, and S. Crochet (2018). "Reward-based learning drives rapid sensory signals in medial prefrontal cortex and dorsal hippocampus necessary for goal-directed behavior". In: *Neuron* 97.1, pp. 83–91.

Lee, A. K. and M. A. Wilson (2002). "Memory of sequential experience in the hippocampus during slow wave sleep". In: *Neuron* 36.6, pp. 1183–1194.

Lee, S. W., S. Shimojo, and J. P. O'Doherty (2014). "Neural computations underlying arbitration between model-based and model-free learning". In: *Neuron* 81.3, pp. 687–699.

Lesaint, F., O. Sigaud, S. B. Flagel, T. E. Robinson, and M. Khamassi (2014). "Modelling individual differences in the form of Pavlovian conditioned approach responses: a dual learning systems approach with factored representations". In: *PLoS computational biology* 10.2, e1003466.

Lieberman, M. D., R. Gaunt, D. T. Gilbert, and Y. Trope (2002). "Reflexion and reflection: a social cognitive neuroscience approach to attributional inference." In.

Liénard, J. and B. Girard (2014). "A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection". In: *Journal of computational neuroscience* 36.3, pp. 445–468.

Liénard, J., A. Guillot, and B. Girard (2010). "Multi-objective evolutionary algorithms to investigate neurocomputational issues: the case study of basal ganglia models". In: *International Conference on Simulation of Adaptive Behavior*. Springer, pp. 597–606.

Lin, L.-J. (1992). "Self-improving reactive agents based on reinforcement learning, planning and teaching". In: *Machine learning* 8.3, pp. 293–321.

Little, D. Y. and F. T. Sommer (2013). "Learning and exploration in action-perception loops". In: *Frontiers in neural circuits* 7, p. 37.

Llofriu, M., G. Tejera, M. Contreras, T. Pelc, J.-M. Fellous, and A. Weitzenfeld (2015). "Goal-oriented robot navigation learning using a multi-scale space representation". In: *Neural Networks* 72, pp. 62–74.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2, pp. 129–137.

Maffei, G., D. Santos-Pata, E. Marcos, M. Sánchez-Fibla, and P. F. Verschure (2015). "An embodied biologically constrained model of foraging: from classical and operant conditioning to adaptive real-world behavior in DAC-X". In: *Neural Networks* 72, pp. 88–108.

Maingret, N., G. Girardeau, R. Todorova, M. Goutierre, and M. Zugaro (2016). "Hippocampo–cortical coupling mediates memory consolidation during sleep". In: *Nature neuroscience* 19.7, pp. 959–964.

Markowitz, J. E., W. F. Gillis, M. Jay, J. Wood, R. W. Harris, R. Cieszkowski, R. Scott, D. Brann, D. Koveal, T. Kula, C. Weinreb, M. A. M. Osman, S. R. Pinto, N. Uchida, S. W. Linderman, B. L. Sabatini, and S. R. Datta (2023). "Spontaneous behaviour is structured by reinforcement without explicit reward". In: *Nature*.

Massi, E., J. Barthélemy, J. Mailly, R. Dromnelle, J. Canitrot, E. Poniatowski, B. Girard, and M. Khamassi (2022). "Model-Based and Model-Free Replay Mechanisms for Reinforcement Learning in Neurorobotics". In: *Frontiers in Neurorobotics* 16.

Massi, E., L. Vannucci, U. Albanese, M. C. Capolei, A. Vandesompele, G. Urbain, A. M. Sabatini, J. Dambre, C. Laschi, S. Tolu, et al. (2019). "Combining evolutionary and adaptive control strategies for quadruped robotic locomotion". In: *Frontiers in Neurorobotics* 13, p. 71.

Mathis, A., P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge (2018). "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature neuroscience* 21.9, pp. 1281–1289.

Mattar, M. G. and N. D. Daw (2018). "Prioritized memory access explains planning and hippocampal replay". In: *Nature neuroscience* 21.11, pp. 1609–1617.

McNaughton, B. L., J. J. Knierim, and M. A. Wilson (1995). "Vector encoding and the vestibular foundations of spatial cognition: neurophysiological and computational mechanisms." In.

Meyer, J.-A., A. Guillot, B. Girard, M. Khamassi, P. Pirim, and A. Berthoz (2005). "The Psikharpax project: towards building an artificial rat". In: *Robotics and autonomous systems* 50.4, pp. 211–223.

Michon, F., J.-J. Sun, C. Y. Kim, D. Ciliberti, and F. Kloosterman (2019). "Post-learning hippocampal replay selectively reinforces spatial memory for highly rewarded locations". In: *Current Biology* 29.9, pp. 1436–1444.

Milford, M. and G. Wyeth (2010). "Persistent navigation and mapping using a biologically inspired SLAM system". In: *The International Journal of Robotics Research* 29.9, pp. 1131–1153.

Milford, M. J. and G. F. Wyeth (2008). "Mapping a suburb with a single camera using a biologically inspired SLAM system". In: *IEEE Transactions on Robotics* 24.5, pp. 1038–1053.

Miller, E. K., J. D. Cohen, et al. (2001). "An integrative theory of prefrontal cortex function". In: *Annual review of neuroscience* 24.1, pp. 167–202.

Miyamoto, H., M. Kawato, T. Setoyama, and R. Suzuki (1988). "Feedback-error-learning neural network for trajectory control of a robotic manipulator". In: *Neural networks* 1.3, pp. 251–265.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.

Montemerlo, M., S. Thrun, D. Koller, B. Wegbreit, et al. (2002). "FastSLAM: A factored solution to the simultaneous localization and mapping problem". In: *Aaai/iaai* 593598.

Moore, A. W. and C. G. Atkeson (1993). "Prioritized sweeping: Reinforcement learning with less data and less time". In: *Machine learning* 13.1, pp. 103–130.

Morris, R. G. (1981). "Spatial localization does not require the presence of local cues". In: *Learning and motivation* 12.2, pp. 239–260.

Morris, R. G., P. Garrud, J. a. Rawlins, and J. O'Keefe (1982). "Place navigation impaired in rats with hippocampal lesions". In: *Nature* 297.5868, pp. 681–683.

Moser, E. I., E. Kropff, M.-B. Moser, et al. (2008). "Place cells, grid cells, and the brain's spatial representation system". In: *Annual review of neuroscience* 31.1, pp. 69–89.

Moser, M.-B., D. C. Rowland, and E. I. Moser (2015). "Place cells, grid cells, and memory". In: *Cold Spring Harbor perspectives in biology* 7.2, a021808.

Moxon, K. A. (2005). "Neurorobotics". In: *Neural engineering*. Springer, pp. 123–155.

Murata, T., H. Ishibuchi, et al. (1995). "MOGA: multi-objective genetic algorithms". In: *IEEE international conference on evolutionary computation*. Vol. 1. IEEE Piscataway, NJ, USA, pp. 289–294.

O'Doherty, J. P., P. Dayan, K. Friston, H. Critchley, and R. J. Dolan (2003). "Temporal difference models and reward-related learning in the human brain". In: *Neuron* 38.2, pp. 329–337.

O'Keefe, J. (1976). "Place units in the hippocampus of the freely moving rat". In: *Experimental neurology* 51.1, pp. 78–109.

O'Keefe, J., N. Burgess, J. G. Donnett, K. J. Jeffery, and E. A. Maguire (1998). "Place cells, navigational accuracy, and the human hippocampus". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353.1373, pp. 1333–1340.

O'Keefe, J. and D. H. Conway (1978). "Hippocampal place units in the freely moving rat: why they fire where they fire". In: *Experimental brain research* 31.4, pp. 573–590.

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

O'Doherty, J. P., S. W. Lee, R. Tadayonnejad, J. Cockburn, K. Iigaya, and C. J. Charpentier (2021). "Why and how the brain weights contributions from a mixture of experts". In: *Neuroscience & Biobehavioral Reviews* 123, pp. 14–23.

Ólafsdóttir, H. F., C. Barry, A. B. Saleem, D. Hassabis, and H. J. Spiers (2015). "Hippocampal place cells construct reward related sequences through unexplored space". In: *Elife* 4, e06063.

Ólafsdóttir, H. F., D. Bush, and C. Barry (2018). "The role of hippocampal replay in memory and planning". In: *Current Biology* 28.1, R37–R50.

Ólafsdóttir, H. F., F. Carpenter, and C. Barry (2017). "Task demands predict a dynamic switch in the content of awake hippocampal replay". In: *Neuron* 96.4, pp. 925–935.

Otto, A. R., S. J. Gershman, A. B. Markman, and N. D. Daw (2013). "The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive". In: *Psychological science* 24.5, pp. 751–761.

Oudeyer, P.-Y., F. Kaplan, and V. V. Hafner (2007). "Intrinsic Motivation Systems for Autonomous Mental Development". In: *IEEE Transactions on Evolutionary Computation* 11.2, pp. 265–286.

Palminteri, S., M. Khamassi, M. Joffily, and G. Coricelli (2015). "Contextual modulation of value signals in reward and punishment learning". In: *Nature communications* 6.1, pp. 1–14.

Palminteri, S., V. Wyart, and E. Koechlin (2017). "The importance of falsification in computational cognitive modeling". In: *Trends in cognitive sciences* 21.6, pp. 425–433.

Pang, R. and A. L. Fairhall (2019). "Fast and flexible sequence induction in spiking neural networks via rapid excitability changes". In: *Elife* 8.

Pape, L., C. M. Oddo, M. Controzzi, C. Cipriani, A. Förster, M. C. Carrozza, and J. Schmidhuber (2012). "Learning tactile skills through curious exploration". In: *Frontiers in neurorobotics* 6, p. 6.

Park, S. W., H. J. Jang, M. Kim, and J. Kwag (2019). "Spatiotemporally random and diverse grid cell spike patterns contribute to the transformation of grid cell to place cell in a neural network model". In: *Plos one* 14.11, e0225100.

Partridge, J. G., K.-C. Tang, and D. M. Lovinger (2000). "Regional and postnatal heterogeneity of activity-dependent long-term changes in synaptic efficacy in the dorsal striatum". In: *Journal of Neurophysiology* 84.3, pp. 1422–1429.

Pavlides, C. and J. Winson (1989). "Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes". In: *Journal of neuroscience* 9.8, pp. 2907–2918.

Peng, J. and R. J. Williams (1993). "Efficient learning and planning within the Dyna framework". In: *Adaptive behavior* 1.4, pp. 437–454.

Peyrache, A., M. Khamassi, K. Benchenane, S. I. Wiener, and F. P. Battaglia (2009). "Replay of rule-learning related neural patterns in the prefrontal cortex during sleep". In: *Nature neuroscience* 12.7, pp. 919–926.

Pezzulo, G., C. Kemere, and M. A. Van Der Meer (2017). "Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition". In: *Annals of the New York Academy of Sciences* 1396.1, pp. 144–165.

Pezzulo, G., F. Rigoli, and F. Chersi (2013). "The mixed instrumental controller: using value of information to combine habitual choice and mental simulation". In: *Frontiers in psychology* 4, p. 92.

Pfeiffer, B. E. (2020). "The content of hippocampal "replay"". In: *Hippocampus* 30.1, pp. 6–18.

Pfeiffer, B. E. and D. J. Foster (2013). "Hippocampal place-cell sequences depict future paths to remembered goals". In: *Nature* 497.7447, pp. 74–79.

Pickens, C. L., M. P. Saddoris, B. Setlow, M. Gallagher, P. C. Holland, and G. Schoenbaum (2003). "Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task". In: *Journal of Neuroscience* 23.35, pp. 11078–11084.

Quigley, M., K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, et al. (2009). "ROS: an open-source Robot Operating System". In: *ICRA workshop on open source software*. Vol. 3. 3.2. Kobe, Japan, p. 5.

Ranck Jr, J. (1984). "Head direction cells in the deep layer of dorsal presubiculum in freely moving rats". In: *Society of neuroscience abstract*. Vol. 10, p. 599.

Redish, A. D. (2016). "Vicarious trial and error". In: *Nature Reviews Neuroscience* 17.3, pp. 147–159.

Renaudo, E., B. Girard, R. Chatila, and M. Khamassi (2014). "Design of a control architecture for habit learning in robots". In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 249–260.

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

Samsonovich, A. and B. L. McNaughton (1997). "Path integration and cognitive mapping in a continuous attractor neural network model". In: *Journal of Neuroscience* 17.15, pp. 5900–5920.

Sargolini, F., M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser (2006). "Conjunctive representation of position, direction, and velocity in entorhinal cortex". In: *Science* 312.5774, pp. 758–762.

Schaffer, J. D. (1985). "Multiple objective optimization with vector evaluated genetic algorithms". In: *Proceedings of the First International Conference of Genetic Algorithms and Their Application*, pp. 93–100.

Schaul, T., J. Quan, I. Antonoglou, and D. Silver (2015). "Prioritized experience replay". In: *arXiv preprint arXiv:1511.05952*.

Schultz, W., P. Dayan, and P. R. Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599.

Scoville, W. B. and B. Milner (1957). "Loss of recent memory after bilateral hippocampal lesions". In: *Journal of neurology, neurosurgery, and psychiatry* 20.1, p. 11.

Segal, M., J. F. Disterhoft, and J. Olds (1972). "Hippocampal unit activity during classical aversive and appetitive conditioning". In: *Science* 175.4023, pp. 792–794.

Seijen, H. van and R. Sutton (2015). "A deeper look at planning as learning from replay". In: *International conference on machine learning*. PMLR, pp. 2314–2322.

Singer, A. C. and L. M. Frank (2009). "Rewarded outcomes enhance reactivation of experience in the hippocampus". In: *Neuron* 64.6, pp. 910–921.

Solstad, T., C. N. Boccara, E. Kropff, M.-B. Moser, and E. I. Moser (2008). "Representation of geometric borders in the entorhinal cortex". In: *Science* 322.5909, pp. 1865–1868.

Souza Muñoz, M. E. de, M. Chaves Menezes, E. Pignaton de Freitas, S. Cheng, P. R. de Almeida Ribeiro, A. de Almeida Neto, and A. C. Muniz de Oliveira (2022). "xRatSLAM: An Extensible RatSLAM Computational Framework". In: *Sensors* 22.21, p. 8305.

Srinivas, N. and K. Deb (1994). "Muiltiobjective optimization using nondominated sorting in genetic algorithms". In: *Evolutionary computation* 2.3, pp. 221–248.

Stackman, R. W. and J. S. Taube (1998). "Firing properties of rat lateral mammillary single units: head direction, head pitch, and angular head velocity". In: *Journal of Neuroscience* 18.21, pp. 9020–9037.

Stella, F., P. Baracskay, J. O'Neill, and J. Csicsvari (2019). "Hippocampal reactivation of random trajectories resembling Brownian diffusion". In: *Neuron* 102.2, pp. 450–461.

Stensola, T. and E. I. Moser (2016). "Grid cells and spatial maps in entorhinal cortex and hippocampus". In: *Micro-, meso-and macro-dynamics of the brain*, pp. 59–80.

Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming". In: *Machine learning proceedings 1990*. Elsevier, pp. 216–224.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. and A. Barto (1998). *Introduction to Reinforcement Learning. Cambridge, MA*. Cambridge, MA: MIT Press.

Taube, J. S. (1995). "Head direction cells recorded in the anterior thalamic nuclei of freely moving rats". In: *Journal of Neuroscience* 15.1, pp. 70–86.

— (2007). "The head direction signal: origins and sensory-motor integration". In: *Annual review of neuroscience* 30.1, pp. 181–207.

Taube, J. S., R. U. Muller, and J. B. Ranck (1990a). "Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis". In: *Journal of Neuroscience* 10.2, pp. 420–435.

— (1990b). "Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations". In: *Journal of Neuroscience* 10.2, pp. 436–447.

Tchernichovski, O., Y. Benjamini, and I. Golani (1996). "Constraints and the emergence of'free'exploratory behavior in rat ontogeny". In: *Behaviour*, pp. 519–539.

Tchernichovski, O. and I. Golani (1995). "A phase plane representation of rat exploratory behavior". In: *Journal of neuroscience methods* 62.1-2, pp. 21–27.

Thrun, S., W. Burgard, and D. Fox (2006). "Probalistic robotics". In: *Kybernetes*.

Tolman, E. C. (1948). "Cognitive maps in rats and men." In: *Psychological review* 55.4, p. 189.

Tolman, E. C. (1939). "Prediction of vicarious trial and error by means of the schematic sowbug." In: *Psychological Review* 46.4, p. 318.

Touretzky, D. S. and A. D. Redish (1996). "Theory of rodent navigation based on interacting representations of space". In: *Hippocampus* 6.3, pp. 247–270.

Touretzky, D. S., H. S. Wan, and A. D. Redish (1994). "Neural representation of space in rats and robots". In: *Computational Intelligence: Imitating Life*, pp. 57–68.

Treit, D. and M. Fundytus (1988). "Thigmotaxis as a test for anxiolytic activity in rats". In: *Pharmacology Biochemistry and Behavior* 31.4, pp. 959–962.

Tricomi, E., B. W. Balleine, and J. P. O'Doherty (2009). "A specific role for posterior dorsolateral striatum in human habit learning". In: *European Journal of Neuroscience* 29.11, pp. 2225–2232.

Trullier, O., S. I. Wiener, A. Berthoz, and J.-A. Meyer (1997). "Biologically based artificial navigation systems: Review and prospects". In: *Progress in neurobiology* 51.5, pp. 483–544.

Ujfalussy, B., P. Erős, Z. Somogyvári, and T. Kiss (2008). "Episodes in space: A modeling study of hippocampal place representation". In: *International Conference on Simulation of Adaptive Behavior*. Springer, pp. 123–136.

Valenti, O., N. Mikus, and T. Klausberger (2018). "The cognitive nuances of surprising events: exposure to unexpected stimuli elicits firing variations in neurons of the dorsal CA1 hippocampus". In: *Brain Structure and Function* 223.7, pp. 3183–3211.

Valentin, V. V., A. Dickinson, and J. P. O'Doherty (2007). "Determining the neural substrates of goal-directed learning in the human brain". In: *Journal of Neuroscience* 27.15, pp. 4019–4026.

Verschure, P. F., C. M. Pennartz, and G. Pezzulo (2014). "The why, what, where, when and how of goal-directed choice: neuronal and computational principles". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130483.

Viejo, G., M. Khamassi, A. Brovelli, and B. Girard (2015). "Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning". In: *Frontiers in behavioral neuroscience* 9, p. 225.

Walsh, R. N. and R. A. Cummins (1976). "The open-field test: a critical review." In: *Psychological bulletin* 83.3, p. 482.

Wang, Y., S. Li, Q. Chen, and W. Hu (2007). "Biology inspired robot behavior selection mechanism: Using genetic algorithm". In: *International Conference on Life System Modeling and Simulation*. Springer, pp. 777–786.

Watkins, C. J. C. H. (1989). "Learning from delayed rewards". In.

Watson, T. C., N. L. Cerminara, B. M. Lumb, and R. Apps (2016). "Neural correlates of fear in the periaqueductal gray". In: *Journal of Neuroscience* 36.50, pp. 12707–12719.

Welker, W. (1957). ""Free" versus "forced" exploration of a novel situation by rats". In: *Psychological Reports* 3.1, pp. 95–108.

Whelan, M. T., A. Jimenez-Rodriguez, T. J. Prescott, and E. Vasilaki (2022). "A robotic model of hippocampal reverse replay for reinforcement learning". In: *Bioinspiration & Biomimetics* 18.1, p. 015007.

Whelan, M. T., T. J. Prescott, and E. Vasilaki (2020). "Fast Reverse Replays of Recent Spatiotemporal Trajectories in a Robotic Hippocampal Model". In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 390–401.

Whelan, M. T., E. Vasilaki, and T. J. Prescott (2019). "Robots that imagine–can hippocampal replay be utilized for robotic mnemonics?" In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 277–286.

Whyte, H. D. (2006). "Simultaneous localisation and mapping (SLAM): Part I the essential algorithms". In: *Robotics and Automation Magazine*.

Wilson, M. A. and B. L. McNaughton (1994). "Reactivation of hippocampal ensemble memories during sleep". In: *Science* 265.5172, pp. 676–679.

Wilson, R. C. and A. G. Collins (2019). "Ten simple rules for the computational modeling of behavioral data". In: *Elife* 8, e49547.

Wittmann, T. and H. Schwegler (1995). "Path integration—a network model". In: *Biological Cybernetics* 73.6, pp. 569–575.

Wrase, J., T. Kahnt, F. Schlagenhauf, A. Beck, M. X. Cohen, B. Knutson, and A. Heinz (2007). "Different neural systems adjust motor behavior in response to reward and punishment". In: *Neuroimage* 36.4, pp. 1253–1262.

Wu, C.-T., D. Haggerty, C. Kemere, and D. Ji (2017). "Hippocampal awake replay in fear memory retrieval". In: *Nature neuroscience* 20.4, pp. 571–580.

Wunderlich, K., P. Dayan, and R. J. Dolan (2012). "Mapping value based planning and extensively trained choice in the human brain". In: *Nature neuroscience* 15.5, pp. 786–791.

Yacubian, J., J. Gläscher, K. Schroeder, T. Sommer, D. F. Braus, and C. Büchel (2006). "Dissociable systems for gain-and loss-related value predictions and errors of prediction in the human brain". In: *Journal of Neuroscience* 26.37, pp. 9530–9537.

Yin, H. H. and B. J. Knowlton (2006). "The role of the basal ganglia in habit formation". In: *Nature Reviews Neuroscience* 7.6, pp. 464–476.

Yin, H. H., B. J. Knowlton, and B. W. Balleine (2004). "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning". In: *European journal of neuroscience* 19.1, pp. 181–189.

— (2005). "Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning". In: *European Journal of Neuroscience* 22.2, pp. 505–512.

— (2006). "Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning". In: *Behavioural brain research* 166.2, pp. 189–196.

Yin, H. H., S. P. Mulcare, M. R. Hilário, E. Clouse, T. Holloway, M. I. Davis, A. C. Hansson, D. M. Lovinger, and R. M. Costa (2009). "Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill". In: *Nature neuroscience* 12.3, pp. 333–341.

Yin, H. H., S. B. Ostlund, and B. W. Balleine (2008). "Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks". In: *European Journal of Neuroscience* 28.8, pp. 1437–1448.

Yu, X. and M. Gen (2010). *Introduction to evolutionary algorithms*. Springer Science & Business Media.

Zhang, S. and R. S. Sutton (2017). "A deeper look at experience replay". In: *arXiv preprint arXiv:1712.01275*.