Embodying a Computational Model of Hippocampal Replay for Robotic Reinforcement Learning

Matthew T. Whelan

Supervisors: Prof. E. Vasilaki and Prof. T. J. Prescott

Department of Computer Science
The University of Sheffield

This dissertation is submitted for the degree of
Master of Philosophy

August 2020

Abstract

Hippocampal reverse replay has been speculated to play an important role in biological reinforcement learning since its discovery over a decade ago. Whilst a number of computational models have recently emerged in an attempt to understand the dynamics of hippocampal replay, there has been little progress in testing and implementing these models in real-world robotics settings. Presented first in this body of work then is a bio-inspired hippocampal CA3 network model. It runs in real-time to produce reverse replays of recent spatio-temporal sequences, represented as place cell activities, in a robotic spatial navigation task. The model is based on two very recent computational models of hippocampal reverse replay. An analysis of these models show that, in their original forms, they are each insufficient for effective performance when applied to a robot. As such, choosing particular elements from each allows for a computational model that is sufficient for application in a robotic task.

Having a model of reverse replay applied successfully in a robot provides the groundwork necessary for testing the ways in which reverse replay contributes to reinforcement learning. The second portion of the work presented here builds on a previous reinforcement learning neural network model of a basic hippocampal-striatal circuit using a three-factor learning rule. By integrating reverse replays into this reinforcement learning model, results show that reverse replay, with its ability to replay the recent trajectory both in the hippocampal circuit and the striatal circuit, can speed up the learning process. In addition, for situations where the original reinforcement learning model performs poorly, such as when its time dynamics do not sufficiently store enough of the robot's behavioural history for effective learning, the reverse replay model can compensate for this by replaying the recent history. These results are inline with experimental findings showing that disruption of awake hippocampal replay events severely diminishes, but does not entirely eliminate, reinforcement learning.

This work provides possible insights into the important role that reverse replays could contribute to mnemonic function, and reinforcement learning in particular; insights that could benefit the robotic, AI, and neuroscience communities. However, there is still much to be done. How reverse replays are initiated is still an ongoing research problem, for instance. Furthermore, the model presented here generates place cells heuristically, but there are computational models tackling the problem of how hippocampal cells such as place cells, but also grid cells and head direction cells, emerge. This leads to the pertinent question of asking how these models, which make assumptions about their network architectures and dynamics, could integrate with the computational models of hippocampal replay which make their own assumptions on network architectures and dynamics.

List of figures

List of tables

Nomenclature

Roman Symbols

c_1, c_2	Constants in the action cell Sigmoidal activation function
d	Constant determining the place field width
D_k	Short-term depression for neuron k
e_{ij}	Synaptic eligibility trace at the synapse connecting place cell j to action cell i
F_k	Short-term facilitation for neuron k
I	Current
M_{target}	Magnitude of the population vector of action cells
R	Reward value
r_i	Rate for place cell i (model of Haga and Fukai)
s_j	Neuron spikes (= 1 or 0; model of Pang and Fairhall)
U	Constant in the short-term facilitation dynamics
v_i	Cell voltage (model of Pang and Fairhall)
w_{ij}	Weight projecting from neuron j onto neuron i
X_j	Place field x-coordinate centre location for place cell j
x_j	Rate for place cell j
x_ψ	Constant that determines the Sigmoidal midpoint in the calculation for intrinsic plasticity
X_{MiRo}	MiRo's x-coordinate position

y_i	Rate for action cell i
Y_j	Place field y-coordinate centre location for place cell j
Y_{MiRo}	MiRo's y-coordinate position

Greek Symbols

α	Constant in place cell linear rectifier
β	Determines the Sigmoidal slope in the intrinsic plasticity computation
ε	Constant in place cell linear rectifier
η	Learning rate
ψ_j	Intrinsic plasticity for place cell j
σ	Standard deviation
σ_i	Intrinsic plasticity for place cell i (model of Pang and Fairhall)
θ	MiRo's heading in the environment
θ_i	Heading that action cell i codes for
τ	Time constant

Table of contents

Acknowledgements

It is no understatement to say that, over the period in which this work has been conducted, things have been somewhat abnormal. Yet despite both the personal and the societal difficulties I and others have experienced over the past two years, it is only through the support of others that this work has been manageable.

Both my supervisors, Eleni and Tony, have been immensely supportive. From the first months, they have provided me with opportunities to travel to workshops and conferences, meeting along the way new colleagues and friends. It has been an incredibly stimulating experience, learning from them in the art of research and science. But Eleni and Tony have often gone beyond being mere supervisors, and have provided valuable support during rather difficult personal circumstances. I am in a much better place now than I was at the start of this MPhil, and it is only through their support that this has been possible. I can only hope to recompense through the effort I have put into this MPhil work, but will likely always remain indebted to them.

Individual research can often be an isolating experience, but my friends and colleagues from the Machine Learning group, in particular, Luca, August, Chao and Nada, have provided the necessary social support I needed at times. They are all intelligent, charming, and fun, and have made the last two years all the more enjoyable for it.

Finally, to my family. The work presented here was intended to be towards a PhD, and it was not easy to make the decision to leave the PhD. But despite some initial push back, I am incredibly grateful towards my family for supporting this decision. You know it's for the best. ☺

I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

– Isaac Newton

Chapter 1

Introduction

Hippocampal replay has received a great deal of interest since its initial discovery over two decades ago, and there has been much speculation over its role in a number of functions. The hippocampus, and its surrounding areas in the medial temporal lobe, has long been implicated to play an important role in mnemonic functions. As such, a great deal of the speculations pertaining to the functions of hippocampal replay have been associated to memory and learning. Reverse hippocampal replay in particular has been implicated in reinforcement learning, prompted by an original hypothesis that reverse replays might couple with phasic dopamine release in the striatum (Foster and Wilson, 2006). Later findings have provided further evidence to support this hypothesis, that reverse replay plays an important role in reinforcement learning, showing for instance that reverse replays are often modulated by reward, and that there are co-occuring replays in the ventral striatum alongside hippocampal reverse replays.

Reinforcement learning (RL) has been studied in both the contexts of psychology and artificial intelligence. Reinforcement learning in the context of artificial intelligence is defined as the ability of an agent to learn action sequences that maximise cumulative rewards (Sutton and Barto, 2018). Sometimes known as "conditioned reinforcement" in psychology, it takes on a similar definition here, in that animal behaviours can be reinforced through the delivery of rewards or punishment (Shahan, 2010). A classic example of biological reinforcement learning is the Morris Water Maze task in rodents (Morris, Anderson, et al., 1986). Figure 1.1 gives an example of this task, and it is this RL paradigm that is used for testing in the robot as part of this thesis (see Chapters 4 and 5). Many of the challenges in the development of effective and adaptable robotics can be posed as RL problems, and as such there has been no shortage of attempts to apply RL methods to robotics (Kober, Bagnell, and Peters, 2013; Sutton and Barto, 2018). However, robotics often expresses the need to solve some of the most difficult tasks in RL; problems related to factors such as continuous states and actions, end-to-end learning, reward signalling, computational efficiency, limited training examples, non-episodic resetting, and non-convergences due to continuously changing environments (Kober, Bagnell, and Peters, 2013; Kuutti et al., 2020; Zhu et al., 2020). But much of RL theory has been inspired by early behavioural studies in animals (Sutton and Barto, 2018), and for good reason, since biology has found many of the solutions to the control problems we are searching for in our engineered systems. As such, with continued developments in biology, and particularly in neuroscience, we would be wise to continue transferring insights from biology into robotics. Yet equally important is its inverse, by using our computational and robotic models to inform our understanding of the biology (Webb, 2001; Mitchinson, Pearson, et al., 2011). Of the neuroscience modelling approaches, computational models are perhaps the most easily transferable for robotic integration, whilst the robots themselves offer a valuable real-world testing opportunity to validate these computational models (Sheynikhovich et al., 2009; Jauffret, Cuperlier, and Gaussier, 2015; Prescott, Camilleri, et al., 2019). Computational neuroscience models of RL are therefore a potential opportunity to help solve the RL problems faced in robotics, whilst also helping us to deepen our understanding of the neurobiology of RL by testing the models in robots.

1.1 A *Brief* Introduction to the Hippocampus

Investigations of the hippocampus date back centuries, with the term *hippocampus* first being used by Giulio Cesare Aranzi (circa 1564). *Hippocampus* is a term meaning “sea horse” in Greek, and one can see the resemblance of this seafaring creature when comparing it to the shape and structure of the human hippocampus (see Figure 2.1, Chapter 2).

Interest in the functional properties of the hippocampus have received most attention over the past few decades for its role in episodic memory (Hasselmo, 2011). This was perhaps sparked by the well known 1957 study of Scoville and Milner (1957), who showed that patients with lesions of the hippocampal region had significant deficiencies in their ability to form and/or retrieve new episodic memories.[1] Interestingly, similar memory deficiencies can be found without hippocampal lesioning but by disrupting the normal activity of the hippocampal region (Girardeau et al., 2009). (For a fuller review of hippocampal function, see Deshmukh and Knierim, 2012.) But perhaps some of the most exciting experimental work conducted on the hippocampus has been performed over the most recent decades; one

[1] The definition of an episodic memory is best described with the pithy statement “What did you do at time T in place P?” (Hasselmo, 2011). This differs from semantic memory, say, which is defined as the memory we have for facts and general knowledge, or procedural memory, such as knowing how to tie a shoelace.

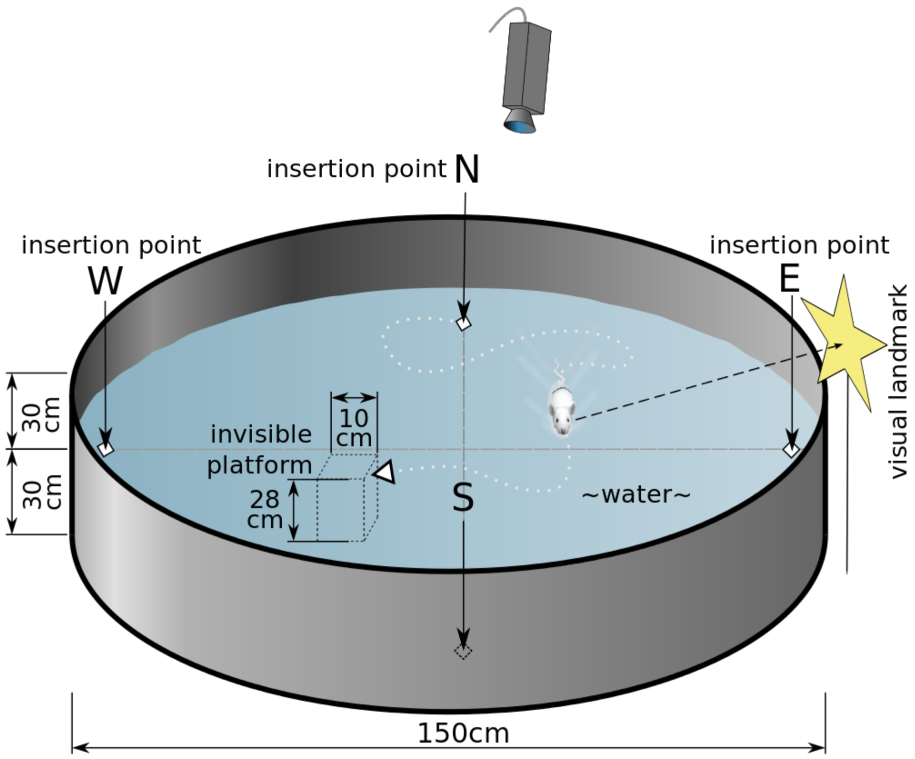


Fig. 1.1 An example of the Morris Water Maze task. This task tests the ability of a rodent to find a hidden goal location, in this instance a hidden platform. The rodent begins each trial in a random location in the water maze, and provided the rodent learns effectively, its time to reach the goal location reduces over a given number of trials. Image freely available under the Creative Commons Licence. Image source: https://commons.wikimedia.org/wiki/File:MorrisWaterMaze.svg

of these being the Nobel prize winning discovery of *place cells*, and another in a phenomenon termed *hippocampal replay*.

1.1.1 Place Cells

Hippocampal place cells, originally discovered in the CA1 region of rodents' hippocampi (O'Keefe and Dostrovsky, 1971a) (see Chapter 2 for a review of hippocampal neuroanatomy), are spatially tuned cells that respond when a rodent is positioned in a specific place in an environment. It was O'Keefe and Dostrovksy who first showed the existence of place cells (O'Keefe and Dostrovsky, 1971a). In their original experiment, they recorded the activity of units in the hippocampal CA1 region, and showed that one of these units responded most strongly when a rat was constrained at a specific orientation in its environment. If the rat was rotated away from this orientation, the activity of that unit reduced, until eventually the activity ceased altogether (relative to a baseline rate of activity) when the rat was rotated beyond approximately 90^{o} of the orientation for maximal activity. They found in total 8 of these orientation-selective hippocampal units. Later work by O'Keefe (1976) extended this result using rats that were no longer constrained, but were free to roam a maze. He showed the existence of 26 *place units*, or place cells, which were CA1 cells with activities that responded preferentially when a rat was in a particular position in the maze. The position in the maze, or any environment, for which a particular place cell preferentially fires is known as the place cell's *place field*.

As such, a collection of place cells, each of which responding to a different location in an environment, could provide a neural mechanism for the encoding of a cognitive map. At least, this was the hypothesis that O'Keefe and Nadel later argued in their book *The hippocampus as a cognitive map* (O'keefe and Nadel, 1978). There is, however, some controversy over this hypothesis. For instance, Buzsáki and Tingley (2018) argue that the hippocampus instead performs more general computations for organising and accessing the sequential structure of sensory experiences. Under this condition place cells could still develop, whose emergence would be the result of similar places in an environment producing similar sensory experiences. However, sensory experiences can be both exteroceptive and interoceptive, and there is evidence showing that place cell behaviour changes with respect to the same place if the context within that place changes, supporting the claim of Buzsáki and Tingley (Pastalkova et al., 2008). Perhaps place cells are better thought of as 'state' cells – abstract representations of sensory and contextual information. Yet, what is perhaps most important is the fact that place cells are, on the whole, rather stable representations of places/states. And due to this stability it is possible to continue measuring the activity of place cells after a rodent has explored an environment, and to infer recent sensory experiences through measurements of those place cells.

1.1.2 Hippocampal Replay

Wilson and McNaughton did just this type of post-exploration place cell monitoring, and in so doing they made a rather interesting discovery. By measuring the activity of place cells whilst a rodent was asleep immediately following a period of exploration in a maze, they discovered that those place cells that were active during awake exploration were more likely to be co-active during the subsequent post-exploration sleep session (Wilson and B. L. McNaughton, 1994). Not long after, Skaggs and B. L. McNaughton (1996) extended this result further, showing that not only were those cells more likely to be co-active during post-exploration sleep, but they also retained their same temporal sequence as was exhibited during exploration. They termed this phenomenon *replay*.

Wilson did not stop in his quest to measure hippocampal replay, and later discovered that replays occurred not only during sleep, but also during states of wakefulness as well. And perhaps even more interestingly, a number of these awake replays were found to be replays propagating in the reverse direction to that in which the prior experience occurred (Foster and Wilson, 2006) – what is now termed *reverse replay*. Reverse replays occur most commonly when a rodent has reached and is consuming an appetitive reward such as food or sugary water (Ambrose, Pfeiffer, and Foster, 2016). Hence it is now hypothesised that hippocampal reverse replay is heavily involved in biological reinforcement learning (Foster, 2017; Ólafsdóttir, Bush, and Barry, 2018).

With this new source of interesting neurophysiological and behavioural data in the form of hippocampal replay, there has been ongoing research attempting to enhance our understanding of it, and particularly to ground it in scientific theory through mathematical and computational modelling.

1.1.3 Computational and Robotic Modelling of Hippocampal Replay

A number of computational models attempting to describe the biophysical mechanisms of hippocampal replay have emerged, predominantly over the past decade (see Chapter 3 and Whelan, Vasilaki, and Prescott, 2019 for reviews), and a smaller selection of those have attempted to prove the potential of their models for solving spatial-memory and goal-oriented navigation tasks (Haga and Fukai, 2018; Cutsuridis and Hasselmo, 2011). In these studies, however, the inputs (mostly place cell activities) that are provided to the models do not represent the noisier and less constrained place cell activity one may find in reality. But more importantly, many of these models do not close the loop between hippocampal replay, its impact on actions and behaviours, and then the subsequent result of these altered behaviours on hippocampal replay (and so on). This leads to the models failing when one attempts to embody them in a simulated or robotic agent.

However, robotics is now playing a progressively more influential role in aiding our understanding of biological systems, and is enabling us to test our hypotheses within a more realistic framework; a framework that allows us to ground theoretical models of cognition with its effects on action and behaviour in physical systems. In the realm of cognitive neuroscience for instance, Prescott, Camilleri, et al. (2019) make the point that "a physical model in the form of a robot can stand as an existence proof ... for the sufficiency of the model and theory." In other words, robots are now allowing us to understand the integration between the body, brain and mind; what Verschure (2013) has termed *closing Vico's loop*, after the 18th century philosopher Giambattista Vico, who first emphasised the idea that "what I cannot create, I do not understand". (*Verum et factum reciprocantur seu convertuntur.*)

And Webb (2001) has eloquently expressed the usefulness of embodying biological models in robots: “understanding biology to build robots, and building robots to understand biology.”. This need not necessarily be in the form of physical robots either, which often contain unnecessary difficulties related to technical, rather than theoretical, issues. Simulated robots for instance can provide a useful platform for testing models that require interaction with an environment, whilst having the advantage of simpler implementation through greater environmental control.

Yet despite a number of the hippocampal computational models being suitable for robotic implementations, there has as of yet been little progress in integrating them within robots. It is within this space in particular that the work presented in this thesis looks to explore.

1.1.4 Hippocampal Replay and its Role in Reinforcement Learning

An interesting yet difficult problem that is challenging both the robotic/AI and neuroscience/psychology communities is the problem of reinforcement learning (RL).

The neural mechanisms of reinforcement learning can be traced back to Schultz’s seminal work on dopamine as a reward-predicting error signal (Schultz, 1998), and a recent review on the ventral basal ganglia (VBG) – a region heavily innervated with dopaminergic neurons (Ikemoto, Yang, and Tan, 2015) – has shown that the hippocampal region projects to and possibly receives projections from the VBG (Humphries and Prescott, 2010). Indeed, experimentally there is strong evidence that interactions between the hippocampus, VBG, and ventral tegmental area support reward-guided memory and conditioned place preference (CPP) (Gomperts, Kloosterman, and Wilson, 2015; Foster and Wilson, 2006; Trouche et al., 2019). For instance, recent experimental results have shown that hippocampal replays and sharp-wave ripples coordinate with bursts of activity in the ventral tegmental area (Gomperts, Kloosterman, and Wilson, 2015) and ventral striatum (Pennartz et al., 2004), and that changes in reward modulates the rate at which hippocampal reverse replays, but not forward replays, occur (Ambrose, Pfeiffer, and Foster, 2016). It has even been shown in a recent study on humans that spatial memory is prioritized for rewarding locations “retroactively", suggesting that reward-prioritized spatial memory appears some time after an event has occurred (Braun, Wimmer, and Shohamy, 2018). Perhaps it is hippocampal replay in the interim that modulates the memory?

Hippocampal replay in coordination with dopaminergic activity therefore seems well suited as a potential mechanism for reinforcement learning. A number of models have looked to incorporate dopamine as a neuromodulatory third factor in three-factor learning rules for synaptic plasticity (see (Gerstner et al., 2018) for a review), successfully showing, for instance, behavioural changes for conditioned place preference in a simulated Morris water maze task (Vasilaki, Frémaux, et al., 2009). Traditionally, reinforcement learning algorithms have only partially resembled biology, even if they take their inspiration from it (Sutton and Barto, 2018)). However, some of the reinforcement learning algorithms, such as DynaQ algorithms and the deep Q-network, seem well suited as explanations for the use case of hippocampal replay with their need for offline *sequence replays* (Aubin, Khamassi, and Girard, 2018; Cazé et al., 2018; Johnson and Redish, 2005; Mattar and Daw, 2018; Mnih et al., 2015; Sutton and Barto, 2018).

1.2 Motivation for this Thesis

Hippocampal reverse replay seems to have an important role in biological reinforcement learning, and whilst there now exists a number of computational models of hippocampal replay in the literature, there is very little in terms of robotic modelling. Given this, I have intended here to make a scientific contribution to our understanding of hippocampal replay by following the philosophy of robotic embodiment, through developing and then embodying a model of hippocampal reverse replay for solving a robotic reinforcement learning task. Or, in rewording the statement from Webb (2001) above for this specific instance – modelling hippocampal replay in order to improve robotic reinforcement learning, whilst using displays of robotic reinforcement learning to better understand hippocampal replay's role in biological reinforcement learning.

1.3 Structure of the Thesis

Chapter 2 begins by exploring the structure of the hippocampus through a review of what is currently known about its neuroanatomy. This sets the necessary groundwork for appropriately developing a computational model of the hippocampus.

Chapter 3 reviews computational models of hippocampal replay developed primarily over the most recent decade. A discussion on their applicability, as well as the challenges, towards integrating the models in robotic applications is given here.

Chapter 4 then sets out to take two very recent models of hippocampal reverse replay and develops them further for implementing in the MiRo robot.

Chapter 5 explores the role that hippocampal reverse replays in particular might play in reinforcement learning, by augmenting a hippocampal-striatal network originally designed to solve an RL spatial navigation task with the model of hippocampal reverse replay.

Chapter 6 concludes by summarising the content of the thesis and then looks towards future research directions.

Chapter 2

Review of Hippocampal Formation Neuroanatomy

One could argue that the literature on the hippocampal region is greater and more vast than most other regions of the brain – whether it pertains to that of rats, monkeys or humans. It has a long history of study dating back to the era of the Alexandrian schools, with the ancients naming it *cornu ammonis*; Latin for “horn of the ram" due to its resemblance to a ram’s horn. Such a naming convention is preserved today, with the three subfields of the hippocampus each named CA1, CA2 and CA3. The term *hippocampus* was coined by Giulio Cesare Aranzi (circa 1564) as a consequence of its resemblance to a seahorse (see Figure 2.1); hippocampus being derived from the Greek for seahorse. Since there is a vast amount of historical and contemporary study on the hippocampal region, it becomes a strenuous challenge to know how to begin with compiling that rich source of neuroanatomical information into a single, concise chapter (and that’s before discussing the literature on the putative functions of the hippocampal region).

This task has been heroically accomplished by David Amaral and Pierre Lavenex, in their chapter on hippocampal neuroanatomy in *The Hippocampus Book* (Amaral and Lavenex, 2007). At 77 pages, it is difficult to deem it a short chapter, but it could certainly be regarded as a comprehensive review of the literature. The review includes the discoveries from the early pioneering work of Santiago Ramón y Cajal (1852 - 1934) and Rafael Lorente de Nó (1902 - 1990), and journeys right up to contemporary experimental results (at the time of publication of the chapter, dated at 2007). The review written here will be an even more concise summary of the literature than Amaral and Lavenex’s, with only those experiments most important to the discussion being cited. This should in no way reflect the quality of Amaral and Lavenex’s chapter which, as a trainee researcher, has been a valuable example of how a literature review should be communicated. The only advantage I have over Amaral

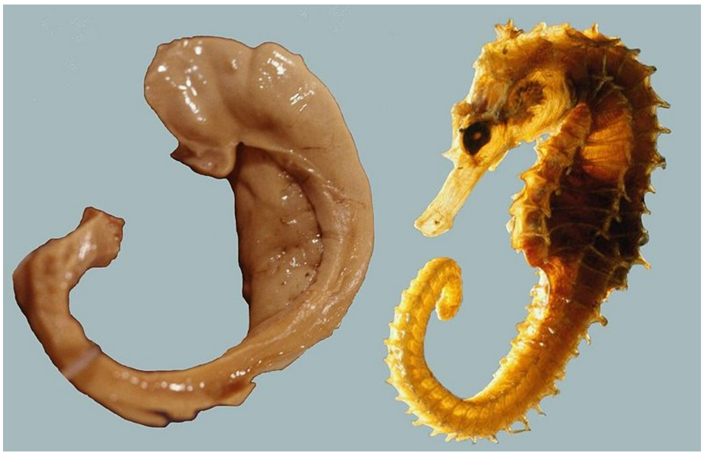


Fig. 2.1 The term *Hippocampus* is derived from the Greek for seahorse, and one can aptly see why given its resemblance to the human hippocampus. Figure freely available under the Creative Commons Attribution-Share Alike 1.0 Generic Licence (image source: https://commons.wikimedia.org/wiki/File:Hippocampus_and_seahorse.JPG).

and Lavenex is that I write now 12 years after the publication of their chapter, giving me access to literature that was not yet published at the time of their writing. One important contribution of the review offered here, that was not included in the review of Amaral and Lavenex, is the existence of projections between the hippocampus and the striatum. This connection however is a key factor in the computational model developed in Chapter 5.

It is, of course, impossible to relate all the information that has been generated over the decades and centuries pertaining to hippocampal anatomy. But with the best of intentions, this review aims to document the most important aspects of hippocampal anatomy in order to help put other literature on functional properties into perspective, and to justify decisions made in modelling.

This Chapter is organised as follows: Section 2.1 is a description of the general layout of the hippocampal formation (HF) and details the principal excitatory projections of the HF. In a sense, it is a summary of the subsequent sections that each present more detailed descriptions of HF anatomy. For some hippocampal modellers, this simple description may be enough for them to get started. The sections following this, sections 2.2-2.5, are a more detailed description of each of the main regions within the HF, starting with the entorhinal cortex (Section 2.2), dentate gyrus (Section 2.3), hippocampus (Section 2.4) and subiculum (including both presubiculum and parasubiculum, Section 2.5).

As mentioned already, acknowledgements must go to Amaral and Lavenex (Amaral and Lavenex, 2007) for their excellent review of the literature on hippocampal neuroanatomy. Their own review, in an elegant way, formed the map that guided my own exploration into the literature on hippocampal neuroanatomy.

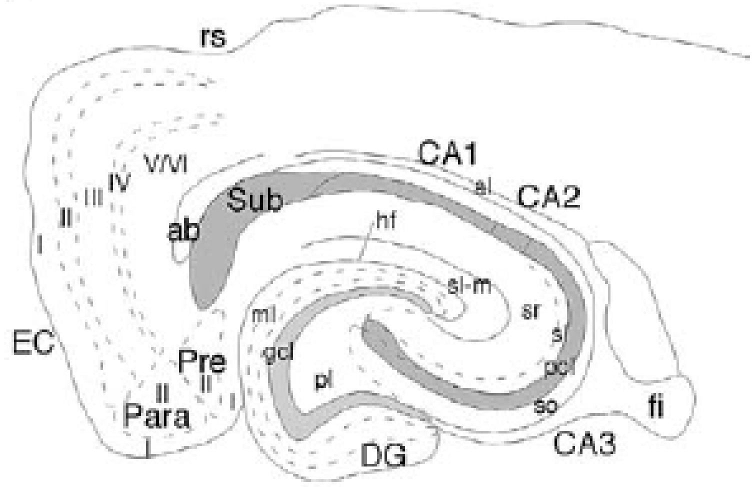


Fig. 2.2 Line drawing depicting a horizontal section of a caudoventral portion of the HF in the rat. All primary sections of the HF are shown, including the cell layers. Abbreviations: EC, entorhinal cortex; DG, dentate gyrus; fi, fimbria; Sub, subiculum; Pre, presubiculum; Para, parasubiculum; ab, angular bundle; ml, molecular layer of DG; gcl, granule cell layer of DG; pl, polymorphic layer of DG; so, stratum oriens of hippocampus; pcl, pyramidal cell layer of hippocampus; sl, stratum lucidum of CA3; sr, stratum radiatum of hippocampus; sl-m, stratum lacunosum moleculare of hippocampus; hf, hippocampal fissure. (*Source*: Adapted from (Amaral and Lavenex, 2007, p. 51).)

2.1 Overview of Hippocampal Formation Neuroanatomy

Firstly, it must be noted that the following description of hippocampal neuroanatomy pertains to that of the rat, due primarily to it being the largest source of experimental data. And whilst there exists many anatomical similarities between rats and, say, humans, there are also some noticeable differences.[1] This should be borne in mind as one reads through this review.

Secondly, the terminology employed here is to use the term *hippocampus* specifically for the regions CA1, CA2 and CA3. *Hippocampal formation* (HF) is used to denote the larger group of regions surrounding the hippocampus, as well as the hippocampus itself. Or more specifically, HF refers to the entorhinal cortex (EC), dentate gyrus (DG), hippocampus (or hippocampus proper), subiculum, presubiculum and parasubiculum. This follows the convention used by Amaral and Lavenex (2007). *Intrahippocampal* and *extrahippocampal* are used to denote connections (or otherwise) as *within* and *outside* the hippocampal formation, respectively.

[1] Rats for instance have a strong commissural connection between regions CA1, CA3 and dentate gyrus (Gottlieb and Cowan, 1973). Macaques and presumably humans on the other hand have much weaker HF commissural connections, with almost no commissural connections to the dentate gyrus (Amaral, Insausti, and Cowan, 1984).

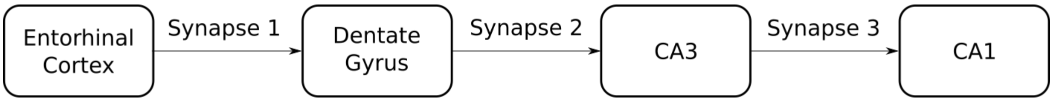


Fig. 2.3 The “trisynaptic circuit”, consisting of excitatory projections in the hippocampal formation as originally discovered by Andersen, Bliss, and Skrede (1971). Synapse 1 arises from axon terminations in the perforant pathway; synapse 2 due to axon terminations in the mossy fibres; and synapse 3 due to axon terminations in the Schaffer collaterals.

The HF is quite unique in its structure when compared to other regions of the cortex. For one, it is highly laminar, and its cells are positioned in distinct layers (Figure 2.2). Furthermore, unlike the neocortex where bidirectional connectivity tends to be the norm, in the HF unidirectional connections are more prominent (Amaral and Lavenex, 2007). One of the first attempts at describing the system of connectivity within the HF was the “trisynaptic circuit”, a series of excitatory HF pathways discovered by Andersen, Bliss, and Skrede (1971) via measurements of neuronal responses to EC stimulation.

The circuit was given as follows: Axons from EC cells travel through the perforant pathway and synapse onto granule cells of the DG. From there, DG cell axons, through the mossy fibres, synapse onto the CA3 pyramidal cells, from which CA3 axons travel through the Schaffer collaterals to synapse onto the CA1 pyramidal cells (Andersen, Bliss, and Skrede, 1971). Finally, via the alveus, axons from CA1 pyramidal cells travel to the fimbria, which subsequently projects to other subcortical structures (such as the septum and hypothalamas, amongst others). This “trisynaptic circuit” is summarised in Figure 2.3.

A second consequence of the experiment conducted by Andersen et al. (Andersen, Bliss, and Skrede, 1971) was the development of the *lamellar hypothesis*. In essence, the lamellar hypothesis states that the excitatory “trisynaptic circuit” exists along “slices”, or “lamellae” (such as the slice shown in Figure 2.2), parallel to the transverse axis of the HF. However, nearly two decades later, as a rebuttal to the lamellar hypothesis, Amaral and Witter (1989) published an updated review of the anatomical structure of the HF and stated clearly that the “*overwhelming consensus in all these studies* [their own as well as others] *is that aside from the mossy fibres, none of the intrinsic connections of the hippocampal formation is organized in a lamellar fashion.*” They found that connections along the longitudinal axis of the HF were just as prominent as those found across the transverse axis. But another two decades later again, Sloviter and Lømo (2012) gave their updated review on the lamellar hypothesis and, unsurprisingly, showed it to be much more nuanced than had previously been thought. They argued that the literature did not undermine the lamellar hypothesis. Yes, there are “translamellar” (connections traversing across the lamellae, i.e. longitudinal) axon projections in the HF[2], but that lamellae functionality could still operate even with longitudinal connections.

Whilst the trisynaptic circuit is of historical importance, it is now seen as only a small portion of the HF connectivity. Subsequent discoveries have shown that there are projections from CA1 to the subiculum and EC (Amaral, Dolorfo, and Alvarez-Royo, 1991; Köhler, 1985), and from the presubiculum and parasubiculum to the EC – a rather important discovery showing that the HF circuit closes back on itself. Further discoveries have shown that there are reciprocal connections between the EC and other cortical regions, suggesting the EC acts as a relay for information to and from the hippocampus (Witter, 1993; Amaral and Lavenex, 2007), as well as discoveries of other hippocampal intrinsic connections. A more complete systems diagram of the HF is shown in Figure 2.4. The origin of this systems diagram will become more apparent as we move through the anatomy of each HF region individually, as we do next.

2.2 Entorhinal Cortex

Amaral and Lavenex (2007), in paraphrasing Ramon y Cajal, said that "*whatever the rest of the hippocampal formation is doing depends on what the entorhinal cortex has done*" (Amaral and Lavenex, 2007, p. 84). It would be a danger to under-appreciate the role the EC plays within the HF. It is largely the gateway through which sensory information reaches the hippocampus, and through which the HF's processed information is then relayed to the neocortex (Amaral and Lavenex, 2007). It seems fitting therefore that we start with this region of the HF.

2.2.1 Cytoarchitecture

The EC can be divided, roughly speaking, into two primary regions: the lateral entorhinal area (LEA) and the medial entorhinal area (MEA) (Witter, 1993), with each being composed of six layers (Figure 2.2). Layers I and IV are considered to be generally cell-poor (acellular), with the remaining cell layers (II, III, V and VI) being more cell dense (cellular).

The predominant cell types and projections of the cellular layers are as follows: Layer II is composed mostly of pyramidal cells and stellate cells, with both projecting to DG and CA3 through the angular bundle (Klink and Alonso, 1997); layer III is composed of various cells that project to CA1 and subiculum through the perforant pathway (Gloveli et al., 1997),

[2] Though the mossy fiber pathway is undoubtedly lamellar, something that even Amaral and Witter (1989) admitted to.

with the most numerous cell type being pyramidal cells. Whilst the superficial layers II - III project *into* the HF, the deep layers V-VI project primarily *away* from the HF and to the neocortex (and therefore have afferents from the HF). The principle cell types of the layer V neurons are pyramidal (the most numerous), horizontal and polymorphic, with little morphological differences between the LEA and MEA[3] (Hamam, Kennedy, et al., 2000; Hamam, Amaral, and Alonso, 2002). Finally, the layer VI cells consist of a wide variety of cell types that project associationally to other layer VI cells, as well as to superficially located cells in layers I-III, and projection neurons that travel towards the deep white matter (Amaral and Lavenex, 2007).

The MEA and LEA, although having some overlap at their border, can largely be divided by their afferent/efferent connections – but also by the cytoarchitectonic characteristics, and histological/histochemical analyses of each region (Blackstad, 1956; Insausti, Herrero, and Witter, 1997). Insausti, Herrero, and Witter (1997) have shown that the EC could perhaps be divided into *six* subregions, adding further to its divisional complexity. In terms of afferent connections, most of the differences lie in the superficial layers of the EC, which are the principle terminating layers from other cortical regions as well as the principle projection layers to the DG and hippocampus (Hamam, Amaral, and Alonso, 2002). We will see next how the afferent connections, and later how the efferent connections, differ in these two areas.

2.2.2 Extrahippocampal Connections

EC superficial layers receives its major afferent inputs from the postrhinal and perirhinal cortices,[4] but there is some distinction in how strongly these two regions project into the LEA and MEA. Perirhinal cortex projects preferentially into the LEA, whereas postrhinal cortex projects more strongly into the MEA but with some projections to the LEA as well (Burwell, 2000). EC also reciprocates the projections back to the perirhinal and postrhinal cortices, with perirhinal cortex receiving return projections almost exclusively from the LEA, whereas the postrhinal cortex receives projections back from both the LEA and MEA (Burwell and Amaral, 1998). Thus these return projections largely follow the topography of the forward projections.

[3] This suggests that most of the functional differences between LEA and MEA are within the superficial layers of the EC (Hamam, Amaral, and Alonso, 2002)

[4] Which, along with the EC, is sometimes collectively called the “parahippocampal region", not to be confused with the “parahippocampal area" of monkeys and humans, which can be thought of as the postrhinal equivalent in the rat (Burwell, 2000).

2.2.3 Intrahippocampal Connections

There is a distinct EC caudolateral-rostromedial relationship to septotemporal levels of the DG, in which EC areas that are more caudolaterally situated have stronger projections to more septal regions of the DG, whilst rostromedially situated EC regions preferentially project to more temporal potions of the DG (Dolorfo and Amaral, 1998a). Given the LEA is situated more rostrolaterally, whilst the MEA is more caudomedial, one will notice that both MEA and LEA project to portions along the whole septotemporal axis of the DG (see Dolorfo and Amaral (1998a) for figures that help clarify the topography described here). But, there is a non-overlapping effect of projections from caudolateral EC portions to the septal half of the DG, medial EC portions to the third quarter of the DG, and rostromedial portions of the EC to the remaining temporally located quarter of the DG. Although the work of Dolorfo and Amaral (1998a) is cited as evidence for hippocampus afferent EC connections being topographically equivalent to DG (see for instance (Hargreaves et al., 2005)), the author is currently unaware of any experiment explicitly supporting this claim. Rather, earlier experimental evidence showed that these areas do not project equally, and that caudomedial EC (the location of MEA) projects more strongly to DG but weakly to CA1, whilst more rostrolateral EC (the location of LEA) projects more strongly to CA1 but weakly to DG. The most rostrolateral portion of EC projects only to CA1 (Witter, Griffioen, et al., 1988). More on the entorhino-hippocampus connections is discussed in Section 2.4.

Finally, the EC shows to have intrinsic associational connections, both within each of the MEA and LEA, but also across the two regions as well (Dolorfo and Amaral, 1998b). It was shown that projections originate in layers II, III and the deep layers, but preferentially terminate within the superficial layers of the EC. To quote Dolorfo and Amaral (1998b), this "*raises the possibility that the cells of origin of the perforant path projections* [i.e. DG, hippocampus and subiculum] *receive substantial input from cells in both the deep and superficial layers*." How this could affect HF function remains unclear.

2.3 Dentate Gyrus

2.3.1 Cytoarchitecture

The dentate gyrus (DG) is composed of only three cell layers, as opposed to the six found in the EC (Figure 2.2). The number of cells in the rat DG has been shown to be on the scale of 0.7–1.2 $\times 10^6$ depending on the rat species and age (Boss, Peterson, and Cowan, 1985; West, Slomianka, and Gundersen, 1991). The most superficial layer, as with the EC, is acellular, and is termed the molecular layer. The middle layer in contrast is cell-dense, called the granule cell layer, and is the principle cell layer of the DG. As is obvious from its name, this layer is composed principally of granule cells, and are the only neurons that give rise to axons that leave the DG (Amaral and Lavenex, 2007). The third layer, the deepest, is called the polymorphic cell layer. Lorente de Nó originally termed this cell layer as the CA4 region of the hippocampus, with this term sometimes still being used today. But, as noted by Blackstad (1956) in his histological analysis of the HF using the silver impregnation method, this layer is better described to be a part of the *area dentata*, or dentate gyrus.

2.3.2 Intrahippocampal Connections

We saw previously that the DG has extensive afferent input from the EC, originating mainly from layer II of the EC (Klink and Alonso, 1997). These EC axonal projections travel through the perforant pathway and terminate preferentially on the dendritic spines of the granule cells – the granule cell dendritic spines are found in the molecular layer (Hjorth-Simonsen and Jeune, 1972). Its only output is to the CA3 region through the mossy fibres (Gaarskjaer, 1978). The border of CA3 where the mossy fibres do not project to is part of what distinguishes CA3 from CA2, since CA2 receives no mossy fibre input (Amaral and Lavenex, 2007).

Interneurons

It is worth spending some time here to point out the existence of interneurons within the DG that sit intermingled amongst the principle cells and form synapses with the DG granule cells. Freund and Buzsáki (1996) give a comprehensive overview of the interneurons within the HF. Many HF models will often simplify the role of interneurons, having them act simply as inhibitory modulators. But interneurons are highly heterogeneous and, paraphrasing Freund and Buzsáki (1996), this heterogeneity allows them to perform multiple tasks in parallel and to act as important assistants in managing the cooperative activity of large cell populations. Most interneurons are indeed "GABAergic", producing the inhibitory neurotransmitter GABA, and in fact Freund and Buzsáki (1996) take the stance that hippocampal interneurons should be termed simply "GABAergic nonprincipal cells". And Amaral and Lavenex (2007) argue that given the current definition of interneurons, then the mossy cells, found in the polymorphic layer, should also be classed as interneurons given that their axons rarely escape the confines of the DG (but do extend ipsilaterally and commissurally across the DG) (Berger, Semple-Rowland, and Bassett, 1981). But, most distinctly, the mossy cells are glutamatergic, an excitatory neurotransmitter, and are thus highly atypical interneurons. The mossy cells act therefore as excitatory associative connections within the DG (Amaral and Lavenex, 2007).

Despite the homogeneous sounding name of “interneuron”, the interneurons are rather diverse, with axo-axonic, basket, HICAP, MOPP and HIPP cells being found within the DG, and many more within other HF regions (Freund and Buzsáki, 1996). Furthermore, one should be careful in collating these various interneurons into a single, uniform cell population having only a single function, such as global inhibition, when modelling, as interneurons appear to be much more varied than this. Although this review will not go into a great deal of detail with respect to hippocampal interneurons as that requires a lengthy review in its own right, it is important to be aware of their existence and the potential roles they could play in hippocampal formation functioning.

2.3.3 Extrahippocampal Connections

The most prominant and well studied extrahippocampal projection to the dentate gyrus is from the septal nuclei. The two mains areas from which septal afferents project into the DG (and also the hippocampus) are the medial septum and the nucleus of the diagonal band of Broca (Nyakas et al., 1987). The connections are topographically organised however, such that the more medially situated diagonal band of Broca cells project preferentially to septal or dorsal portion of DG, whilst more laterally situated medial septum nucleus cells and diagonal band of Broca cells project to temporal portions of DG (Nyakas et al., 1987; Amaral and Lavenex, 2007). A good proportion of these projections, but not the majority, are cholinergic, producing the neurotransmitter acetylcholine (Wainer et al., 1985). It has been proposed that the non-cholinergic projections could contain the neurotransmitters substance P (Baisden, Woodruff, and Hoover, 1984) and/or GABA (Köhler, Chan-Palay, and Wu, 1984).

There is also a somewhat substantial though rather diffuse projection from the supramammillary area of the hypothalamas onto the DG principle cells, or the granule cells (Maglóczky, Acsády, and Freund, 1994), with these most likely being excitatory glutamatergic projections (Kiss et al., 2000). Other projections arrive from the brain stem, including noradrenergic from the pontine nucleus locus coeruleus (Loughlin, Foote, and Grzanna, 1986), serotenergic input from the raphe nuclei (Vertes, Fortin, and Crane, 1999), and rather weakly diffuse dopaminergic projections from the ventral tegmental area (Scatton et al., 1980). All the mentioned projections from the brain stem terminate predominantly in the polymorphic layer of DG (Amaral and Lavenex, 2007).

2.4 Hippocampus

The hippocampus proper has had much of the attention when it comes to the medial temporal lobe area. It was here that place cells were first discovered, in area CA1 (O'Keefe and Dostrovsky, 1971a), whilst area CA3 garners a great deal of interest due to its high level of recurrent connectivity and its subsequent functional roles in auto-associative memory (Le Duigou et al., 2014). Much of the modelling will be based upon the hippocampus, so care should be taken in ensuring we represent its anatomy with the correct level of detail.

2.4.1 Cytoarchitecture

The hippocampus, or hippocampus proper, has conventionally been divided into at least three primary subregions, CA1, CA2 and CA3 (CA being Cornu Ammonis). The number of cells in regions CA1 and CA3 is around $0.32 - 0.42 \times 10^6$ and $0.21 - 0.33 \times 10^6$ (Boss, Turlejski, et al., 1987), which is 2-3 times less than the number of DG cells (Boss, Peterson, and Cowan, 1985). Regions CA1 and CA2 are composed of four cell layers, with CA3 containing all four layers of regions CA1/2 but also having an additional layer, thus distinguishing it from CA1/2 (Amaral and Lavenex, 2007).

The four cell layers occupying all CA regions are given as follows, in order from deep to superficial (see Figure 2.2): Stratum oriens, the infrapyramidal acellular region containing a smaller proportion of the CA3-CA3 associational and CA3-CA1 Schaffer collateral connections (Amaral and Lavenex, 2007); pyramidal cell layer, containing the principle pyramidal cells of the hippocampus (Amaral and Lavenex, 2007); stratum radiatum, the suprapyramidal region containing the larger number of CA3-CA3 associational and CA3-CA1 Schaffer collateral connections; stratum lacunosum-moleculare, the region in which EC afferents terminate.

The additional layer within region CA3 is known as the stratum lucidum, occupying a space between the pyramidal cell layer and stratum radiatum (Figure 2.2), and it is within this region that the mossy fibers from DG terminate (Amaral and Lavenex, 2007). Stratum lucidum also contains a number of interneurons that have been speculated to be important for local circuitry in CA3, with recurrent projections to stratum lucidum and stratum radiatum (Spruston, Lübke, and Frotscher, 1997).

Though it is the additional layer of CA3 that distinguishes it from CA1/2, a noticeable feature that distinguishes CA1 from CA2/3 is the homogeneity of its dendritic branches and the size of its soma population. CA1 cells, on the whole, have average total dendritic lengths (sum of apical and basal dendrites) of $13 - 17$mm and a soma size averaging 15μm (Ishizuka, Cowan, and Amaral, 1995; Pyapali et al., 1998). CA2/3 regions however have much more heterogeneous dendritic branches, as well as much larger soma sizes. Their dendritic length and soma size increases as one moves from proximal CA3, near the DG, to distal CA2, near CA1, with dendritic lengths varying from $8-18$mm, and soma size varying from $20-30\mu$m (Ishizuka, Cowan, and Amaral, 1995) along the proximal-distal axis.

Finally, as with the DG, it is worth noting that there is a large variety of interneurons within the hippocampus, such as axo-axonic, basket, bistratified, horizontal trilaminar, radial trilaminar and O-LM cells (Freund and Buzsáki, 1996).

2.4.2 Intrahippocampal Connections

As mentioned previously, it is only CA3 that receives mossy fibre input from the DG, and is thus a major feature distinguishing CA3 from CA1 and CA2. The EC however projects to the stratum lacunosum-moleculare in all CA areas of the hippocampus, but not all is equal. For one, CA2/3 receive their input from layer II of EC whilst CA1 receives its input from layer III of EC (Witter, 1993; Amaral and Lavenex, 2007), indicating that CA2/3 receives a similar EC input to that of the DG's EC input. Furthermore, there is a distinct topographic projection from EC to CA1 – lateral EC projects to more distally located CA1 cells (close to the subiculum) whilst medial EC projects to more proximally located CA1 cells (near CA2) (Amaral and Lavenex, 2007). Finally, CA1 is the only CA area that projects back to the EC, terminating in the deep layers, V and VI, of the EC. The same topography as the forward projections is preserved for the return projections (Naber, Lopes da Silva, and Witter, 2001).

Perhaps one of the most intriguing findings of the hippocampus is the heavy associational connections of CA3. A typical CA3 cell has shown to be innervated by approximately 5% of all other CA3 cells (Buzsáki, 1989). Furthermore, the CA3 pyramidal cells that send their axons out to make associational connections also contain axon collaterals that project to all the other CA regions of the hippocampus, both ipsilaterally and commissurally (Ishizuka, Weber, and Amaral, 1990). Thus, whatever information CA3 is sending associationally to itself, it seems it is also sending to both CA2 and CA1. The terminating layers of the CA3 projections are primarily stratum radiatum and stratum oriens (for all CA regions), indicating terminations both on the apical and basal dendrites of the pyramidal cells (Ishizuka, Weber, and Amaral, 1990).

The axon collaterals that branch off from the assocational CA3 connections and that lead to the CA3 to CA1 projections are known as the *Schaffer collaterals* (Amaral and Lavenex, 2007), and a single CA3 axon has been shown to travel along at least two-thirds of the septotemporal length of the hippocampus (Li et al., 1994). Thus the collaterals along the septotemporal (longitudinal) axis travel far and wide. Along the transverse axis however, there is a more point-to-point projection, such that more proximally located CA3 cells (close to DG) project more heavily to distally located CA1 cells, whilst more distally located CA3 cells (near CA1) project more to proximally located CA1 cells (which are thus closer to CA2/3) (Ishizuka, Weber, and Amaral, 1990).

Finally, we note CA1's projection to the subiculum, which follows the same transverse topographic projections of the Schaffer collaterals, that is, proximal CA1 projects more to distal subiculum and distal CA1 projects more to proximal subiculum (Amaral, Dolorfo, and Alvarez-Royo, 1991). Septotemporally however, a single cell in the septal portion of CA1 projects divergently to the septal third portion of the subiculum (i.e. it could project to any point in this portion, there is no one-to-one mapping), with this also being true of mid and temporal CA1 cells projecting to middle and temporal thirds of subiculum, respectively (Tamamaki, Abe, and Nojyo, 1987). This type of topography has been described as "columnar organisation" by Tamamaki, Abe, and Nojyo (1987).

2.4.3 Extrahippocampal Connections

In the rat, there is perhaps a small monosynaptic connection to the neocortex from the CA1 region, but there are at least no known connections between neocortex and CA2/3 (Amaral and Lavenex, 2007). Interestingly, one can contrast this with the primate brain where, for instance, the marmoset monkey has been shown to have sparse reciprocal connections between all CA regions and the neocortex (Schwerdtfeger, 1979). This may indicate something important regarding the difference in cognitive abilities between rodents and primates.

Both the dorsal and ventral portions of the CA1 region have been found to send projections to the shell of the ventral striatum in the basal ganglia (Swanson and Cowan, 1977; Trouche et al., 2019). This is an important connection that seemed not to be included in the review of (Amaral and Lavenex, 2007). Later, in Chapter 5, is an hippocampal-striatal model of reinforcement learning for a spatial navigation task. This connection is of special importance therefore, and more is mentioned regarding this in the model description of Chapter 5.

The septum is largely the only subcortical structure that has connections with the hippocampus, with connections occurring between the septum and all CA regions (Amaral and Lavenex, 2007). Like the DG septal inputs, the hippocampus septal inputs, through the fimbria/fornix pathway, originate from the medial septal nucleus and the diagonal band of Broca (Meibach and Siegel, 1977). The hippocampus then returns projections back to the lateral septal nucleus (Swanson and Cowan, 1977).

The temporal portion of the hippocampus receives projections from, and sends projections to, the amygdaloid complex. More precisely, CA3 and CA1 receive inputs from the basal nucleus of the amygdala, terminating heavily in stratum oriens and stratum radiatum (Pikkarainen et al., 1999), and the return projections back to the amygdaloid complex originate from the pyramidal cells of CA1 (Pitkänen et al., 2000).

The hypothalamas too projects to the hippocampus, or more specifically, the supramammillary area of the hypothalamas projects to areas CA2/3 (Maglóczky, Acsády, and Freund, 1994). Interestingly, though CA2 and CA3 receive roughly the same inputs, the supramammillary input has a heavier projection to CA2 than to CA3 (Amaral and Lavenex, 2007). The significance of this heavier projection is still rather unclear.

It seems that both the septal connections and the hypothalamic connections are important in modulating the two hippocampal wave activities known as theta wave and slow wave activity (Buzsáki, 1989). When the medial septal area is lesioned, for instance, theta waves in the hippocampus seize to occur (Rawlins, Feldon, and Gray, 1979). Furthermore, supramammillary activity exhibits slow waves that are phase locked with hippocampal slow waves, but only when the septal connection is intact (Kirk and N. McNaughton, 1991). Kirk and N. McNaughton (1991) proposed that hippocampal wave activity (specifically encoding of the frequency) is modulated by the supramammillary area, which influences the medial septal area, following which the medial septal area modulates the hippocampus.

The nucleus reuniens of the midline thalamas also projects into area CA1, terminating in an overlapping region with the EC inputs (Dolleman-Van der Weel and Witter, 2000). And lastly, the brain stem has inputs into the hippocampus, namely noradrenergic inputs from the locus coeruleus and serotonergic inputs from the raphe nuclei into regions CA3 and CA1 (Loughlin, Foote, and Grzanna, 1986; Vertes, Fortin, and Crane, 1999).

2.5 Subiculum, Presubiculum and Parasubiculum

Unfortunately, and perhaps unjustifiably, the subiculum, presubiculum and parasubiculum – collectively termed the *subicular complex* – has not been exposed to the same level of study has have other HF regions. The subiculum particularly has numerous extrahippocampal connections, and plays a major role in closing the loop in the HF system. Amaral and Lavenex (2007) make the point that *“evidence has mounted that the subiculum is one of two primary output structures of the hippocampal formation"* – presumably the other structure being the entorhinal cortex – whilst H. Groenewegen et al. (1987) claimed the subiculum to be *“the main output structure of the hippocampal formation.”* And O’Mara et al. (2001) highlight an important detail in that the subiculum *“plays an important but ill-defined role both in spatial navigation and in mnemonic processing.”* Whilst some interesting work in the subiculum has since emerged, such as the potential discovery of boundary vector cells (Lever et al., 2009), let us hope that this seemingly important HF structure receives greater attention in the near future.

2.5.1 Cytoarchitecture

Subiculum

The subiculum is formed of three cell layers, and is largely contiguous with CA1. Layers strata radiatum and lacunosum moleculare of CA1 converge to become the molecular layer in the subiculum (O'Mara et al., 2001). CA1's pyramidal cell layer widens in size as it is replaced by the subiculum's pyramidal cell layer. Interestingly, it has been proposed that the pyramidal cell layer be divided further by two, due to the firing characteristics of its cells – it was found that there exists a set of regular spiking cells and a set of bursting cells, located preferentially in the superficial and deep portions of the pyramidal cell layer, respectively (Greene and Totterdell, 1997). The final layer is the polymorphic layer.

Not much is known about the interneurons of the subiculum, but it is known that there exists GABAergic interneurons that are speculated to play an important modulatory role in local subicular computations (Prida et al., 2006).

Presubiculum and Parasubiculum

The presubiculum is differentiated from the subiculum due to its tightly packed but small pyramidal cells, whilst the parasubiculum has similarly tightly packed and dense cells, but contains larger pyramidal cells (Amaral and Lavenex, 2007).

2.5.2 Intrahippocampal Connections

All regions of the subicular complex contain longitudinal associational connections, yet interestingly, whilst the subiculum does not contain commissural connections, the pre- and parasubiculum regions do (Groen and Wyss, 1990). Furthermore, unlike any other region of the HF, the axonal organisation of the bursting cells and regular spiking cells in the subiculum appears to be columnar and laminar, respectively: the bursting cells send their axons only to local cells within their local columns, whereas regular spiking cells send their axons more distributively across columns (Harris et al., 2001).

The subiculum projects to both the pre- and parasubiculum, following a topology such that septal regions of subiculum project to dorsal and caudal portions of pre-/parasubiculum, and temporal regions of subiculum project to venral and rostral portions of pre-/parasubiculum, whilst both pre- and parasubiculum project to one another (Groen and Wyss, 1990).

In the process of closing the hippocampal loop, the subicular complex projects reciprocally with EC. The perforant path, in its journey towards DG, CA3 and CA1 from EC, passes through the subiculum and forms synapses with the subiculum's pyramidal cells. Subsequently, the subiculum sends return projections to the deep layer of the EC. These forward and return projections largely follow the same topographically organised pattern as the EC-CA1 connections (Witter and Henk J Groenewegen, 1990; Amaral and Lavenex, 2007) (see Section 2.4.2). EC also receives projections from both presubiculum and parasubiculum, though whilst the parasubiculum projects to both medial and lateral EC, the presubiculum projects only to the medial EC (Caballero-Bleda and Witter, 1993).

2.5.3 Extrahippocampal Connections

As mentioned, the subicular complex has rather extensive extrahippocampal connections. This includes subiculum neocortical projections, specifically projections to the lateral and medial prefrontal cortex (Verwer et al., 1997) and retrospenial cortex (Wyass and Van Groen, 1992).

Beyond the neocortex, temporal portions of subiculum project to the amygdaloid complex, whilst the amygdaloid complex returns these projections back to the temporal portions of both subiculum and parasubiculum (Pitkänen et al., 2000). Subiculum and presubiculum both project to the mammillary nuclei of the hypothalamus, receiving return projections from the supramammillary region (Swanson and Cowan, 1975; H. Groenewegen et al., 1987; Groen and Wyss, 1990). In fact, subiculum's projection provides the mammillary body's heaviest input (Amaral and Lavenex, 2007).

Both pre- and parasubiculum are reciprocally connected with the anterior thalamic nuclei, whilst subiculum has no connectivity (Robertson and Kaitz, 1981; Kaitz and Robertson, 1981; Amaral and Lavenex, 2007). The subiculum does however have reciprocal connectivity with the nucleus reuniens, with the subiculum's afferents being similar to those of CA1 (Herkenham, 1978).

The most substantial subicular projection towards subcortical regions is towards the septal complex and the nucleus accumbens of the ventral striatum (along with the mammillary body, discussed above) (H. Groenewegen et al., 1987). Whilst the nucleus accumbens does not reciprocate these projections, the septal complex does return rather weak cholinergic projections to the subiculum, but heavy cholinergic projections to the pre- and parasubiculum (Amaral and Lavenex, 2007).

2.6 Hippocampal Neuroanatomy: Summary

Learning of the numerous intra- and extrahippocampal connectivities and neuron types of the HF, it can often feel one is simply more lost in understanding the HF than one was beforehand. In much of science it is often necessary to simplify a process in order to aid understanding – thermodynamics, for instance, simplifies the individual kinetic energy of millions of molecules to 'averages' so that we may understand the activity of those millions of molecules. And indeed the leaky-integrate and fire neuron is an example closer to home, where the dynamics of ions across neuron membranes are simplified to a single term(s) representing the 'average' behaviour of the ions on changes in neuron electric charge.

Therefore, in keeping with simplification in order to aid understanding, a simplified systems diagram of the HF, more detailed than the trisynaptic circuit of Figure 2.3, is presented in Figure 2.4. This figure represents the principle projections of the HF as discussed in the sections above. With this simplified yet explanatory description of the HF, we may feel more confident in answering where, why, and how hippocampal models are structured and developed as they are.

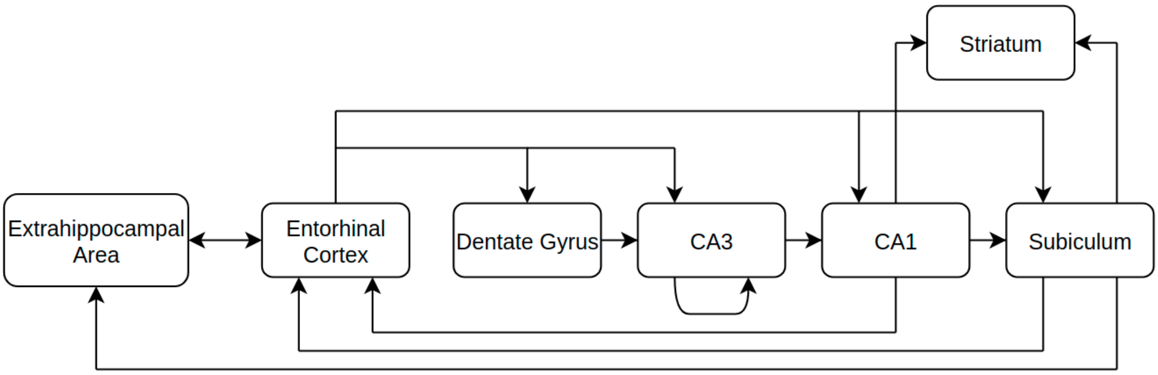


Fig. 2.4 A more complete hippocampal systems diagram, showing the principle excitatory projections throughout regions of the hippocampal formation. Most importantly, the transfer of information shows to loop back upon itself; there are prominent associational connections in CA3 (Amaral and Lavenex, 2007); and the connection from regions CA1 and Subiculum to the striatum, which will be important in a model later developed in Chapter 5. The extrahippocampal area represents mostly the rest of the brain and though, whilst not shown here, other regions apart from the EC and Subiculum do also connect with extrahippocampal regions (described in the text).

Chapter 3

Computational Models of Hippocampal Replay

This chapter provides a review of current computational models of hippocampal replay. A general review of the computational models is first provided, following which a more thorough examination of two computational models, (Haga and Fukai, 2018; Pang and Fairhall, 2019), is made. The dynamics and architectures of these two models form much of the basis for the robotic hippocampal replay model developed in the next chapter. The chapter concludes by analysing these models in the context of robotic applications.

Much of this chapter is based on a submission to the 8th International Conference on Biomimetic and Biohybrid Systems (Nara, Japan, July 9–12, 2019). See Whelan, Vasilaki, and Prescott (2019).

3.1 Overview of the Models

As discussed in the introduction chapter, reverse replays have been speculated to play an important role in biological reinforcement learning. Understanding the dynamics of hippocampal replay is therefore an important undertaking. Models of hippocampal replay are almost exclusively composed of neural networks with either rate-based or spiking-based neural dynamics, and most, if not all, necessitate the use of recurrent networks in order to store memory traces for later reinstatement. Furthermore, they mostly simplify the problem of place cell activation by assuming evenly distributed place fields, usually overlapping, in an environment for which specific place cells fire as a function of the agent's distance from the centre of the respective place field.

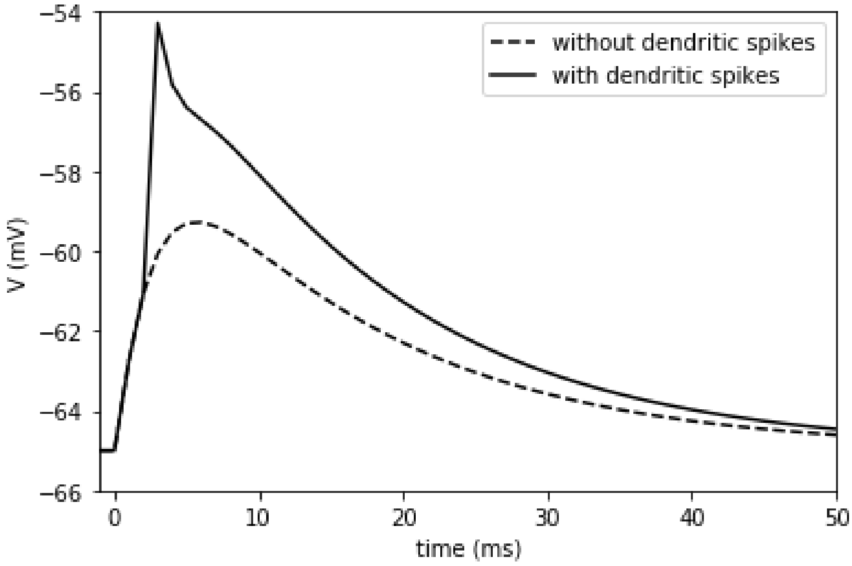


Fig. 3.1 Dendritic spiking causes supralinear responses to synchronous inputs (solid line) above what would be expected with a simple summation of inputs only (dashed line), important in the model by Jahnke, Timme, and Memmesheimer (2015) for modelling forward/reverse replays and sharp-wave ripples. This plot was modelled using a standard leaky-integrate and fire neuron receiving instantaneous synchronous inputs at t=0, with and without dendritic spiking (see main text).

We start with a spiking-based model of leaky-integrate and fire neurons by Jahnke, Timme, and Memmesheimer (2015). Here they exploit theta phase precession (O'Keefe and Recce, 1993) to generate memory traces via spike-timing dependent plasticity. But the key inclusion in their model is to use dendritic spiking, which occurs when a high number of synchronous inputs exceed some threshold Θ_b within a time interval of ΔT^s. This then causes a dendritic current impulse which causes an increase in membrane voltage above what would be expected without dendritic spiking (Figure 3.1).

Once a dendritic spike is initiated, the dendrite enters a refractory period during which time it cannot transmit any spikes. In a linear sequence of place cells with bidirectional connections, this refractory period is important for restricting replays to only travel in a single direction, without reversing back on itself (Gauy et al., 2018). Furthermore, the supralinear nature of the dendritic impulse generates activity pulses that are reminiscent of sharp-wave ripples.

Dendritic spiking, found to occur in CA1 pyramidal cells of the hippocampus (Ariav, Polsky, and Schiller, 2003), offers a unique explanation for the occurrence of both sharp wave ripples and replay, and Gauy et al. (2018) have extended the use of dendritic spikes, as modelled by Jahnke, Timme, and Memmesheimer (2015), but invented a new cell type termed 'sequence cells'. The reason for this inclusion is that Jahnke et al.'s model could not

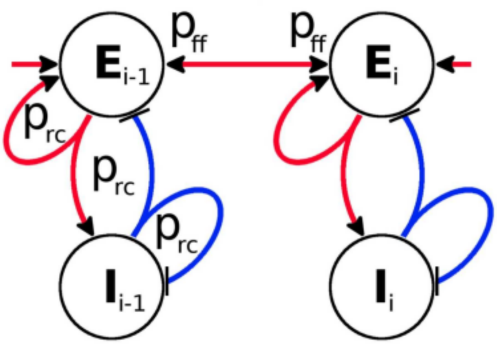


Fig. 3.2 A model by Chenkov, Sprekeler, and Kempter (2017) of a synfire-like chain of cell assemblies containing excitatory (**E**) and inhibitory (**I**) cell populations, recurrently connected to each other with probability P_{rc}, and connected feedforwardly with probability P_{ff}. Replay events are characterized as activity propagation from one cell assembly to another and so on, with activity modulated by the inhibitory interneurons. Figure adapted from (Chenkov, Sprekeler, and Kempter, 2017).

account for different trajectories containing the same place cells. Having multiple trajectories emanating from the same place cell would cause replays of multiple trajectories at once. Rather, sequence cells, activated in sequential order as an agent traversed an environment, were paired with place cells via Hebbian learning. As such, one needs only save individual trajectories of sequence cells, and reactivate them in order to reinstate the place cell sequences learned during exploration. It is worth noting that the assumption of sequence cells causes two possible issues: 1) this may necessitate an indefinite number of distinct sequences of sequence cells to account for all trajectories experienced; 2) there is no biological evidence for the existence of sequence cells (though for bioinspiration this may be irrelevant).

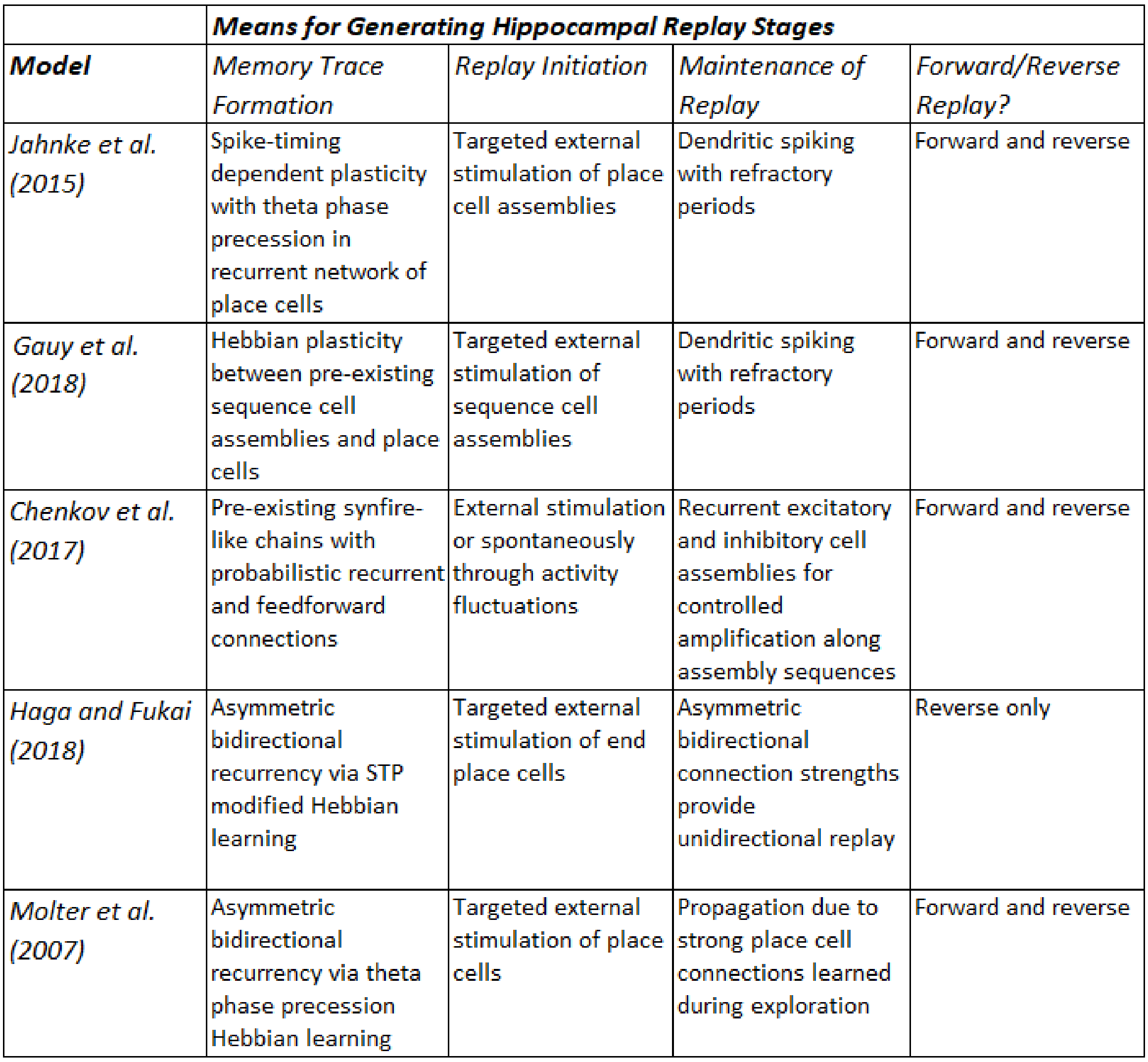
Matheus Gauy et al.'s model above had sequences of sequence cells arranged in a similar fashion to synfire chains, and Chenkov, Sprekeler, and Kempter (2017) designed a similar synfire-like chain of cell assemblies. Within each cell assembly of the synfire chain was a collection of excitatory and inhibitory cells with recurrent connections (Figure 3.2). By carefully designing each assembly such that there was increased inhibition for accumulating excitatory spikes, the model was able to successfully amplify activity down through the assembly sequences, mimicking replay events, but avoids explosions of activity reminiscent of synfire chain explosions and bursting. This controlled amplification allows weak memory traces, such as those that might be generated during one-shot learning episodes, to successfully re-fire. Furthermore, the increase in inhibition due to accumulating excitatory activity causes replay events to travel in a single direction only.

Refractory periods and inhibitory effects with symmetric bidirectional connections are two methods that allow reverse replays to occur, but Haga and Fukai (2018) have shown that the effects of short-term plasticity could also be an explanation for reverse replay. By modelling short-term depression and facilitation at synapses, it is possible to long-term potentiate bidirectional connections in an asymmetric fashion, such that the reverse direction is potentiated more than the forward direction following a forward activation of a sequence. However, it is not clear how this model accounts for forward replay without first generating reverse replays, nor how it prevents continuous reversals in the memory trace strength. For instance, reverse replays cause potentiations to strengthen more in the forward direction again, thus undoing the reversed potentiation.

For completeness, it is necessary that a model can support both forward and reverse replays. Perhaps the earliest model of a network incorporating both forward and reverse replay was from Molter, Sato, and Yamaguchi (2007). Their original model was more typical in that a traversal through a set of place cells would potentiate that trajectory more in the forward than the reverse direction, but still has non-negligible reverse connections necessary for reverse replays. They also, like in Jahnke et al.'s model, employed theta phase precession during memory trace formation. The model was also somewhat simpler and computationally cheaper than the above models, as it was rate-based as apposed to spiking (though both Haga and Fukai (2018) and Gauy et al. (2018) include rate-based and spiking-based models). But the replays themselves in a 2D environment were similar to a wave-like propagation across the entire environment emanating from the position of replay initiation – as such it does not hold an accurate model of traversal for the environment, though it can provide replays of inexperienced paths.

Following memory trace formation it is then necessary to initiate replay events, and all models suggest that an external stimulus be input to the first (last) cell/cell assembly to initiate forward (reverse) replays. Chenkov, Sprekeler, and Kempter (2017), however, through control of recurrent and feedforward connection probabilities, show that asynchronous-irregular spiking can spontaneously initiate replay events – whether this is of use is unknown, but a recent study with a DynaQ neural network algorithm suggests 'random' hippocampal replays are not only useful, but necessary, for converging Q-values (Aubin, Khamassi, and Girard, 2018).

To summarize, there have been a small number of computational models, rate-based and spiking-based, that aim to capture the dynamics of hippocampal replay. Most networks require recurrent connections, either across the whole network or within sub-assemblies that are then connected as synfire-like chains, so that memory traces can be effectively stored and, as a consequence of an external stimulus, reinstated later as a replay event. The mechanisms through which each model forms memory traces and then initiates and maintains replay events is summarized in Table 3.1.



	Means for Generating Hippocampal Replay Stages			
Model	*Memory Trace Formation*	*Replay Initiation*	*Maintenance of Replay*	*Forward/Reverse Replay?*
Jahnke et al. (2015)	Spike-timing dependent plasticity with theta phase precession in recurrent network of place cells	Targeted external stimulation of place cell assemblies	Dendritic spiking with refractory periods	Forward and reverse
Gauy et al. (2018)	Hebbian plasticity between pre-existing sequence cell assemblies and place cells	Targeted external stimulation of sequence cell assemblies	Dendritic spiking with refractory periods	Forward and reverse
Chenkov et al. (2017)	Pre-existing synfire-like chains with probabilistic recurrent and feedforward connections	External stimulation or spontaneously through activity fluctuations	Recurrent excitatory and inhibitory cell assemblies for controlled amplification along assembly sequences	Forward and reverse
Haga and Fukai (2018)	Asymmetric bidirectional recurrency via STP modified Hebbian learning	Targeted external stimulation of end place cells	Asymmetric bidirectional connection strengths provide unidirectional replay	Reverse only
Molter et al. (2007)	Asymmetric bidirectional recurrency via theta phase precession Hebbian learning	Targeted external stimulation of place cells	Propagation due to strong place cell connections learned during exploration	Forward and reverse

Table 3.1 Summary of the hippocampal replay models. The means by which each model performs the stages required for hippocampal replay are summarized here. See main text for full details.

It is worth noting that a small number of studies have modelled the process of sharp-wave/ripples in the hippocampus, which occurs simultaneously with a replay event (Diba and Buzsáki, 2007). Particularly they model the generation of sharp-wave/ripples via interactions of inhibitory interneurons, extra-hippocampal inputs such as septal inputs, and/or neuromodulators like acetylcholine, and the relationship between sharp-wave/ripples and replay events (Cutsuridis and Hasselmo, 2011; Cutsuridis and Taxidis, 2013; Saravanan et al., 2015; Taxidis et al., 2012).

3.2 A Further Examination of Two Computational Models

There are two computational models of hippocampal reverse replay that together form the robotic model developed in the next chapter. These were chosen for two reasons – they are the state of the art in computational modelling of hippocampal reverse replay, and they both showed that their models performed sufficiently in a spatial navigation task. The first of these two models is by Haga and Fukai (2018), whilst the second is by Pang and Fairhall (2019). The details of each will be taken in turn, including the results of re-implementing Haga and Fukai's model. This re-implementation of Haga and Fukai's model shows that due to unbounded network weights, the network eventually becomes unstable. It is for this reason that the model of Pang and Fairhall is introduced, since their model generates reverse replays without the need for synaptic plasticities, thus eliminating the cause of the instability in Haga and Fukai's model.

3.2.1 Model of Haga and Fukai (2018)

The model of (Haga and Fukai, 2018) was developed to provide a potential biophysical mechanism for the emergence of reverse replays. Their novel contribution was in generating a modified Hebbian learning rule that is modulated by the effects of short-term plasticity. The model itself is composed of two parts. The first is an arrangement of neurons positioned in 1D space (i.e. spatial distances between each cell are given only in one dimension, see below) that allows a sequence of activity to propagate along the network. This part is used as a proof of concept for the model. The second part is an application of this model to goal-directed spatial navigation in a W-maze and an open arena, by scaling the neurons into a 2D space. Only the first of these two parts is described here, which provides the necessary details to understand the network dynamics. All mathematical details described here are those of Haga and Fukai (2018).

Mathematical Details

The model is composed of 500 neurons connected all-to-all, but with normally distributed synaptic weights. Thus, the initial weight between neuron i and neuron j is given as

$$w_{ij} = w_{max} \exp\left(-\frac{|i-j|}{d}\right) \tag{3.1}$$

Since the neurons in this model represent place cells, the term $|i-j|$ represents the magnitude of the one dimensional distance between two neuron's place fields. It is trivial to expand

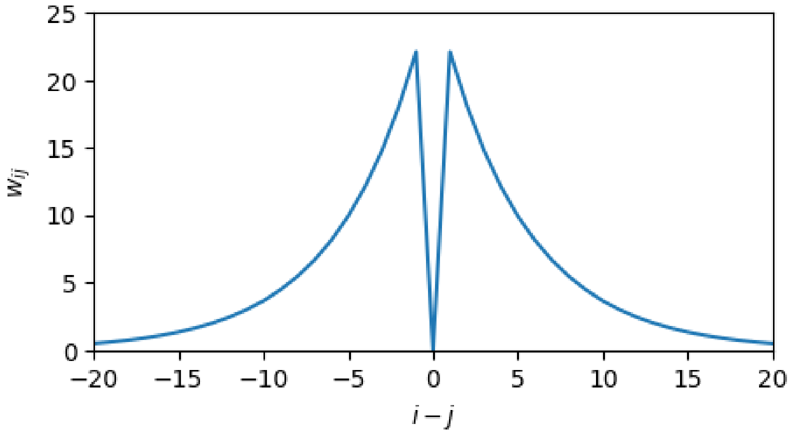


Fig. 3.3 Magnitudes of the weights when the network model of Haga and Fukai is first initialised. The chosen values of $w_{max} = 27$ and $d = 5$ results in significant weights for pairs of cells whose absolute distances are approximately < 10.

this to the two dimensional case. w_{max} and d are the max weight value and the spread of the weight distribution respectively, and are set at $w_{max} = 27$ and $d = 5$ in the model. Figure

Any weight change between a pre-synaptic neuron j and a post-synaptic neuron i occurs via a modified second order Hebbian learning rule given as

$$\tau_w \frac{d^2 w_{ij}}{dt^2} = -\frac{dw_{ij}}{dt} + \eta r_i r_j D_j F_j \tag{3.2}$$

with $\tau_w = 1000ms$ and $\eta = 20$ setting the time constant and learning rate, respectively; r specifies the rate of activity; whilst D and F are the short-term plasticity values for short-term depression and short-term facilitation, respectively.

The short-term plasticity terms are given according to the following coupled dynamical equations

$$\frac{dD_j}{dt} = \frac{1-D_j}{\tau_{STD}} - r_j D_j F_j \tag{3.3}$$

$$\frac{dF_j}{dt} = \frac{U-F_j}{\tau_{STF}} + U\left(1-F_j\right) r_j \tag{3.4}$$

with $\tau_{STD} = 500ms$, $\tau_{STF} = 200ms$, and $U = 0.6$, a parameter which sets the steady state baseline (when $r_j = 0$) for F_j. The steady state baseline for D_j is 1.

Finally, the rate of a neuron r_i is a function of its excitatory, inhibitory and external current sources

$$r_i = f\left(I_i^{exc} - I^{inh} + I_i^{ext}\right) \tag{3.5}$$

The firing rate function is a rectified linear unit of the form

$$f(I) = [\rho\,(I-\varepsilon)]^{+} = \max\{0, \rho\,(I-\varepsilon)\} \tag{3.6}$$

where ρ and ε are constants. The excitatory current and the global inhibitory current is given as follows

$$\frac{d}{dt}I_i^{exc} = -\frac{I_i^{exc}}{\tau^{exc}} + \sum_j w_{ij} r_j D_j F_j \tag{3.7}$$

$$\frac{d}{dt}I^{inh} = -\frac{I^{inh}}{\tau^{inh}} + w_{inh} \sum_j r_j D_j F_j \tag{3.8}$$

with the time constants set as $\tau^{exc} = \tau^{inh} = 10ms$, and the global inhibitory weight set to $w_{inh} = 1$. One should note here that including the same STP terms in both the excitatory and inhibitory neurons may be biologically unrealistic. This is due to short-term plasticity occurring at the synapses, but the excitatory and inhibitory synapses are different in each neuron.

Conversion to Discrete Time

In order to simulate the above set of differential equations, they are first converted into discrete time equations.[1] For this Euler's method is used (Chapra, Canale, et al., 2010). Time constants and discrete time steps were checked to ensure stability upon implementation. Given some first order differential equation of the form

$$\frac{dy}{dt} = f(t, y) \tag{3.9}$$

an estimate can be made for the value of y at time step t_{i+1} given the value and slope of y at t_i

$$y_{i+1} = y_i + f(t_i, y_i)\delta \tag{3.10}$$

where $\delta = t_{i+1} - t_i$ is some small finite step size. Thus, one can solve the first order differential of Equation 4.7 for instance in a discrete manner as follows

$$F_{t+1}^{j} = F_t^{j} + \left[\frac{\left(U - F_t^{j}\right)}{\tau_{STF}} + U\left(1 - F_t^{j}\right) r_t^{j}\right]\delta \tag{3.11}$$

[1] This method of discretisation is also followed for the models developed in Chapters 4 and 5.

For second order differential equations, one can simplify things down to solving for two coupled first order differential equations and solving each in the same way as above. Thus, for a second order differential equation of the form

$$\frac{d^2y}{dt^2} = \frac{dy}{dt} + f(t,y) \tag{3.12}$$

we let $\frac{dy}{dt} = z$ so that

$$\frac{dz}{dt} = z + f(t,y) \tag{3.13}$$

Solving these two first order differentials for discrete time as before provides us with the solution

$$z_{i+1} = z_i + [z_i + f(t_i, y_i)]\,\delta \tag{3.14}$$

$$y_{i+1} = y_i + z_i\delta \tag{3.15}$$

Haga and Fukai's model has only a single second order differential, given by Equation 3.2. Thus, to solve Equation 3.2 in discrete time, we let

$$\frac{dw_{ij}}{dt} = \Delta_{ij} \tag{3.16}$$

so that

$$\tau_w \frac{d\Delta_{ij}}{dt} = -\Delta_{ij} + \eta r_i r_j D_j F_j \tag{3.17}$$

which gives us the two first order differential equations we need to generate the discrete time version of Equation 3.2. Applying the discrete time solution to these then provides us with the following

$$\Delta_{t+1}^{ij} = \Delta_t^{ij} + \frac{1}{\tau_w}\left[-\Delta_t^{ij} + \eta r_t^i r_t^j D_t^j F_t^j\right]\delta \tag{3.18}$$

$$w_{t+1}^{ij} = w_t^{ij} + \Delta_t^{ij}\delta \tag{3.19}$$

It is important to note that these are approximations, and intuitively it can be seen that the accuracy of these approximations relies on the step size δ. To get a more rigorous understanding of the error that can arise in the discrete solutions, we can use the definition of the Taylor expansion (Chapra, Canale, et al., 2010)

$$f(t_{i+1}) = f(t_i) + \frac{df(t_i)}{dt}\delta + \frac{1}{2!}\frac{d^2f(t_i)}{dt^2}\delta^2 + \cdots + \frac{1}{n!}\frac{d^nf(t_i)}{dt^n}\delta^n \tag{3.20}$$

If we were to ignore any higher order terms, say n and beyond, then the truncation error as a result of this would be on the order of δ^n. For a first order differential, we ignore any second order terms and higher to yield the following

$$f(t_{i+1}) = f(t_i) + \frac{df(t_i)}{dt}\delta - O(\delta^2) \tag{3.21}$$

where $O(\delta^2)$ is the truncation error, on the order of δ^2, as a result of the approximation.

For a second order differential, we notice that one takes two steps in δ, and thus Equation 3.20 becomes

$$f(t_{i+2}) = f(t_i) + \frac{df(t_i)}{dt}2\delta + \frac{1}{2}\frac{d^2 f(t_i)}{dt^2}(2\delta)^2 - O(\delta^3) \tag{3.22}$$

But importantly we see in this instance that the truncation error is on the order of $O(\delta^3)$. Hence, one can get a sense as to how step size can influence the degree of error that could accumulate in the discrete approximations.

Results of Re-implementation

The simulation was run according to the system of equations above with the 500 neurons, for a total simulation time of 4 seconds in discrete time steps of 0.5ms. Given the smallest time constant in the system of equations above is 10ms, a step size of 0.5ms was deemed reasonable to avoid large approximation errors.

For the first 10ms of the simulation, an external current of 5 units was input into the first 10 neurons, with indices $i < 10$, eliciting a forward sequence of activity (Figure 3.4). At $t = 3s$, another 5 units of external current was input into the middle 10 neurons, with indices $244 < i < 255$, initiating a reversed direction sequence of activity (Figure 3.4). These results, allowing for slight differences in simulation step sizes, are qualitatively similar to the results presented in (Haga and Fukai, 2018). Furthermore, as in (Haga and Fukai, 2018), Figure 3.5 shows the effect of weight asymmetries as a consequence of the forward sequence of activity. These results, which are in accordance with the results of Haga and Fukai's, suggest the re-implementation of the neuron model is valid.

Network Instability

Haga and Fukai showed only the results following a single forward sequence followed by a single reverse sequence. It is sensible to ask, therefore, what would happen under a second (or more) forward/reverse sequence of activity. Figure 3.6 displays the activity of the network after applying the same external inputs as in Figure 3.4 immediately following the point for which the first simulation ends. As previously, the forward sequence is initiated following an external input, with the notable difference that the rates have increased and the speed with

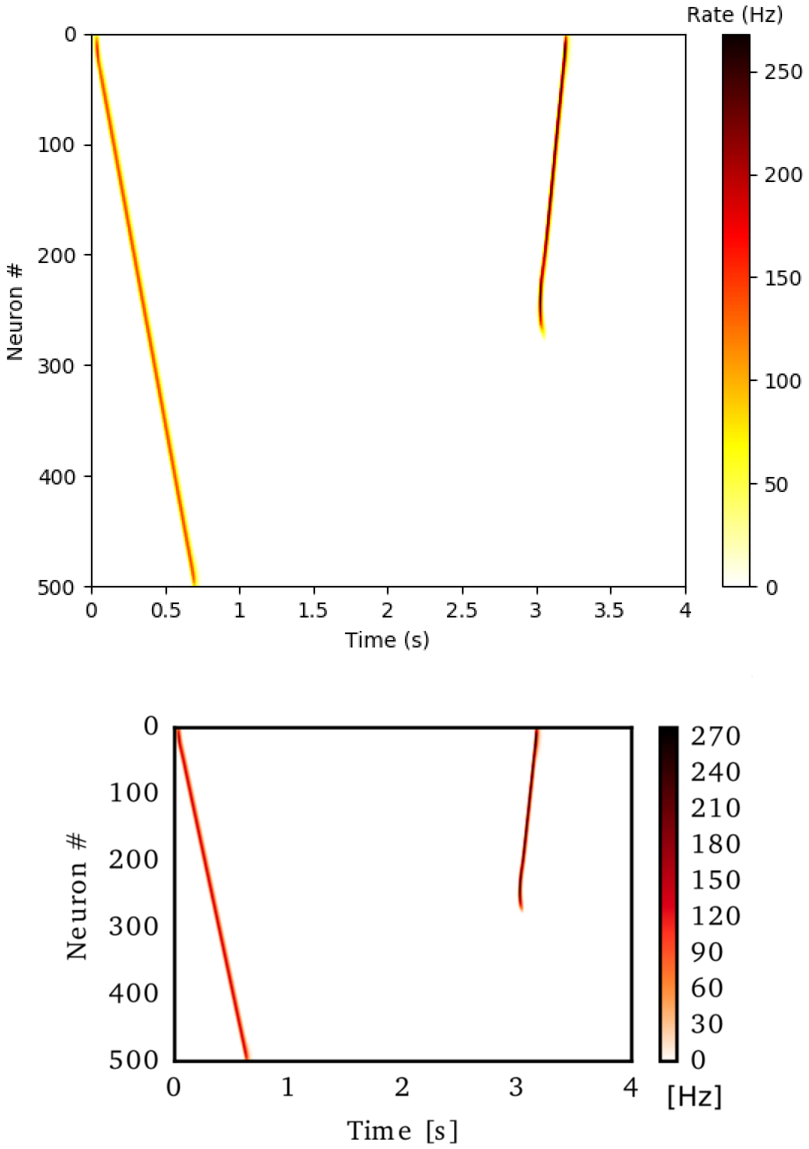


Fig. 3.4 Top: The results of simulating the linear recurrent model of Haga and Fukai (Haga and Fukai, 2018). Shown here is the activity of a network of 500 neurons. At t=0s, the first 10 neurons are stimulated initiating a sequence of activity that travels along the 500 neurons. Then at t=3s, the middle 10 neurons are stimulated. Due to the specific synaptic learning rules of the model, the synaptic weights are stronger in the reverse direction to the initial sequence of activity. This causes the activity to travel backwards – a 'reverse replay' event. Bottom: The equivalent results as published in Haga and Fukai's original paper, extracted from (Haga and Fukai, 2018).

which activity travels through the network has likewise increased. This is due to the increase in weight changes induced by the previous activity.

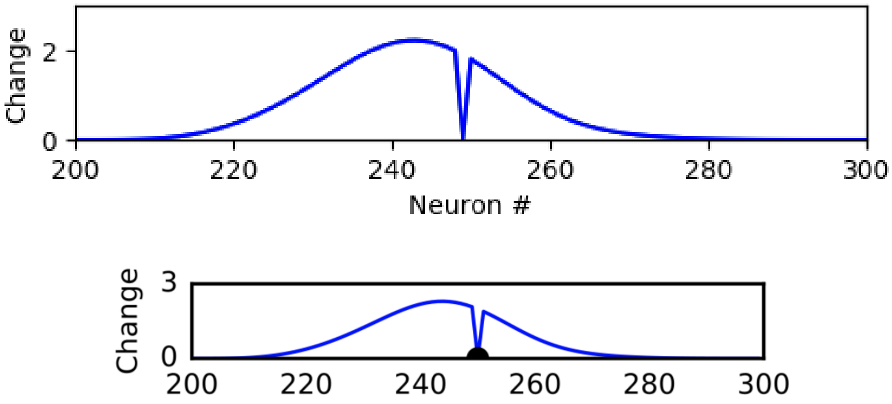


Fig. 3.5 Top: Figure showing the weight change that occurred on neuron #250, between t=0s and t=3s in Figure 3.4. I.e. the effects of weight changes after the forward sequence of activity. Notice that the weight changes between neuron #250 and neurons with indices smaller than it are larger than those between between neuron #250 and those with larger indices. This larger, asymmetric weight change in the reverse is the reason for the backward sequence of activity occurring at t=3s in Figure 3.4. Bottom: Similar results shown by Haga and Fukai in their original study, extracted from (Haga and Fukai, 2018).

Following the initiation of the reversed sequence, it becomes clear here that the network has become unstable. As before, excitation of the centre neurons initiates a sequence of activity that travels in the reverse. The activity is so strong, however, that it eventually spreads throughout the network which results in a second, uninitiated, forward sequence of activity.

Figure 3.7 displays the change of weights following the second forward sequence of activity. Whereas previously (Figure 3.5) the max weight change was at 220%, the max weight change in the second instance is 820%. It is thus reasonable to ask whether the learning rule used in this network is inherently unstable, with weight changes increasing exponentially and without bound. Recall that the synaptic learning rule for this model is given by Equation 3.2, which is a modified Hebbian learning rule. Yet, since all of $\eta, r_i, r_j, D_j, F_j \geq 0$, this learning rule contains the same instabilities of the standard Hebbian learning rule, such that weight changes are non-negative leading to uncontrolled growth (Dayan and Abbott, 2001, p. 284).

This instability would be a problem in particular if this network were to be embodied and used in a robot, since one would expect the robot to traverse the same or similar paths multiple times, and hence multiple reactivations of the same sequences. Despite this, Haga and Fukai have extended their model to a simple agent performing spatial navigation tasks, without having these same instabilities. They do this by both increasing and decreasing drastically the time constant and learning rate of Equation 3.2, respectively, and re-normalising the synaptic weights when the sums grow beyond unity. This fix is not wholly satisfactory however, due to the biological implausibility of re-normalisation. Hence, we next turn to a model by Pang and Fairhall, who show that reverse replays can initiate without the need for synaptic plasticity.

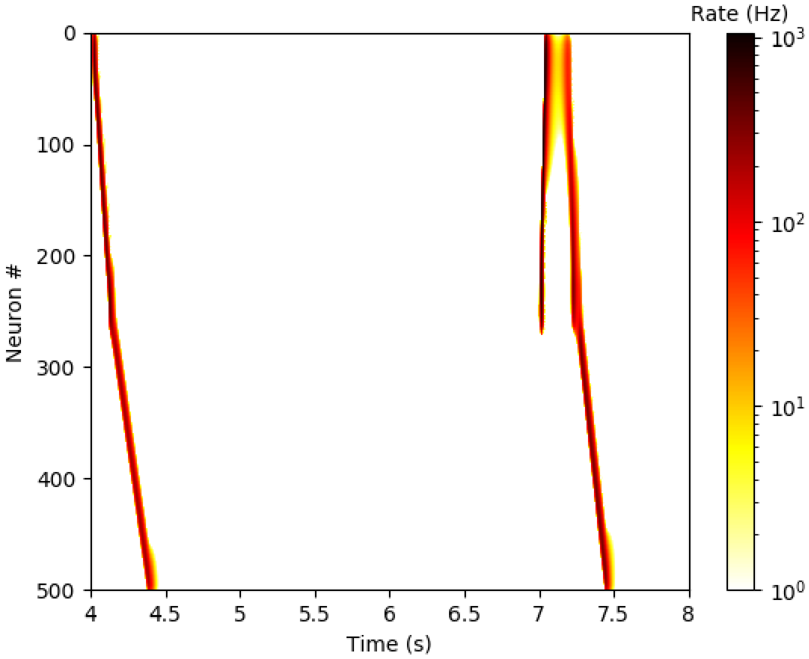


Fig. 3.6 Following immediately from the end point of Figure 3.4 at t=4s, and applying the same inputs as in Figure 3.4 at t=4s and t=7s. Due to the large weight changes induced by the previous activity, the rates are high and activity propagates much quicker. Importantly, the network becomes unstable at around t=7.1s, in which the high rates of activity causes a second and unwanted forward sequence of activity.

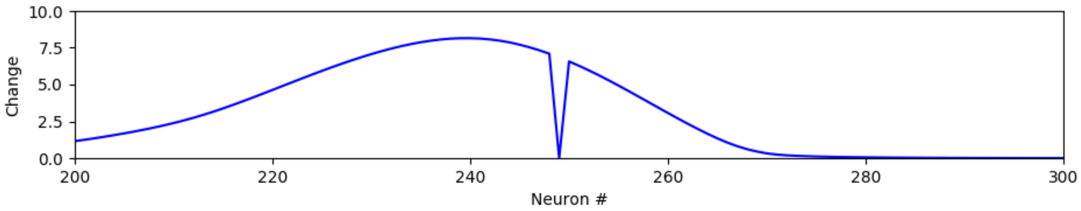


Fig. 3.7 The change in weights between neuron #250 and the rest of the neurons, this time taken between points t=0s (Figure 3.4) and t=7s (Figure 3.6). The weight changes in this instance are more than double those in Figure 3.5, indicating the possibility of exponentially increasing weight changes.

3.2.2 Extracting Intrinsic Plasticity from Pang and Fairhall (2019)

Pang and Fairhall (2019) offered another solution to the problem of how hippocampal replays might emerge, by employing the phenomenon of *intrinsic plasticity*. Intrinsic plasticity is described as the ability of a neuron, following a period of increased activity, to then increase heterosynaptic long-term potentiations in that cell (Zhang and Linden, 2003; Hyun et al., 2015), so that regardless of which pre-synaptic neuron(s) initiated the cell's activity, all synapses on that cell (or local regions of the cell, such as a dendritic branch) are long-term potentiated.

The approach taken by Pang and Fairhall, in their spiking neuron model, was to increase the potential for a neuron whose intrinsic plasticity had been increased due to it being recently active. This increased the intrinsic potentiation for that cell, making it more likely to fire in response to synaptic inputs. This ultimately had the effect of generating reverse replays following a spatial trajectory. Though they produced results for a network of leaky integrate-and-fire neurons with conductance-based synapses, they also developed a reduced spiking model. For ease of understanding then, their reduced model is described here. In the reduced model, which is a network of all-to-all connected neurons similar to Haga and Fukai's (so that weights between neurons are dependent on the spatial distance between them), they specify that for neuron i, its voltage would obey the following dynamics,

$$v_i(t) = \sum_j w_{ij} s_j(t-1) + I_i^g \sigma_i + I_i^{ext}(t) \tag{3.23}$$

where w_{ij} is the weight connecting neuron j to neuron i, $s_j(t-1)$ is 1 if neuron j spiked at time $t-1$ and 0 otherwise, and I_{ext}^i is an external current input. Intrinsic plasticity is represented as σ_i, with I_i^g being a gated current input that was a constant current injection throughout. If $v_i(t)$ exceeds some threshold voltage v_{thresh}, then neuron i spikes, enters a refractory period and then is reset. Hence, as intrinsic plasticity increases, the neuron is taken closer to spiking threshold. For computing intrinsic plasticity, they used a Sigmoidal function with lower and upper limits of 1 and 2,

$$\sigma_i = 1 + \frac{1}{1 + \exp\left[-\beta(r_i^{max} - r_\sigma)\right]} \tag{3.24}$$

where β and r_σ are constants that determine the shape of the Sigmoid, whilst r_i^{max} is the maximum firing rate that the neuron experiences over a given trial. The gating input, I_i^g, is absent during the encoding stage, but is turned on after encoding. With the gating input on, pairs of neurons that have stronger connections (and are hence close to one another in space) and who have higher levels of σ_i (and hence were active during the encoding stage) are more likely to pass on their activity, giving rise therefore to replay events.

In order to prevent replays reversing on themselves, the refractory period following a spike was important, ensuring the replay travelled in a single direction only. But Haga and Fukai's short-term plasticity approach performed a similar function. The model of hippocamapal replay developed for MiRo was deemed to be most suitable using continuous rate based, rather than spiking based, neurons. Hence, refractory periods were not a suitable option for enabling controlled replays. This led to the development of combining Haga and Fukai's short-term plasticity rule with Pang and Fairhall's intrinsic plasticity mechanism. Chapter 4 describes this model in full.

3.3 Hippocampal Replay for Robotic Applications

The models reviewed here are computational models with the primary intent of replicating experimental findings. But, as seen above, each on their own does not prove immediately useful for robotic applications. Furthermore, they all require place cell representations readily available prior to replay, and offer no useful outputs post-replay, for instance in modulating behaviour. What could be missing then is a unified model of place cell, or state, representations at the input end of a hippocampal replay model, and a means for action-selection improvement at the output end.[2]

Though a few recent studies are worth mentioning here that could integrate well with hippocampal replay. On the place cell representation end, the first is a biologically inspired SLAM algorithm, or RatSLAM, developed by Milford, Wyeth, and Prasser (2004), which has proven effective at capturing state representations in the form of 'pose cells'. With an accurate map represented in the form cell values, this offers itself as a candidate for replay models based upon neural networks. Alternatively, Byrne, Becker, and Burgess (2007) model hippocampal place cells, boundary vector cells and head direction cells, all neurophysiological features of the hippocampal region (O'Keefe and Dostrovsky, 1971b; Taube, 1998), which could provide a more biologically plausible model of place cell representations, whilst Jauffret, Cuperlier, and Gaussier (2015) have recently developed a model of grid cells (Hafting et al., 2005) and place cells that was successfully applied for spatial navigation in a robot.

[2] The replay model developed for MiRo attempts to address the latter issue, on how hippocampal replay may modulate behaviour, but does not attempt to address the former issue of place cell emergence. More is discussed regarding this in Chapter 6.

For action-selection improvement, the first is a DynaQ neural network algorithm developed by Aubin, Khamassi, and Girard (2018), which is a reinforcement learning model using Q-learning and the Dyna algorithm. It too is composed of a neural network that represents states, but pairs the states with (discrete) actions. They indeed integrated a version of hippocampal replay and showed that where Q-values could not converge online due to similarities in state values, they could however converge offline via 'random' hippocampal replays. The deep Q-network (DQN) of Mnih et al. (2015), in a similar fashion to the DynaQ neural network, utilized *experience replays*, which is conducted by selecting from a random uniform distribution groupings of state, action, reward and next state experiences. A list of experiences could then be denoted by $D_t = \{e_1, ..., e_t\}$ with $e_t = (s_t, a_t, r_t, s_t')$ being an individual experience, and applying the Q-learning update for each $e_{\text{rand}} \sim U(D)$ where $U(\cdot)$ is the uniform distribution. But both these algorithms suffer from perhaps one minor issue, in that Q-values here are learned for a discrete set of actions. Though perhaps rectifiable, this could prove problematic for states that are represented continuously.

Recent work by Mattar and Daw (2018) developed a Q-learning based reinforcement learning model that prioritizes Bellman backups. Such a prioritization (for which something similar is found in the model by Aubin, Khamassi, and Girard (2018) and termed *prioritized sweeping*) determines whether the agent should prioritize the evaluation of upcoming decisions, or whether to perform updates in order to capture newly learned information about a reward. Prioritization of the former increases the number of forward replays, whilst for the latter reverse replays become more prominent. In this way, the model favours forward replays at the start of a trial, whilst reverse replays are favoured at the end of a trial – effects similar to that found with hippocampal replay (Diba and Buzsáki, 2007).

Another challenge for robotics is the number of trials required for reinforcement learning algorithms to properly converge. This was a problem addressed by Vasilaki, Frémaux, et al. (2009) (see also (Richmond et al., 2011)) in a spike-based model of hippocampal place cells for a reinforcement learning Morris water maze task. They showed that whereas policy-gradient methods require either a high number of learning trials (due to small learning rates) or cause noisy eligibility traces (when learning rates are high), their model could account for effective learning within a small number of trials, as is found experimentally with rats. Interestingly, they modelled "action cells", which could possibly be found in the basal ganglia as an action selection mechanism (Redgrave, Prescott, and Gurney, 1999), and further, unlike the models discussed above, they were able to represent actions and states as continuous, rather than discrete. Yet importantly for our discussion here, they did not employ hippocampal replay. The model developed for MiRo will advance the model of Vasilaki, Frémaux, et al. (2009) by augmenting it with hippocampal replay (see Chapter 5).

Hippocampal replay could offer another means to achieve the low number of learning trials required – learning is done “offline" as (perhaps noisy) repetitions of previous experience. This could therefore offer an effective and highly efficient mechanism that converges state-action values “offline", which could prove useful for robotic learning, as well as offer bioinspired learning mechanisms for biomimetic robotics.

Chapter 4

Developing a Model of Hippocampal Replay in the MiRo Robot

Presented in this chapter is a hippocampal CA3 inspired, continuous rate-based network model of reverse replay implemented on a simulated version of the biomimetic robot MiRo (Mitchinson and Prescott, 2016). The work here is largely based on a recently accepted submission to the International Conference on Biomimetic and Biohybrid Systems (Online, July 29 - 30, 2020). See (Whelan, Prescott, and Vasilaki, in press).

Reverse replays in this model occur as a consequence of two modes of transient neural states. The first is due to the implementation of a time decaying model of *intrinsic plasticity*. Intrinsic plasticity is the ability of a cell to increase heterosynaptic long-term potentiation of post-synaptic potentials following recent activity (Zhang and Linden, 2003; Hyun et al., 2015), and has recently been proposed as a potential mechanism for the occurrence of reverse replays by Pang and Fairhall (2019). The second transient neural state implementation is in short-term plasticity, which acts to ensure unidirectional, stable replays, developed by Haga and Fukai (2018). This is due to short-term depression suppressing synaptic currents after a given amount of continuous firing, thus preventing unbounded synaptic transmissions.

As described in the previous chapter, these components from each of the two models are selected to overcome the shortcomings in the other. For instance, Haga and Fukai's model employs short-term plasticity in order that replays are unidirectional, but requires long-term plasticity for learning memory traces which can lead to instabilities. Pang and Fairhall's model meanwhile employs intrinsic plasticity rather than synaptic plasticity for learning memory traces, thus overcoming the network instabilities of Haga and Fukai, but requires refractory periods in their spiking neurons to ensure unidirectional replays. Refractory periods are difficult to employ in a network of continuous rate neurons, thus short-term plasticity provides a suitable substitute for solving this issue.

The details of the model are presented first, which includes a description of the network architecture and its dynamics. Following this, a description of the experimental setup used to test the model on the MiRo is given, followed by a presentation of the results from the experiment. Finally, a detailed discussion of the model and the experimental results is provided.

4.1 Model Details

4.1.1 Network Architecture

The network consists of 100 rate-based neurons representing place cells, arranged in a grid of size 10×10, each of which has its place fields spread evenly across an open circular environment. Each cell forms a bidirectional and symmetric synaptic connection to its 8 nearest neighbours, with all weights fixed at a value of 1. Figure 4.1 gives an example of the network architecture for a subset of cells.

4.1.2 Network Dynamics

The rate for each place cell neuron, represented by x_j, is given as a linearly rectified rate with upper and lower bounds of 0Hz and 100Hz,

$$x_j^{'} = \alpha \left(I_j - \varepsilon\right)$$

$$x_j = \begin{cases} 0 & \text{if } x_j^{'} < 0 \\ 100 & \text{if } x_j^{'} > 100 \\ x_j^{'} & \text{otherwise} \end{cases} \tag{4.1}$$

where α and ε are constants. I_j is the cell's activity, which evolves according to time decaying first order dynamics,

$$\tau_I \frac{d}{dt} I_j = -I_j + \psi_j I_j^{syn} + I_j^{place} - I^{inh} \tag{4.2}$$

where τ_I is the time constant, I_j^{syn} is the synaptic inputs from the cell's neighbouring neurons, and I_j^{place} is the place specific input calculated as per a normal distribution of MiRo's position from the place field's centre point. ψ_j represents the place cell's *intrinsic plasticity*, as discussed above and detailed further below. I^{inh} is a global inhibitory input.

Each place cell has associated with it a place field in the environment defined by its centre point and width, with place fields distributed evenly across the environment (100 in

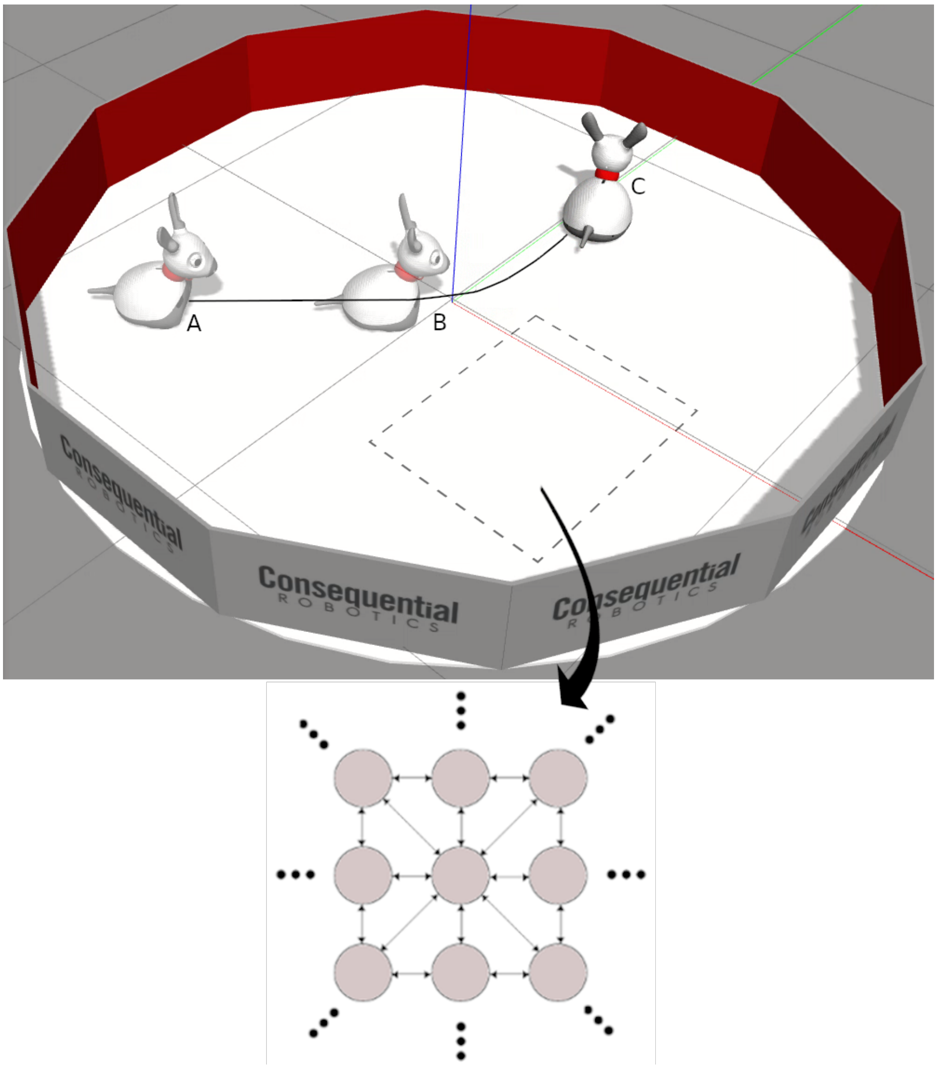


Fig. 4.1 The simulated environment used to test the model with the MiRo robot (see Section 4.2.1). The network architecture consists of a 10x10 array of place cells with place fields uniformly covering the environment. Bidirectional symmetric connections exist between each cell's eight nearest neighbours in space, as shown for a small patch of the environment here. An example trajectory is shown here, in which MiRo begins at the start position in A, passes through location B and ends in the goal location at C.

total). As stated, the place specific input, I_j^{place}, is computed from a two-dimensional normal distribution determined by MiRo's distance from the place field's centre point,

$$I_j^{place} = I_{max}^{p} \exp\left[-\frac{(X_{MiRo}-X_j)^2+(Y_{MiRo}-Y_j)^2}{2d^2}\right] \tag{4.3}$$

where I_{max}^{p} determines the max value for the place cell input. (X_{MiRo}, Y_{MiRo}) represents MiRo's (x, y) coordinate position in the environment, whilst (X_j, Y_j) is the location of the place field's centre point. The term d in the denominator is a constant that determines the width of the place field's distribution. In this case, d is chosen to be 0.1m, which ensures overlapping of place fields with nearest neighbours only.

The synaptic inputs, I_j^{syn}, are computed as a sum over neighbouring synaptic inputs modulated by the effects of short-term depression and facilitation, D_k and F_k respectively,

$$I_j^{syn} = \lambda \sum_{k=1}^{8} w_{jk}^{place} x_k D_k F_k \tag{4.4}$$

where w_{jk}^{place} is the weight projecting from place cell k onto place cell j. In this model, all these weights are fixed at a value of 1. λ takes on a value of 0 or 1 dependent on whether MiRo is exploring ($\lambda = 0$) or is at the reward ($\lambda = 1$). This therefore prevents there being any synaptic transmissions during exploration, but not whilst MiRo is at the reward (the point in which reverse replays occur). Whilst not biologically realistic, this two-stage approach can be found in similar models as a means to separate an *encoding* stage during exploration from a *retrieval* stage (Saravanan et al., 2015), and was a key feature of some of the early associative memory models (Hopfield, 1982). Experimental evidence also supports this two-stage process due to the effects of acetylcholine. Acetylcholine levels have been shown to be high during exploration but drop during rest (Kametani and Kawamura, 1990), whilst acetlycholine itself has the effect of suppressing the recurrent synaptic transmissions in the hippocampal CA3 region (Hasselmo, Schnell, and Barkai, 1995).

The inhibitory input, I_i^{inh}, is a global term given as a summation of the whole network's activity,

$$\frac{d}{dt} I^{inh} = -\frac{I^{inh}}{\tau^{inh}} + w_{inh} \sum_j r_j D_j F_j \tag{4.5}$$

D_k and F_k in Equation 4.4 are respectively the short-term depression and short-term facilitation terms, and for each place cell these are computed as (as in (Haga and Fukai, 2018), but see (Tsodyks, Pawelzik, and Markram, 1998; Vasilaki and Giugliano, 2014; Esposito, Giugliano, and Vasilaki, 2015)),

$$\frac{d}{dt} D_k = \frac{1 - D_k}{\tau_{STD}} - x_k D_k F_k \tag{4.6}$$

$$\frac{d}{dt} F_k = \frac{U - F_k}{\tau_{STF}} + U\,(1 - F_k)\, x_k \tag{4.7}$$

where τ_{STD} and τ_{STF} are the time constants, and U is a constant representing the steady-state value for short-term facilitation when there is no neuron activity $(x_k = 0)$. D_k and F_k each take on values in the range $[0,1]$. Notice that when $x_k > 0$, short-term depression is driven steadily towards 0, whereas short-term facilitation is driven steadily upwards towards 1.

Turning finally to the intrinsic plasticity term in Equation 4.2, represented by ψ_j. Its behaviour, as observed in Equation 4.2, is to heterosynaptically scale all incoming synaptic inputs. To model intrinsic plasticity in (Pang and Fairhall, 2019), a heuristically developed sigmoid whose output was determined as a function of the neuron's rate was used, but it did not have time decaying dynamics. Given our robot often travels across most of the environment, we needed a time decaying form of intrinsic plasticity to avoid potentiating all cells in the network. The simplest form of time decaying intrinsic plasticity is therefore,

$$\frac{d}{dt}\psi_j = \frac{\psi_{ss} - \psi_j}{\tau_\psi} + \frac{\psi_{max} - 1}{1 + \exp\left[-\beta(x_j - x_\psi)\right]} \tag{4.8}$$

with again, τ_ψ being its time constant, and ψ_{ss} being a constant that determines the steady state value for when the sigmoidal term on the right is 0. All of ψ_{max}, β and x_ψ are constants that determine the shape of the sigmoid. Since ψ_j could potentially grow beyond the value of ψ_{max}, we restrict ψ_j so that if $\psi_j > \psi_{max}$, then ψ_j is set to ψ_{max}.

In order to initiate a replay event then, place cell inputs as computed using Equation (4.3) need inputting for only a short time period. Due to the effects of the intrinsic plasticity and increased synaptic connectivity within the recurrent hippocampal network, this initiates a fast replay of the most recent place cell trajectory, as shall be shown next.

4.1.3 Model Parameters

The model parameters used across the experiments are given in Table 4.1 below.

Parameter	Value
α	$1C^{-1}$
ε	$2A$
τ_I	$0.05s$
I^p_{max}	$50A$
d	$0.1m$
λ	0 or 1, see text
τ^{inh}	$0.05s$
$_{inh}$	0.1
τ_{STD}	$1.5s$
τ_{STF}	$1s$
U	0.6
ψ_{ss}	0.1
ψ_{max}	4
τ_ψ	$10s$
β	1
x_ψ	$10Hz$

Table 4.1 Model parameter values used in the experiments for hippocampal reverse replays.

4.2 Experimental Setup and Results

4.2.1 MiRo Robot and the Testing Environment

For testing the model, it is implemented using a simulated version of the MiRo robot (Figure 4.2). The MiRo robot, a commercially available robot developed by Consequential Robotics in partnership with the University of Sheffield, is a biomimetic robot whose design has been inspired by biology, psychology and neuroscience. For mobility it is differentially driven, whilst for sensing only its front facing sonar is used for the detection of approaching walls and objects, though there are a number of additional available sensing options. The Gazebo physics engine is used to perform simulations, where the readily available open-arena environment is used (Figure 4.1). It is run using the Kinetic Kame distribution of the Robot Operating System (ROS). Full specifications for the MiRo robot, including instructions for simulator setup, can be found on the MiRo documentation web page (Consequential Robotics, 2019).

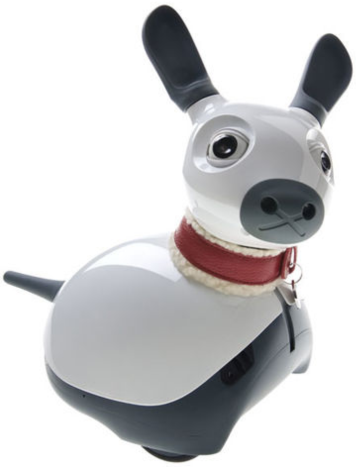


Fig. 4.2 The biomimetic MiRo robot is differentially driven and has a number of sensing options, though for this experiment only the sonar sensor, located in MiRo's nose, is used for detecting walls. Whilst it is available both in physical form and simulated form, for ease of experimental setup, the simulated version of MiRo is used for testing our network model.

4.2.2 Searching for a Hidden Reward

The model is run on the MiRo robot in a simulated open arena environment, having a diameter of 2m (Figure 4.1) and using simulation time steps of 10ms. Model equations are discretised using the Euler method with time steps of $\Delta t = 10ms$ to match the simulated time steps.[1] From a random start location, MiRo is left to freely explore its environment via a basic implementation of a random walk, with the goal of finding a hidden reward. It is worth noting that in biology, rodents often take better characterised paths when performing search strategies (Vouros et al., 2018). For instance, they can often concentrate their searches in target areas, scan across an environment efficiently, or stick to walls (Vouros et al., 2018). For simplicity however, these have not been implemented in the robot. This random walk is the *active exploration* phase, and during this phase the network rates are driven solely by the place specific inputs with no recurrent synaptic transmissions. There is no synaptic plasticity implemented in this experiment, and so all weights, w_{ij}, are fixed at a value of 1. Figures 4.3A and 4.3B show the activity of the network during active exploration. Due to the distribution of the place-specific input, no more than 4 cells are active at any one time, though most often this amounts to no more than 2 or 3 cells being simultaneously active. This sparse representation during exploration provides a neural representation of space. Neurons that become active due to the place specific input then undergo increases in intrinsic plasticity, decaying exponentially (according to Equation 4.8) when activity in the neuron drops.

[1] Full code for the model (using Python 2.7) can be found at https://github.com/mattdoubleu/robotic_reverse_replay.

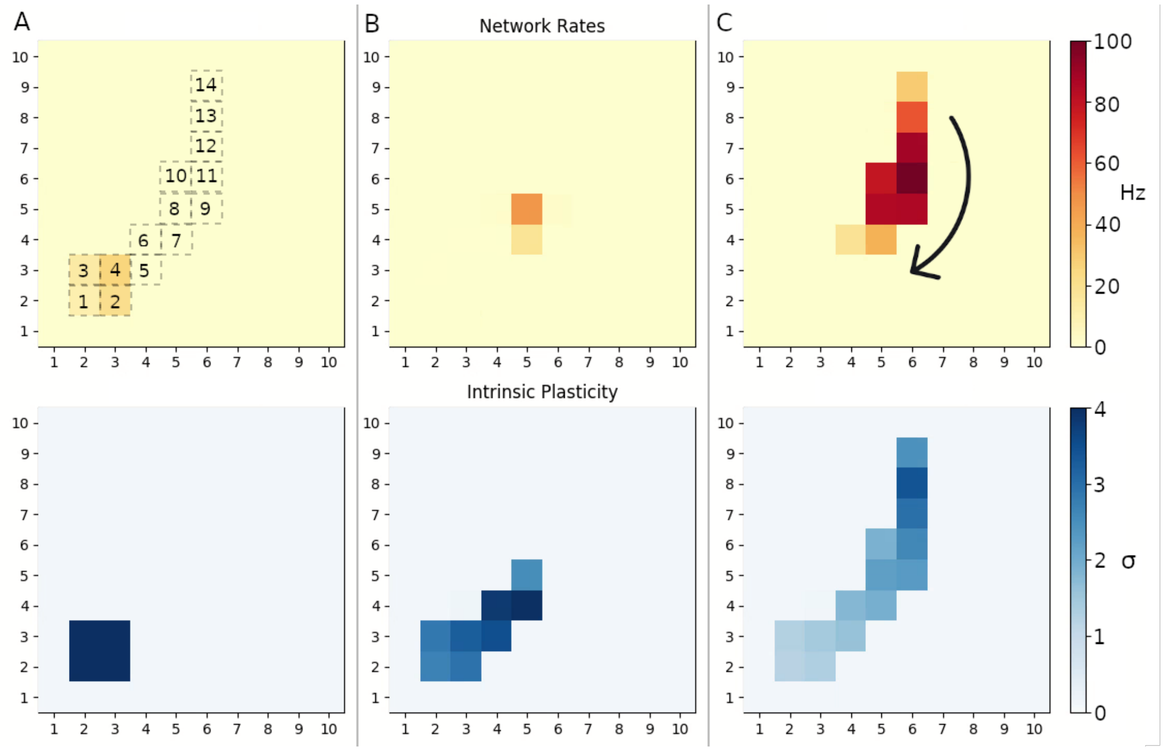


Fig. 4.3 Rates (top plots) and intrinsic plasticities (bottom plots) for the 10x10 network are shown here for the locations marked in the trajectory of Figure 4.1. These are: A) MiRo is at the start location. The numbered boxes ranging from 1 to 14 here represent all cells that were active during the exploration phase and the temporal order in which they fired during exploration (i.e. cell 1 fired first, cell 14 last). Note however that at the start point, only the first 4 cells were active. B) MiRo is exploring the environment. C) MiRo has reached the reward and reverse replays are being initiated. The arrow indicates the temporal order of firing during this replay event.

Upon reaching the hidden reward location, MiRo pauses and enters the *quiescent reward* phase. Place specific inputs are computed using Equation 4.3 and are input into the network via pulses of 0.1s-ON and 1.9s-OFF. Recall that during this phase, recurrent synaptic conductances are allowed. Due to the increase in synaptic recurrent conductance and post-synaptic activity being scaled by the intrinsic plasticity, activity propagates quickly through the network, reinstating the most recently active cells in a temporally reversed order to that seen during exploration. Figure 4.3C shows the activity of the network midway through a replay event. Notice the trace in the intrinsic plasticity plots, which transiently stores the most recent sequence of activity in the network and provides the mechanism for faithful replays of the recent trajectory. In this instance, many more cells are found to be simultaneously active, but their time points for peak activity retain the temporal ordering seen during exploration (Figure 4.4).

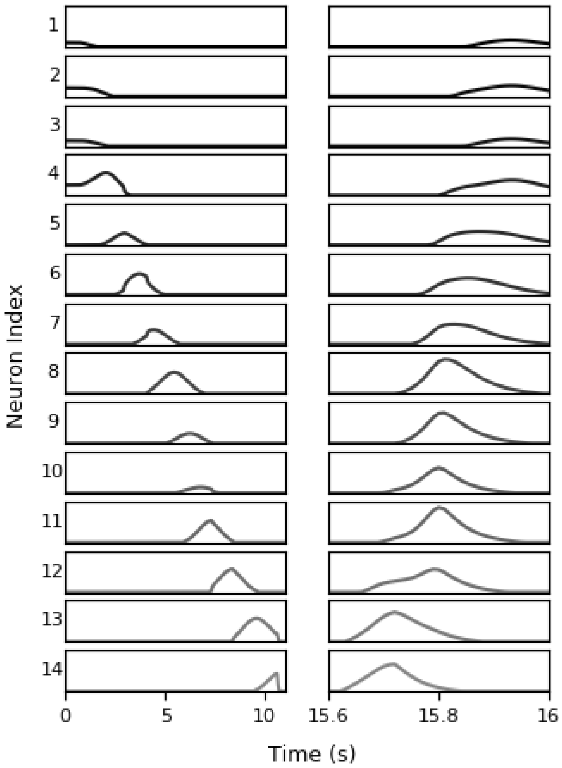


Fig. 4.4 A time course plot of the cell rates for the cells indexed in Figure 4.3A. The lower and upper limits in each box plot is 0Hz and 100Hz. Plots on the left show the activities during exploration, occurring over a time period of approximately 12s. The plots on the right show the activity during a reverse replay event. Note that Figures 4.3A, B and C are snapshots of the network's activity at times 0s, 5s and 15.8s, respectively.

In order to provide a more detailed comparison of the network's activity during the exploration phase versus the quiescent phase in which reverse replays occur, Figure 4.4 displays a time course plot of the rates for the 14 cells that were active during exploration in Figure 4.3A. It is clear in Figure 4.4 that the temporal ordering of cell firing during a reverse replay event is preserved in comparison to the ordering during exploration.

4.2.3 Removing intrinsic plasticity

To show the effects of removing intrinsic plasticity from the model, σ_i is set to 1 for all cells and the model is run once more on a similar trajectory (Figure 4.5). In this instance, rather than a direct replay of the recent trajectory, the activity in the network displays a divergent replay event across the whole network from the point of initiation. This effect was similarly seen in the model of Haga and Fukai (2018), who assumed a similar network architecture to this one but did not model intrinsic plasticity. This shows that the intrinsic plasticity

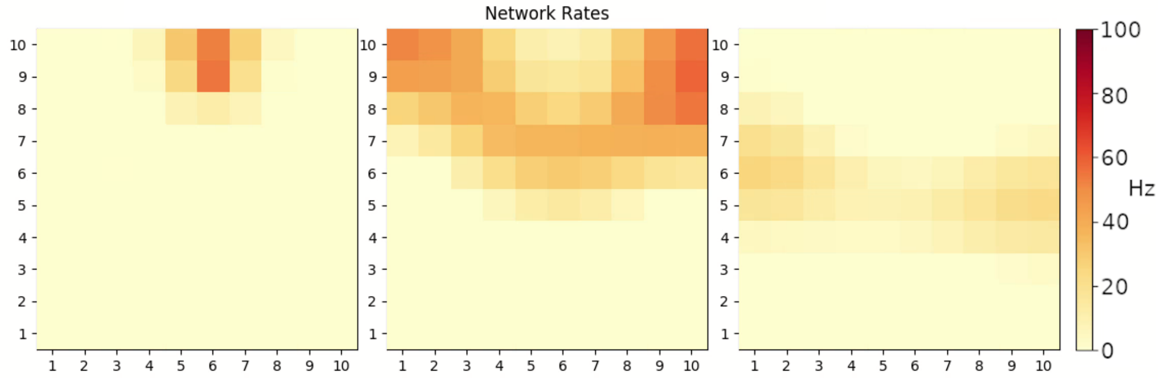


Fig. 4.5 Example of a replay event without intrinsic plasticity, where $\sigma_i = 1$ for all neurons. A similar trajectory as in Figures 4.1/4.3 is taken here, with reverse replay events initiated at the same location. The heat maps, from left to right, show the temporal ordering of network activity during a replay event. As intrinsic plasticity is homogeneous across the network, there is no preferential trajectory for the sequence of cell activities to follow. As such a divergent wave propagates across the whole network from the point of initiation.

is important for restricting the replay event to the previously experienced trajectory only. However, divergent replays could have potential benefits in the learning of goal-oriented paths (see Section 4.3 below).

4.3 Analysing the Model and Results

The model of a hippocampal CA3 inspired network presented here produces fast reverse replays of recently active place cell trajectories. Whilst the network connectivity remains static and symmetric, the implementation of intrinsic plasticity produces asymmetries in the network that amplifies incoming synaptic currents, enabling activity to travel through the network along a trajectory determined by levels of intrinsic plasticity. Intrinsic plasticity was first introduced as a potential mechanism for hippocampal replays by Pang and Fairhall (2019), but as we are running the model on the MiRo robot, for which it can very quickly cover a whole area, time decaying dynamics have had to be included so that the whole network does not become intrinsically potentiated. Given only a subset of the network becomes potentiated by intrinsic plasticity (i.e. those cells most recently active), this creates a certain level of sparsity in the network, and is interesting to compare with a previous computational model of replay dynamics by Chenkov, Sprekeler, and Kempter (2017) who showed that sparsity in their network was important for generating effective and controlled replays. Yet, whilst they achieved sparsity by changing the number of synaptic connections, here it is achieved through intrinsic plasticity changes. These results nonetheless point towards a level of sparsity that is important for specific and controlled replays.

Another important component in this model for generating stable propagations of replay sequences is short-term plasticity effects, first shown by Haga and Fukai (2018) in their reverse replay model. It is perhaps a useful analogy to consider short term plasticity in this instance having the effect of a 'refractory period' for activity propagation, in that it prevents further transmission of activity after a given amount of continuous activity. Refractory periods have been shown in previous models to ensure stable, unidirectional replays (Jahnke, Timme, and Memmesheimer, 2015; Pang and Fairhall, 2019). However, implementing refractory periods requires a model of spiking neurons, and so modelling short-term plasticity lends itself to rate-based implementations of replay. This is of course particularly useful in real-time robotic applications where spiking neuron models may be computationally inefficient. But short-term plasticity could have a more interesting property during reverse replays. Haga and Fukai (2018) showed that short-term plasticity could generate reversed synaptic weight changes. This enables reverse replays to strengthen synaptic traces in the forwards direction, despite the replay event occurring in the reverse. Thus, whilst their model produced divergent replay events similar to that seen here when intrinsic plasticity is removed, the reversed synaptic potentiations proved useful in generating synaptic traces towards a goal location, even if particular place cells had not been active during exploration. These could prove useful if, for instance, the network connectivity provides a neural map of the environment. Replays could then provide a means to explore trajectories towards goal locations even for trajectories that have never been physically explored.

A third component of the model that was necessary for appropriately timed replays was the implementation of a two-stage dynamic, which prevented the network from transmitting recurrent synaptic currents during the *exploration phase*, but allowed synaptic transmission during the *quiescent reward* phase (where MiRo sat quietly at the reward location). This was based on findings that suggest different levels of acetylcholine during active exploration and sleep states (Kametani and Kawamura, 1990), which alters CA3 synaptic conductances (Hasselmo, Schnell, and Barkai, 1995) – higher levels of acetylcholine inhibit synaptic conductance. However, what is not clear is that acetylcholine levels drop significantly enough during the quiescent reward state for which reverse replays occur, given it follows immediately after exploration (Foster and Wilson, 2006). Whilst levels of acetylcholine have been found to change quickly on the time scale of a few seconds, at least in the prefrontal cortex (Parikh et al., 2007), it is unclear as to whether this occurs in the hippocampal CA3 region. What is perhaps interesting to note, however, is that cholinergic stimulation, which leads to an increase in acetylcholine, has been shown to suppress hippocampal sharp-wave ripples yet promote theta oscillations (Vandecasteele et al., 2014). Given theta activity is found to co-occur with exploratory states (Vanderwolf, 1969), whilst replays occur usually during sharp-wave ripple events (Diba and Buzsáki, 2007), this suggests that for reverse replays to arise, acetylcholine levels must phasically drop during a quiescent reward state to enable sharp-wave ripples.

Chapter 5

Employing Reverse Replays in a Robotic Reinforcement Learning Task

Presented here is a unification of the hippocampal reverse replay model presented in Chapter 4 with a hippocampal-striatal inspired reinforcement learning model. The reinforcement learning (RL) model is based on the spiking neuron model of Vasilaki, Frémaux, et al. (2009). However, the synaptic learning rule derived in Vasilaki, Frémaux, et al. (2009) has here been re-derived for continuous rate neurons, such that a novel learning rule has been developed. The newly developed learning rule offers interesting properties that is later compared with the action selection hypothesis of the basal ganglia.

This chapter is broken down as follows: First, background material relating to RL is re-introduced but also extends upon the discussion in the Introduction (Chapter 1). This is followed by a short review of three-factor learning rules and synaptic eligibility traces, mechanisms for which the original RL model of Vasilaki, Frémaux, et al. (2009) is based upon. It is then contrasted with the hypothesis of hippocampal replay's involvement in RL. Following this background material is a description of the complete details of the model. This description includes a model of action cells that intend to represent striatal cells, but does not include the hippocampal replay model where the reader is referred to Chapter 4 for those details. The developed model is tested again with MiRo and in the same environment as in that of Chapter 4, but this time MiRo is tasked with learning the location of a hidden goal. Performance is compared both with and without reverse replays, in order to show the effect reverse replays have on performance. The chapter concludes with a discussion on the model.

It is perhaps worth noting that the content of this chapter is currently in preparation for a journal paper submission.

5.1 Background

Though the neurobiology of RL has largely centred on the role of dopamine as a reward-prediction error signal (Schultz, 1998; Redgrave, Vautrelle, et al., 2017), there are still questions surrounding how brain regions might coordinate with dopamine release for effective learning. Particularly given the fact that dopamine is released at the point of reward administration, which typically follows the behaviour that led to the reward. Another way to state this problem is as follows: Behavioural timescales evolve over seconds, perhaps longer, whilst the timescales for synaptic plasticities in mechanisms such as spike-timing dependent plasticity (STDP) evolve over milliseconds (Bi and Poo, 1998) – how does the nervous system bridge these time differentials so that rewarded behaviour is reinforced at the level of synaptic plasticities?

Three-Factor Learning Rules and Synaptic Eligibility Traces

One recent hypothesis as an explanation to this problem has been in three-factor learning rules (Vasilaki, Frémaux, et al., 2009; Richmond et al., 2011; Frémaux and Gerstner, 2016; Gerstner et al., 2018). In the three-factor learning rule hypothesis, learning at synapses occurs only in the presence of a third factor, with the first and second factors being the typical pre- and post-synaptic activities. This can be stated in a general form as follows,

$$\frac{d}{dt}w_{ij} = \eta f(x_j)g(y_i)M^{3rd}(t) \tag{5.1}$$

where η is the learning rate, x_j represents a pre-synaptic neuron with index j, y_i a post-synaptic neuron with index i, and $f(\cdot)$ and $g(\cdot)$ being functions mapping respectively the pre- and post-synaptic neuron activities. $M^{3rd}(t)$ represents the third factor, which here is not specific to the neuron indices i and j and is therefore a global term. This third factor is speculated to represent a neuromodulatory signal, which in this case is best thought of as a dopamine, or more generally a reward, signal. Equation 5.1 in its current form still appears to possess the problem stated above, of how learning can occur for neurons that were co-active prior to the introduction of the third factor. To solve this, a synaptic eligibility trace is introduced, which is a time-decaying form of the pre- and post-synaptic activities (Gerstner et al., 2018),

$$\begin{aligned} \frac{d}{dt}e_{ij} &= -\frac{e_{ij}}{\tau_e} + \eta f(x_j)g(y_i) \\ \frac{d}{dt}w_{ij} &= e_{ij}M^{3rd}(t) \end{aligned} \tag{5.2}$$

The eligibility trace time constant, τ_e, modulates how far back in time two neurons were co-active for in order for learning to occur – the larger τ_e is, the more of the behavioural time history will be learned and therefore reinforced. To effectively learn behavioural sequences over the time course of seconds then, τ_e is set to be in the range of a few seconds (Gerstner et al., 2018). Work conducted by Vasilaki, Frémaux, et al. (2009) successfully applied such a learning mechanism in a spiking network model for a simulated agent learning to navigate in a Morris water maze task, in which they used a value of 5s for τ_e.

Hippocampal Replay as an Alternative?

But there might be an alternative hypothesis within the three-factor learning rule framework to synapse-specific eligibility traces for learning on the order of behavioural timescales, via the phenomenon of *hippocampal replay*. It was the original experiment by Foster and Wilson (2006) that began the speculations that hippocampal reverse replays might be significantly involved in RL. Since reverse replays occur immediately after reaching a reward, and replays the most immediate experience (Diba and Buzsáki, 2007), it has been speculated that reverse replays, coupled with phasic dopamine release, might be such a mechanism to reinforce behavioural trajectories and thus solving the problem stated above.

Whilst it has been well established that hippocampal neurons project to the nucleus accumbens (Humphries and Prescott, 2010), the proposal that reverse replays may play an important role in RL has since received further support. For instance, there are experimental results showing that reverse replays often co-occur with replays of the ventral striatum (Pennartz et al., 2004) as well as there being increased activity in the ventral tegmental area during awake replays (Gomperts, Kloosterman, and Wilson, 2015), which is an important region for dopamine release. Furthermore, rewards have been shown to modulate the frequency with which reverse replays occur, such that increased rewards promotes more reverse replays, whilst decreased rewards suppresses reverse replays (Ambrose, Pfeiffer, and Foster, 2016).

In terms of theoretical support, the role of hippocampal replays in relation to RL methods, and particularly the role of replays in speeding up the learning process, has recently been examined (Johnson and Redish, 2005; Mattar and Daw, 2018; Cazé et al., 2018). Indeed, the popular RL algorithm Deep-Q Network (Mnih et al., 2015) utilises the concept of "experience replays", such that experiences, characterised by state transitions, are stored throughout an episode and then those experiences are replayed uniformly at random during Q-learning updates. Interestingly there are instances, such as in a navigational setting similar to the one we test here, where non-random experience replays that are determined by the magnitudes of the temporal difference errors performs better than uniform random sampling (Karimpanal and Bouffanais, 2018). There is therefore growing experimental and theoretical support for hippocampal replays playing an active and even central role in RL. This then offers a complimentary, or potentially an alternative, solution to the synaptic eligibility trace in solving the problem of learning on behavioural timescales.

5.2 A Hippocampal-Striatal Model

5.2.1 Network Architecture

The network is composed of a layer of 100 bidirectionally connected *place cells*, which connects feedforwardly to a layer of 72 *action cells* via a weight matrix of size 100×72 (Figure 5.1B). The place cells each encode for a specific region of the environment (O'Keefe and Dostrovsky, 1971a; O'Keefe, 1976), and in this model place cell activities are generated heuristically using two dimensional normal distributions of activity inputs determined as a function of MiRo's position from each place field's centre point (Figure 5.1A), similar to other approaches of place cell activity generation (Vasilaki, Frémaux, et al., 2009; Haga and Fukai, 2018). The action cells are driven by the place cells, with each action cell encoding for a particular heading that MiRo moves towards – 72 action cells encoding for 360 degrees means each action cell encodes for 5 degree increments (for simplicity, MiRo's forward velocity is kept constant at 0.2m/s).

5.2.2 Hippocampal Place Cells

The network model of place cells is as detailed in Chapter 4, but with one minor modification. Here, the place cells have from their dynamics the inhibitory inputs of Equation 4.2 removed. It was found in further analysis that, due to the λ term in Equation 4.4, inhibitory inputs are not necessary for ensuring stable activity clusters in the network. Appendix A.1 presents an analysis showing that there is no difference in the performance of the network either during exploration or reverse replays, with or without inhibition. The updated place cell dynamics for the network presented in this chapter is therefore,

$$\tau_I \frac{d}{dt} I_j = -I_j + \psi_j I_j^{syn} + I_j^{place} \tag{5.3}$$

All other dynamics and parameter values are kept the same as in the experiments in Chapter 4. Finally, it is worth reiterating that each place cell encodes for a region in the environment, and that the rates for each place cell is represented by the notation x_j after passing Equation 5.3 through a linear rectifier.

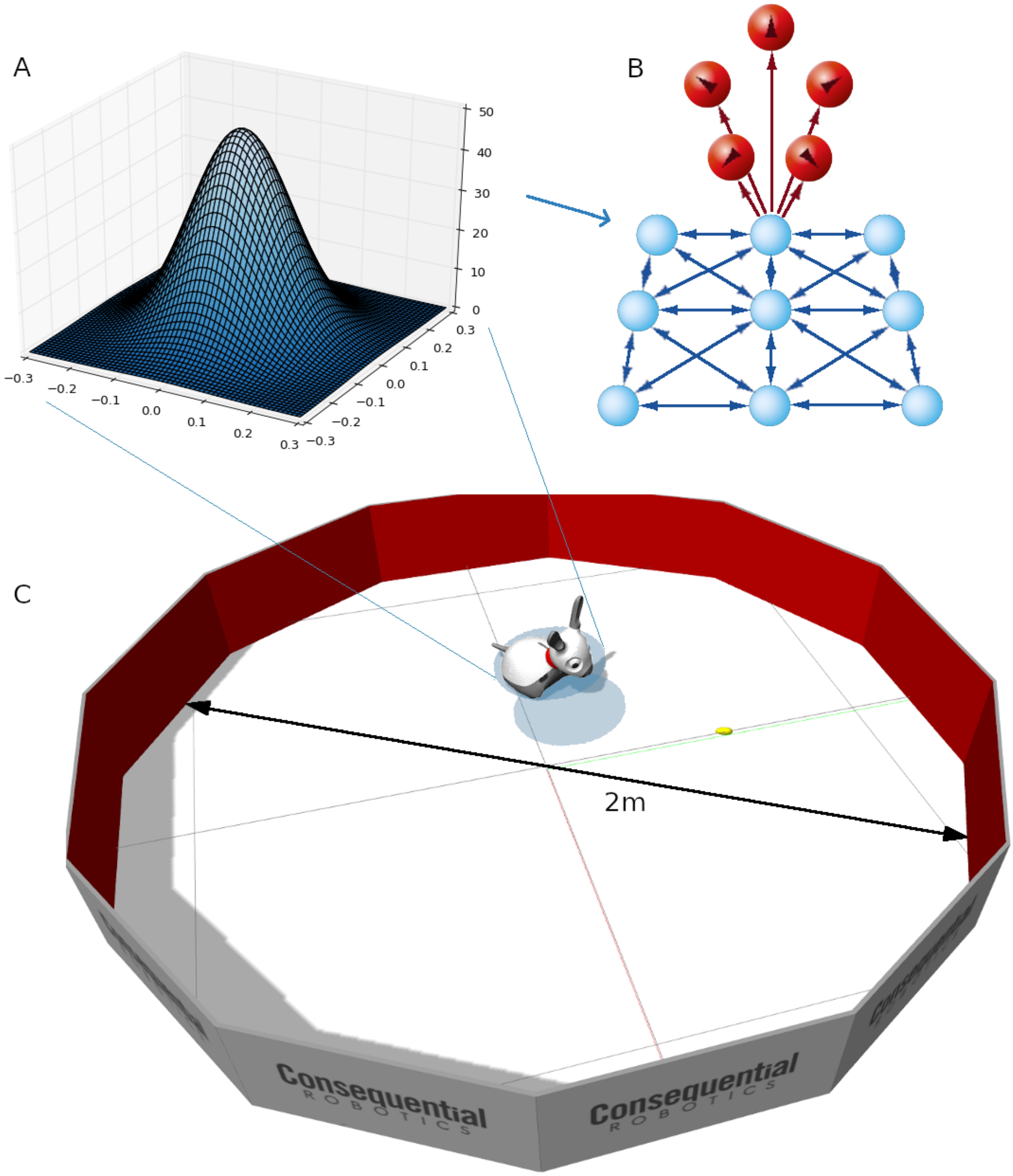


Fig. 5.1 The testing environment, showing the simulated MiRo robot in a circular arena. A) Place fields are spread evenly across the environment, with some overlap, and place cell rates are determined by the normally distributed input computed as a function of MiRo's distance from the place field's centre. B) Place cells (blue, bottom set of neurons) are bidirectionally connected to their eight nearest neighbours, and each connects feedforwardly to a network of action cells (red, top set of neurons). In total there are 100 place cells and 72 action cells.

5.2.3 Striatal Action Cells

The action cell values determine how MiRo moves in the environment. All place cells project feedforwardly to all action cells, as shown in Figure 5.1B, and recall that there are 72 action cells so that each action cell represents a 5 degree heading (360°/72). MiRo moves at a constant forward velocity, whereas the output of the action cells sets a target heading for MiRo to move in. This target heading is allocentric, in that the heading is relative to the arena. The activity for each action cell is denoted as y_i and the target heading as θ_{target}. To find the heading from the action cells, the population vector of the action cell values is computed as follows,

$$\theta_{target} = \arctan\left(\frac{\sum_i y_i \sin\theta_i}{\sum_i y_i \cos\theta_i}\right) \tag{5.4}$$

where θ_i is the angle coded for by action cell i. It is also possible to compute the magnitude of the population vector, which denotes how strongly the action cell activities are promoting a particular heading,

$$M_{target} = \sqrt{\left(\sum_i y_i \sin\theta_i\right)^2 + \left(\sum_i y_i \cos\theta_i\right)^2} \tag{5.5}$$

The action cells are restricted to take on values between 0 and 1, i.e. $y_i \rightarrow [0,1]$, with one useful interpretation for this value being a probability of cell spiking. This of course is different to the representation of the place cell activities, which is rate-coded, and this change is due to the nature of how the action cells are computed from the place cells, which we will turn to shortly.

Though it is most natural in this network setup for the action cells to be computed solely from the place cell network, doing this is not always that effective, particularly in the early stages when the network weights are random. The action cells are therefore also computed not only from the place cell network, but also by a separate module, termed a *correlated random walk* module. The reason for this is that the place cell network, particularly in the early stages of exploration when the weights are randomised, is often unable to make sensible directional decisions. I.e., with random network weights, the actions that the network chooses is *too* random. A simple implementation of a semi-random walk module therefore allows MiRo to explore the environment sensibly, as opposed to erratically when the randomised network weights are used. The details of the *correlated random walk* implementation is given below. But first we turn to a description of how they are computed from the place cell network.

Computing Action Cell Values from Place Cells

As mentioned, the action cells are restricted to taking on values between 0 and 1. This is because in computing the action cell values from the place cells, the incoming activity is passed through a sigmoidal activation function with upper and lower limits of 0 and 1,

$$\tilde{y}_i = \frac{1}{1 + \exp\left[-c_1 \sum_{j=1}^{100} w_{ij}^{PC\text{-}AC} x_j - c_2\right]} \tag{5.6}$$

with c_1 and c_2 determining the shape of the sigmoid. $w_{ij}^{PC\text{-}AC}$ represents the weight projecting from place cell j onto action cell i.

It is of course possible to select other types of activation functions for $\tilde{y}_i$, and more is discussed on this output selection in the learning rule derivation below (Section 5.2.4). The reason for setting the activity as $\tilde{y}_i$ is that the final activity as computed from the place cell inputs, termed y_i^{PC}, is drawn from a probability distribution that has $\tilde{y}_i$ as its mean. Using a probability distribution in this manner has the effect of encouraging MiRo to explore, as opposed to always selecting the actions computed deterministically from the network. Formally then, y_i^{PC} is drawn from a Gaussian distribution with mean $\tilde{y}_i$ and variance σ^2,

$$Y_i^{PC} \sim \mathcal{N}\left(\tilde{y}_i, \sigma^2\right) \tag{5.7}$$

where Y_i^{PC} is a random variable from which a specific value for y_i^{PC} is chosen.

The action cells as computed by the place cells do not always give strong preferences for any direction. For instance, at the start of a new experiment, the weights connecting the place cells to the action cells are randomised. Under these initial randomised conditions, computing the magnitude of the action cell population vector as per Equation 5.5 usually gives a very small magnitude. Furthermore, the nature of the randomness results in erratic behaviour from MiRo if these action cell values were to be chosen. Given this, it is more sensible under these conditions to compute a desired heading using a less chaotic *correlated random walk* method.

Computing Action Cell Values using the Correlated Random Walk Module

To compute the heading as determined by the correlated random walk implementation, a small but random value, θ_{noise}, is added to MiRo's current heading,

$$\theta_{random\_walk} = \theta_{current} + \theta_{noise} \tag{5.8}$$

where θ_{noise} is a random variable taken from the uniform distribution $\theta_{noise} \sim \text{unif}(-50^{\text{o}}, 50^{\text{o}})$. This ensures that MiRo generally keeps moving in its current direction, but is capable of changing slightly to the left or right, though by no more than 50$^{\text{o}}$.

To convert this into the form of action cell values, each action cell is computed as a function of its angular distance from θ_{random_walk}, in a similar manner to how the place cell activities were computed as the Cartesian distance of MiRo from the place cell centres,

$$y_i^{random\_walk} = y_i^{max} \exp\left[-\frac{(\theta_{random\_walk} - \theta_i)^2}{2\theta_d^2}\right] \tag{5.9}$$

where y_i^{max} determines the maximum value for y_i, in this case 1, and θ_d determines the distribution width. Applying Equation 5.4 on the resultant action cell values will return the value of θ_{random_walk}.

Choosing Between the Place Cell Proposal or Correlated Random Walk Module

In order to select whether the final action cell values should be computed using the place cell inputs or the correlated random walk module, the magnitude of the population vector of place cell inputs is first computed. Using Equation 5.6, the proposed action cell values from the place cell network is found, after which their population vector's magnitude is determined from Equation 5.5. If this magnitude is greater than 1, the final action cell values are computed according to Equation 5.7, otherwise they are computed using the correlated random walk module. To state this more formally, let the magnitude of the place cell network proposal be (using Equation 5.5),

$$M_{PC\_proposal} = \sqrt{\left(\sum_i \tilde{y}_i \sin\theta_i\right)^2 + \left(\sum_i \tilde{y}_i \cos\theta_i\right)^2} \tag{5.10}$$

then the final action cell values are computed as,

$$y_i = \begin{cases} y_i^{PC} & \text{if } M_{PC\_proposal} \geq 1 \\ y_i^{random\_walk} & \text{otherwise} \end{cases} \tag{5.11}$$

using Equations 5.7 and 5.9 to determine y_i^{PC} and $y_i^{random\_walk}$, respectively.

Computing Action Cells During Reverse Replays

The computation for y_i in Equation (5.11) is suitable for the exploration stage, but requires a minor modification in order for the action cells to properly replay during reverse replay events. Thus far, y_i is computed either by taking the network's output as determined by the place cell inputs or, if this output is weak, by using a correlated random walk module. In order for the y_i term to compute properly in the reverse replay case then, we add a third method for computing y_i,

$$y_i^{replay} = \frac{1}{1 + \exp\left[-c_1 \sum_{j=1}^{100} \left(w_{ij}^{PC\text{-}AC} + 0.1\frac{e_{ij}^r}{|e_{ij}^r|}\right) x_j - c_2\right]} \tag{5.12}$$

which is the same computation as Equation (5.6), with the only difference being that the place cell to action cell weights have added to them the eligibility trace *at the time of reward retrieval* for that synapse, normalised and multiplied by a value of 0.1 (i.e. the $0.1\frac{e_{ij}^r}{|e_{ij}^r|}$ term). The term e_{ij}^r represents the value of e_{ij} at the moment of reward retrieval.

As will be shown below in the weight update rule, there is an important term, $(y_i - \tilde{y}_i)$, that takes the difference between the actual action cell values and the place cell proposed action cell values. Therefore, for cases in which $e_{ij} > 0$ at the moment of reward retrieval (so that $e_{ij}^r > 0$), the term $(y_i - \tilde{y}_i)$ becomes greater than 0, whereas the opposite is of course true in the case in which $e_{ij} < 0$ at reward retrieval. If $e_{ij} = 0$, then $(y_i - \tilde{y}_i)$ also equals 0 and there is no change to the eligibility trace. In this way then, the action cell values, y_i, replay the appropriate history during a reverse replay event, so that the eligibility trace can update appropriately given MiRo's recent experience.

5.2.4 Place Cell to Action Cell Synaptic Plasticity

The weights connecting the place cells to the action cells, $W^{PC\text{-}AC}$, determine how, given the place cell activities $\mathbf{x}$, the action cells, $\mathbf{y}$, respond. The goal of the network then is to learn a set of weights that, through modulating the action cell outputs, minimises the time MiRo takes to reach the hidden goal location. To do this, a learning rule of the following form is used,

$$\frac{dw_{ij}^{PC\text{-}AC}}{dt} = R\frac{\eta}{\sigma^2}e_{ij} \tag{5.13}$$

where R is a reward signal and is a scalar value, η is a learning rate, and σ is the standard deviation as per Equation 5.7. Again, $w_{ij}^{PC\text{-}AC}$ represents the weight projecting from place cell j onto action cell i. The term e_{ij} is an eligibility trace, and is a time decaying function of the place cell and action cell values, determined by,

$$\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + (y_i - \tilde{y}_i)(1 - \tilde{y}_i)\tilde{y}_i x_j \tag{5.14}$$

Notice that these two sets of equations are in the form of a three-factor learning rule as proposed in Equation (5.2). It is simple to intuit how this learning rule behaves. Firstly, in Equation (5.13), learning only occurs for cases in which the reward signal, R, is nonzero, and when the eligibility trace, e_{ij}, is also nonzero. In the eligibility trace of (5.14), there are 4 terms that influence its state. The two terms at the end, $\tilde{y}_i$ and x_j, ensures that changes in the eligibility trace (and therefore in learning) only occurs for instances in which both the pre- and post-synaptic neurons are co-active – this establishes correlation in the learning rule, following the standard Hebbian protocol for learning (Hebb, 1949). $(1 - \tilde{y}_i)$ acts as a saturating term, so that as $\tilde{y}_i$ approaches 1 (its max value) the eligibility trace is prevented from increasing. Finally, $(y_i - \tilde{y}_i)$ is important in determining how close the network's output proposal, $\tilde{y}_i$, was from the chosen action cell value, y_i. For instance, if the chosen action cell value of y_i led to a reward, but the proposal by the network was small, such that $(y_i - \tilde{y}_i) > 0$, the learning rule would behave in a way to increase the weights so that $\tilde{y}_i$ is closer to the rewarded action y_i. Conversely, for the case in which $(y_i - \tilde{y}_i) < 0$, this suggests that the action y_i did not lead to the reward, instead it being another action that was responsible. Hence in this instance, the weights that led to a high value for $\tilde{y}_i$ should be appropriately reduced in order to account for their lower responsibility in leading to the reward.

Having provided the intuition behind the learning rule, a formal derivation of how this learning rule was arrived at is given next.

Derivation of the Learning Rule

The following derivation follows the same lines of reasoning as in Vasilaki, Frémaux, et al. (2009). The primary difference is that whilst they performed their derivation on spiking neurons, this one is performed on continuous valued neurons.

The derivation follows a policy gradient method (Sutton and Barto, 2018), where the performance measure is taken to be the average accumulated reward. The expectation for the rewards earned from time $t = 0$ to $t = T$ (which is taken to be the time for a single trial) for a given sequence of place cell activities and action cell activities can be computed according to,

$$\langle R_T \rangle = \int d\mathbf{x}_T d\mathbf{y}_T \, R(\mathbf{x}_T, \mathbf{y}_T) P_w(\mathbf{x}_T, \mathbf{y}_T) \tag{5.15}$$

where $\mathbf{x}_T = (x_0, x_1, x_2, ..., x_T)$ and $\mathbf{y}_T = (y_0, y_1, y_2, ..., y_T)$ are the sequence of place cell and action cell activities up to time T, respectively. $R(\mathbf{x}_T, \mathbf{y}_T)$ is therefore the total amount of reward earned in relation to a given sequence of place cell and action cell activities, with $P_w(\mathbf{x}_T, \mathbf{y}_T)$ being the probability that for a given set of place cell to action cell weights, w, the sequence of place cell and action cell activities arise. The task is then to take the partial derivative of this with respect to a particular weight, and update the weight proportionally to this,

$$\frac{dw_{ab}}{dt} = \eta \frac{\partial \langle R_T \rangle}{\partial w_{ab}} \tag{5.16}$$

Rewriting the partial derivative above using the definition in (5.15) gives,

$$\frac{\partial \langle R_T \rangle}{\partial w_{ab}} = \int \mathrm{d}\mathbf{x}_T \mathrm{d}\mathbf{y}_T \; R(\mathbf{x}_T, \mathbf{y}_T) \frac{\partial P_w(\mathbf{x}_T, \mathbf{y}_T)}{\partial w_{ab}} \tag{5.17}$$

from which the partial derivative on the right hand side is what we aim to derive in the following. Now the probability, $P_w(\mathbf{x}_T, \mathbf{y}_T)$, can be re-written as (decomposition of the probability given in Vasilaki, Frémaux, et al., 2009),

$$P_w(\mathbf{x}_T, \mathbf{y}_T) = \prod_j g_j(\mathbf{x}_T, \mathbf{y}_T) \prod_i h_i(\mathbf{x}_T, \mathbf{y}_T) \tag{5.18}$$

where g_j is the single neuron probability of a place cell with index j taking on a particular rate value, and is determined by the activity of the other place and action cell activities. This is implicit, since place cell activities determine action cell activities, which generates actions which in turn affects place cell activities. Likewise, h_i is the single neuron probability of an action cell with index i taking on a particular rate. This is more explicit since the place cell activities determine the values of the action cells through their connection matrix.

Recall from Equation (5.7) that the probability distribution for the action cells is Gaussian centred on the output of the place cell to action cell network. Taking h_i to be a Gaussian distribution then, with a mean centred on the action cell outputs and a variance σ^2,

$$h_i(\mathbf{x}_T, \mathbf{y}_T) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \tilde{y}_i)^2}{2\sigma^2} \right) \tag{5.19}$$

where $\tilde{y}_i$ is a sigmoidal function of the incoming weighted place cells, $\tilde{y}_i = f_s\left(\sum_j w_{ij} x_j\right)$, as per Equation (5.6). With this in mind, for the single weight, w_{ab}, multiplying the partial derivative of the probability distribution above by $\frac{h_a(\mathbf{x},\mathbf{y})}{h_a(\mathbf{x},\mathbf{y})}$ gives,

$$\begin{aligned}\frac{\partial P_w(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}} &= \frac{\partial}{\partial w_{ab}}\left[\frac{P_w(\mathbf{x}_T,\mathbf{y}_T)}{h_a(\mathbf{x}_T,\mathbf{y}_T)}h_a(\mathbf{x}_T,\mathbf{y}_T)\right] \\ &= \frac{P_w(\mathbf{x}_T,\mathbf{y}_T)}{h_a(\mathbf{x}_T,\mathbf{y}_T)}\frac{\partial h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}}\end{aligned} \tag{5.20}$$

The second equality arises due to $\frac{P_w(\mathbf{x}_T,\mathbf{y}_T)}{h_a(\mathbf{x},\mathbf{y})}$ being no longer dependent on the weight w_{ab}, as it is divided out of the product in (5.18). In fact, it is no longer dependent on any of the weights in the weight vector w_b that projects onto action cell a. Using the definition that,

$$\frac{\partial \ln h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial h_a(\mathbf{x}_T,\mathbf{y}_T)} = \frac{1}{h_a(\mathbf{x}_T,\mathbf{y}_T)} \tag{5.21}$$

and substituting this into (5.20) gives,

$$\begin{aligned}\frac{\partial P_w(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}} &= P_w(\mathbf{x}_T,\mathbf{y}_T)\frac{\partial \ln h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial h_a(\mathbf{x}_T,\mathbf{y}_T)}\frac{\partial h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}} \\ &= P_w(\mathbf{x}_T,\mathbf{y}_T)\frac{\partial \ln h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}} \\ &= P_w(\mathbf{x}_T,\mathbf{y}_T)\frac{(y_a-\tilde{y}_a)}{\sigma^2}\frac{\partial \tilde{y}_a}{\partial w_{ab}} \\ &= P_w(\mathbf{x}_T,\mathbf{y}_T)\frac{(y_a-\tilde{y}_a)}{\sigma^2}(1-\tilde{y}_a)\tilde{y}_a x_b\end{aligned} \tag{5.22}$$

The third equality is the result of performing the differentiation, via the chain rule, of $\frac{\partial \ln h_a(\mathbf{x}_T,\mathbf{y}_T)}{\partial w_{ab}}$ using the definition for $h_a(\mathbf{x}_T,\mathbf{y}_T)$ in Equation (5.19). The last equality arises by performing the differentiation of $\frac{\partial \tilde{y}_a}{\partial w_{ab}}$ using the sigmoidal activation function definition for $\tilde{y}_a$ from Equation (5.6), with the constant (c_1) later being absorbed into the learning rate, hence left out here. Substituting back into (5.17) provides the final expression,

$$\frac{\partial \langle R_T\rangle}{\partial w_{ab}} = \int d\mathbf{x}_T d\mathbf{y}_T\, R(\mathbf{x}_T,\mathbf{y}_T)P_w(\mathbf{x}_T,\mathbf{y}_T)\frac{(y_a-\tilde{y}_a)}{\sigma^2}(1-\tilde{y}_a)\tilde{y}_a x_b \tag{5.23}$$

To approximate this then, one could take an average of

$$R(\mathbf{x}_T,\mathbf{y}_T)\frac{(y_a-\tilde{y}_a)}{\sigma^2}(1-\tilde{y}_a)\tilde{y}_a x_b \tag{5.24}$$

over a number of trials and use this to update the synaptic weight after enough trials have occurred, in order to accurately average. However, the biologically plausible update method would be to update "online" instead (omitting the $(\mathbf{x}_T, \mathbf{y}_T)$ dependencies for simplification),

$$\frac{dw_{ab}}{dt} = \eta R \frac{(y_a - \tilde{y}_a)}{\sigma^2} (1 - \tilde{y}_a) \tilde{y}_a x_b \tag{5.25}$$

One last modification is needed to include an eligibility trace as shown in Equations (5.13) and (5.14). To do so, one can absorb an exponential kernel into the learning rate in order to effectively capture the decaying time dynamics that result from the decaying effects of the eligibility trace (see Methods in Vasilaki, Frémaux, et al., 2009).

It is worth noting here that the computation for $\tilde{y}_i$ can be replaced with any other function, but that the $(y_i - \tilde{y}_i)/\sigma^2$ will always remain, due to the Gaussian probability distribution assumption of Equation (5.19). For instance, if one replaces the computation for $\tilde{y}_i$ with a simple summation such as $\tilde{y}_i = \sum_j w_{ij}^{PC\text{-}AC} x_j$, then the final learning rule will instead be (using the indices ij rather than ab),

$$\frac{dw_{ij}}{dt} = \eta R \frac{(y_i - \tilde{y}_i)}{\sigma^2} x_j \tag{5.26}$$

For the reasons discussed above, the $(y_i - \tilde{y}_i)$ plays an important role in the learning rule, and as seen remains regardless of the choice for the output function of $\tilde{y}_i$. But it is perhaps more interesting to consider this difference term with regards to an action selection mechanism in the basal ganglia. More is discussed regarding this in Section 5.4.

5.2.5 Review of the Implementation

A description of the full implementation process is provided here, with an overview of the algorithmic implementation presented in the Algorithmic Implementation box below. This is the procedure taken for a single experiment, which usually consists of 20 trials.

Initialisation – At the start of a new experiment, the weights that connect the place cells to the action cells are randomised and then normalised. All the variables for the place cells are set to their steady state conditions for when no place specific inputs are present, and the action cells are all set to zero. MiRo is then placed into a random location in the arena.

Taking Actions – There are three main actions MiRo can make, depending on whether the reward it receives is positive +1 and is therefore at the goal, negative -1 such that MiRo has reached a wall, or 0 for neither of these two cases. If the reward is 0, the action cell values, y_i, are computed from either y_i^{PC} or $y_i^{random\_walk}$ according to Equation (5.11), from which a heading is computed using Equation (5.4). MiRo moves at a constant forward velocity and a new heading is computed every 0.5s, for which it then changes course and moves in this new heading direction. If MiRo reaches a wall, a wall avoidance procedure is used which turns

MiRo round 180^{o}. Finally, if MiRo reaches the goal, it pauses there for 2s, after which it heads to a new random starting location.

Determining Reward Values – There are three reward values that MiRo can collect. If MiRo has reached a wall, a reward of R = -1 is presented to MiRo for a period of 0.5s, which tends to occur during MiRo's wall avoidance procedure. If MiRo has found the goal, a reward of R = +1 is presented to it for a period of 2s. And if neither of these conditions are true, then MiRo receives no reward, i.e. R = 0.

Initiating Reverse Replays – Reverse replays are only inititiated when MiRo reaches the goal location, but not for when MiRo is avoiding a wall. For the case in which reverse replays are initiated, λ is set to 1 to allow hippocampal synaptic conductance, and the place specific input for MiRo's position whilst at the goal, I_j^{place}, is injected 1s after MiRo first reaches the goal for a total time of 0.5s. With synaptic conductance enabled, and due to intrinsic plasticity, this initiates reverse replay events initiating at the goal location and traveling back through the recent trajectory in the place cell network.

Updating Network Variables – Regardless of whether MiRo is exploring, avoiding a wall, or is at the goal and is initiating replays (or not), all the network variables, including the weight updates, occur for every time step of the simulation. It is only when MiRo has reached the goal, gone through the 2s of reward collection, and is making its way to a new random starting location that all the variables are reset as in the Initialisation step above (though excluding the randomisation of the weights). This would then begin a new trial in the experiment.

Python Code

Full code for the model, which includes Python 2.7 code for model implementation (using ROS Kinetic) and Python 3.5 code for plotting visualisations, can be found at https://github.com/mattdoubleu/robotic_RL_replay.

Algorithmic Implementation

1. Initialisation:
   - MiRo is placed into a random start location.
   - All place cell variables set to steady state conditions for zero place cell input.
   - All action cell values set to zero.
   - Weights $w_{ij}^{PC\text{-}AC}$ randomised and normalised: $w_{ij}^{PC\text{-}AC} \leftarrow \frac{w_{ij}^{PC\text{-}AC}}{\sum_i w_{ij}^{PC\text{-}AC}}$.
2. Determine MiRo's movement and reward values:
   - If found_goal:
     - For 2s: R = 1; $\lambda = 1$; MiRo_movement = stalled.
     - If this experiment includes replays: initiate reverse replay event after 1s; $y_i = y_i^{replay}$ during replay.
     - After 2s: $\lambda = 0$; MiRo_movement = move_to_random_location.
   - Else If detected_wall:
     - For 0.5s: R = -1.
     - MiRo_movement = wall_avoidance_procedure.
   - Else:
     - R = 0.
     - If 0.5s has passed since last action:
       - If $M_{PC\_proposal} > 1$: $y_i = y_i^{PC} \, \forall i$
       - Else: $y_i = y_i^{random\_walk} \, \forall i$.
       - Compute θ_{target} from y_i and set MiRo_movement to move towards this heading with constant forward velocity.
3. Update network variables:
   - Update place cells based on MiRo's position in the environment.
   - Compute action cell outputs.
   - Update eligibility trace: $\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + (y_i - \tilde{y}_i)(1 - \tilde{y}_i)\tilde{y}_i x_j \, \forall i, j$.
   - Update weights and then normalise: $\frac{dw_{ij}^{PC\text{-}AC}}{dt} = R \frac{\eta}{\sigma^2} e_{ij} \, \forall i, j$.
4. Return to Step 2 and repeat.

5.2.6 A Post-Synapse Only Eligibility Trace

The learning rule and eligibility trace of Equations (5.13) and (5.14) follow from the derivation above and in Vasilaki, Frémaux, et al. (2009), with the eligibility trace in particular capturing the co-dependencies of the pre- and post-synaptic activities. But this does not necessarily afford the ability to learn “shortcuts” to rewards, for instance as in the model by Molter, Sato, and Yamaguchi (2007).

It may be possible, however, to create a shortcut mechanism for the open arena environment by taking an average over the previously selected actions up to the moment of receiving a reward. To achieve this, the eligibility trace is altered so that it is dependent only on the difference between the chosen action selection, y_i, and the network’s proposed action selection, $\tilde{y}_i$,

$$\frac{de_i}{dt} = -\frac{e_i}{\tau_e} + (y_i - \tilde{y}_i) \tag{5.27}$$

with the weight update rule now being modified to,

$$\frac{dw_{ij}^{PC\text{-}AC}}{dt} = R\frac{\eta}{\sigma^2}(1 - \tilde{y}_i)\,\tilde{y}_i x_j e_i \tag{5.28}$$

Notice now that the eligibility trace in (5.27) is dependent only on the action cells, although learning will still only occur for instances in which place cells activate when at the reward point, as seen in Equation (5.28). This has the effect that only “reactive” learning, or learning of the most immediate stimulus response, can occur when there are no reverse replays. This is as a result of only those place cells whose place fields occupy the space in which MiRo is positioned, i.e. at the reward, firing during reward retrieval. Replays of the previous trajectory are therefore the only means through which learning can occur over the trajectory history.

The role of τ_e in Equation (5.27) plays a slightly different role in this instance. Rather than storing the history of pre- and post-syanptic co-activities, or state-action pairs, as is the case in Equation (5.14), here it determines how much weight to apportion to actions that were selected further back in time regardless of states. For instance, small values of τ_e results in eligibility traces storing only the most recent actions, whereas large values for τ_e take into consideration actions further back in time, alongside those actions that were taken most recently. The result this has on learning is that each place cell, when activated during a reverse replay event, learns for all the previous actions, and this results in learning of the “average action” across the preceding trajectory.

5.3 Experimental Results

The network is run on a simulated MiRo robot in an open arena and tested using a Morris-water maze like test paradigm (Morris, Garrud, et al., 1982). The arena is circular with a diameter of 2m, and within it MiRo traverses the arena at a constant velocity of 0.2m/s, changing only its heading. At the start of a trial, MiRo is placed into a random position in the arena and its objective is to find a hidden goal location within the shortest time possible. Once MiRo has found the goal location, it pauses there for 2s whilst a reward signal of value 1 is presented to it. If MiRo comes close to hitting a wall, MiRo detects the wall using its sonar sensor and turns around towards the centre of the arena, whilst a reward of -1 is sent to it for a period of 0.5s. All weights connecting the place cells to the action cells are initialised randomly at the beginning of an experiment and are normalised so that the total sum of the weights projecting from a single place cell equals 1,

$$w_{ij}^{PC\text{-}AC} \leftarrow \frac{w_{ij}^{PC\text{-}AC}}{\sum_i w_{ij}^{PC\text{-}AC}} \tag{5.29}$$

The weights and all the other network variables as described in the Methodology are updated every 10ms using a discretised form of the differential equations (discretised via the Euler method with the 10ms time step, see also Section 3.2.1). After updating the weights according to Equation (5.13) they are re-normalised using Equation (5.29) at each time step. All parameters in the model are kept constant across all experiments and trials, except for the learning rate (η) and eligibility trace time constant (τ_e) in Equations (5.13) and (5.14), which are modified in order to examine performance. All parameter values are summarised in Table 5.1, with values for η and τ_e specified appropriately in the results.

This results section is divided into three subsections. Presented first are the results for when running the model without reverse replays, with the aim of showing that the network and the learning rule perform as expected. Following this, the model is then run with reverse replays, with these results being compared to the non-replay case. All model parameters and the learning rule are kept equal between the two cases for fair comparisons. Finally, an heuristic learning rule is tested in which the eligibility trace is updated using only post-synaptic activity. In this scenario, we are testing whether reverse replays are capable of effective learning despite causality – that is, co-activity between pre- and post-synaptic activity during exploration – being removed, such that an average over previous actions is instead learned.

Parameter	Value
c_1	0.1
c_2	20
σ	0.1
θ_d	10
τ_e	*See text*
η	*See text*

Table 5.1 Model parameter values used in the reinforcement learning experiments.

5.3.1 Testing the Learning Rule Without Reverse Replays

The learning rule as derived in Section 5.2.4 is given as,

$$\frac{dw_{ij}^{PC\text{-}AC}}{dt} = R\frac{\eta}{\sigma^2}e_{ij} \tag{5.30}$$

$$\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + (y_i - \tilde{y}_i)(1-\tilde{y}_i)\tilde{y}_i x_j \tag{5.31}$$

To confirm that this derived learning rule performs as expected, the network is first run without reverse replays. Figure 5.2A shows the results for the time taken to reach the hidden goal as a function of trial number, averaged across 40 independent experiments. The time to reach the goal approaches the asymptotic performance at around 5 trials. Note that, despite the larger variance towards the final two trials, given the small experimental size (40 independent experiments) there is no statistically significant difference between performance in trial 20 and performance in trial 10. ($N = 40$; $z = -1.526$; p-value $= 0.126$ in Wilcoxon Signed-Rank test.)

Figure 5.2B displays the population weight vector for the weights projecting from the place cells to the action cells. The weight population vector for a single place cell is computed as,

$$(w_j^x, w_j^y) = \left(\sum_{i=1}^{72} w_{ij}^{PC\text{-}AC}\cos\theta_i, \sum_{i=1}^{72} w_{ij}^{PC\text{-}AC}\sin\theta_i\right) \tag{5.32}$$

where (w_j^x, w_j^y) represents the x and y components for the weight population vector of the j^{th} place cell, $w_{ij}^{PC\text{-}AC}$ is the value of the weight from place cell j onto action cell i, and θ_i is the heading direction that action cell i codes for. The magnitude of the population weight vector can then be computed as,

$$M_{w_j} = \sqrt{\left(w_j^x\right)^2 + \left(w_j^y\right)^2} \tag{5.33}$$

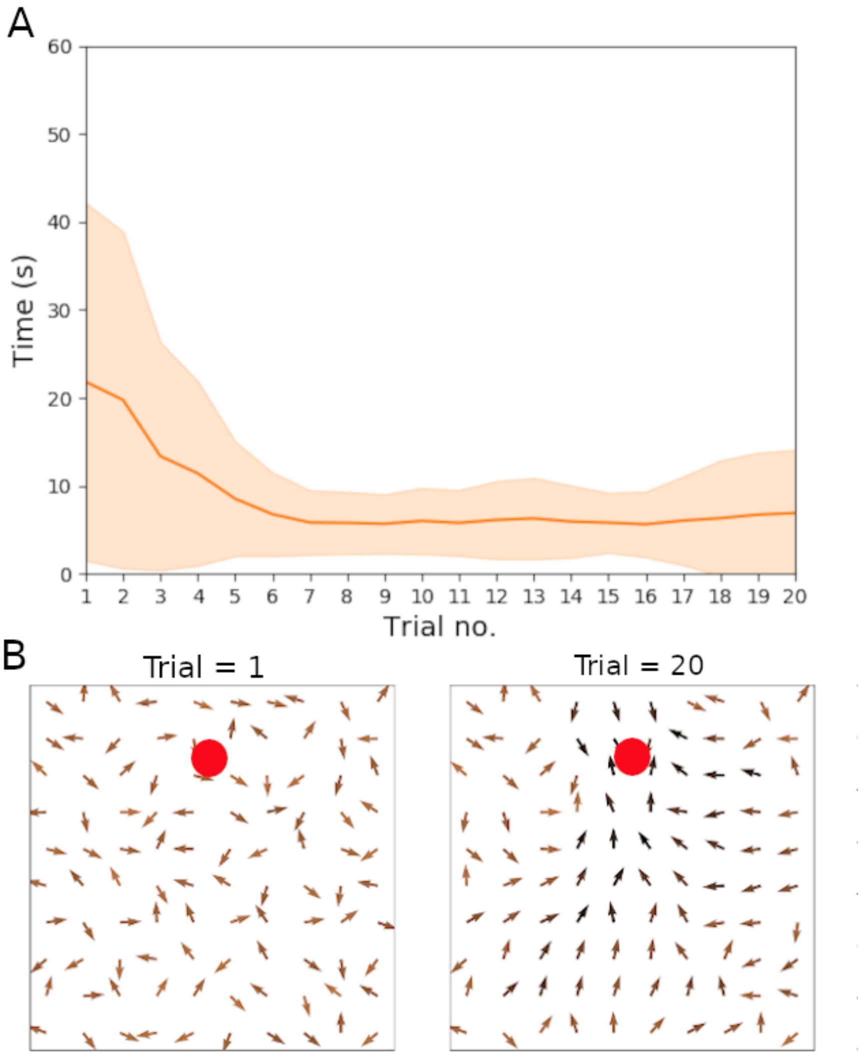


Fig. 5.2 Results for the non-replay case in order to test that the derived learning rule performs well. Parameters used were $\eta = 0.01$ and $\tau_e = 1s$. A) Plot showing the average time to reach goal (red line) and standard deviations (shaded area) over 20 trials. Averages and standard deviations are computed from 20 independent experiments. B) Weight population vectors at the start of trial 1 versus at the end of trial 20 in an example experiment. All vectors are normalised, thus magnitudes for the vectors are represented as a shade of colour; the darker the shade, the larger the magnitude. Red dots indicate the goal location.

Weight population vectors whose directions are aimed towards the goal have the effect of producing actions that move MiRo in the direction of the goal.

5.3.2 Effect of Reverse Replays on Performance

Using first the same learning rate and eligibility trace time constant as in the non-replay case above, the performance average shows not to have any significant difference ($p > 0.05$ across 18 trials and $p > 0.01$ across 19 trials in a Wilcoxon Signed-Rank Test; see Appendix A.2 for all experimental results). Average time to reach goal over the last 10 trials is 6.21s in the non-replay case and 6.92s in the replay case (data not shown, see Appendix A.2). This suggests replays are at least as good when compared to the best case non-replay. Results on performance of varying the learning rate and eligibility trace time constant are presented next.

Reducing the Eligibility Trace Time Constant

Given the standard, non-replay model requires the recent history to be stored in the eligibility trace, it follows that having too small an eligibility trace time constant might negatively impact the performance of the model. Reverse replays however may have the potential to compensate for this, since the recent history is also stored, and then replayed, in the place cell network. Figure 5.3 shows the effects on performance of significantly reducing the eligibility trace time constant (to $\tau = 0.04s$). Both cases, with and without reverse replays, are compared. If the learning rate is too small ($\eta = 0.01$) then for neither case is there any learning. But as the learning rate is increased, having reverse replays shows to significantly improve performance. Similar but less significant results are found for a learning rate of $\tau_e = 0.2s$ (see Appendix A.2).

To explore why replays perform significantly better at small eligibility trace time constants, Figure 5.4 displays example plots of the eligibility trace population vectors at reward retrieval. Population vectors for the eligibility traces are computed according to Equations 5.32 and 5.33, but replacing the weights ($w_{ij}^{PC\text{-}AC}$) with the eligibility traces (e_{ij}). At the first point of reward retrieval ($t^r = 0s$), the eligibility traces with and without reverse replays are the same. However, 1.5s after reward retrieval, or 0.5s after replay initiation, whereas the non-replay eligibility trace has decayed to near zero for all weights, the replay case has re-activated the eligibility trace. This reactivation therefore boosts learning speed.

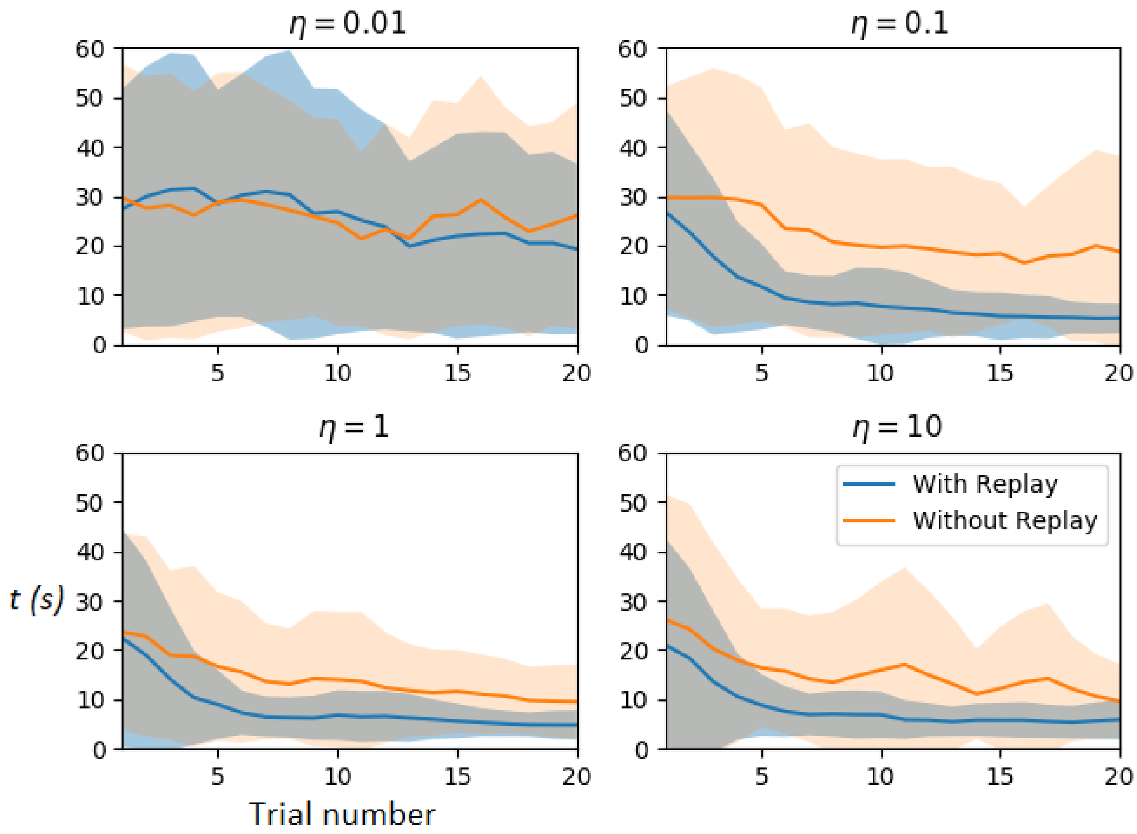


Fig. 5.3 Comparing the effects of a small eligibility trace time constant with and without reverse replays. $\tau_e = 0.04s$ across all figures. Thick lines are averages across 40 independent experiments, with shaded areas representing one standard deviation. The moving averages, averaged across 3 trials, are plotted here for smoothness.

Effects of Small Learning Rate on Performance

Since reverse replays offer additional opportunities for learning during the reward retrieval stage, it may follow that reverse replays improve performance for instances where the learning rate is small. Figure 5.5 shows results for when the learning rate is small ($\eta = 0.001$). Perhaps the only significant result is for $\tau_e = 1s$, where average performance and variance with replays is marginally better than without replays ($p - value < 0.05$ in 5 trials in a Wilcoxon Signed-Rank Test). Importantly though, reverse replays only ever perform equally well or better than the case with no replays. Results are similar for various other small learning rates (see Appendix A.2).

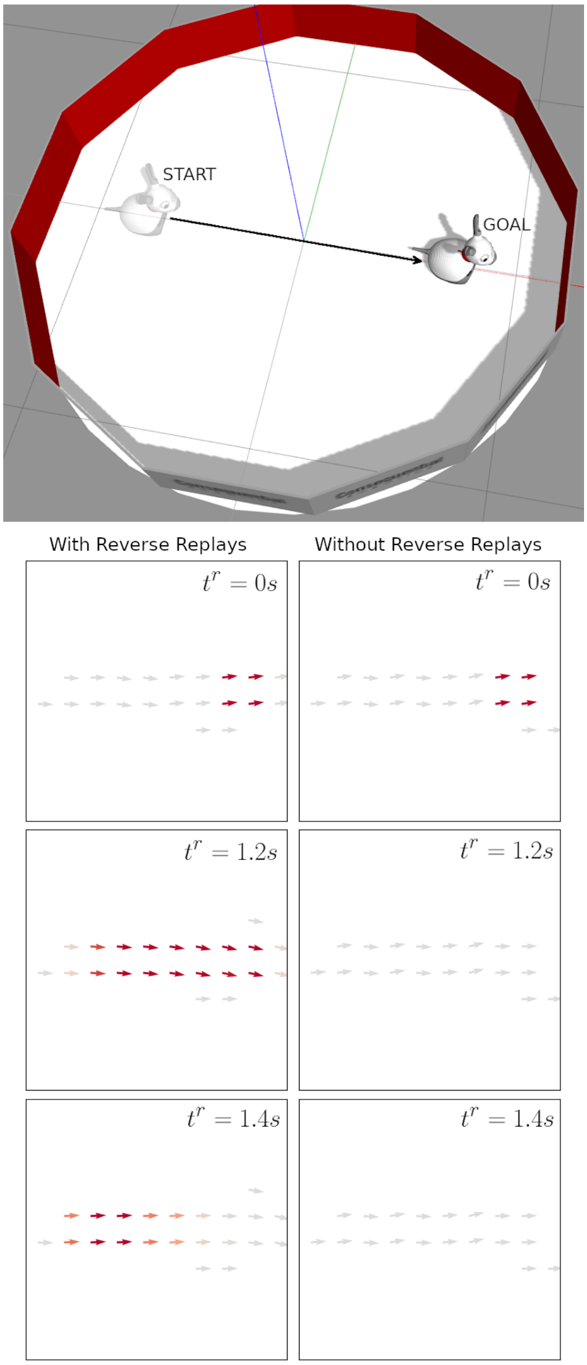


Fig. 5.4 Comparison of the eligibility trace at $t^r = 0s$, $t^r = 1.2s$ and $t^r = 1.4s$, where t^r represents the time after reward retrieval. Top figure shows the straight line path MiRo took towards the goal. Bottom plots show the population vector plots for the eligibility traces with reverse replays (left) and without reverse replays (right). Parameters used are $\eta = 0.1$ and $\tau_e = 0.04s$.

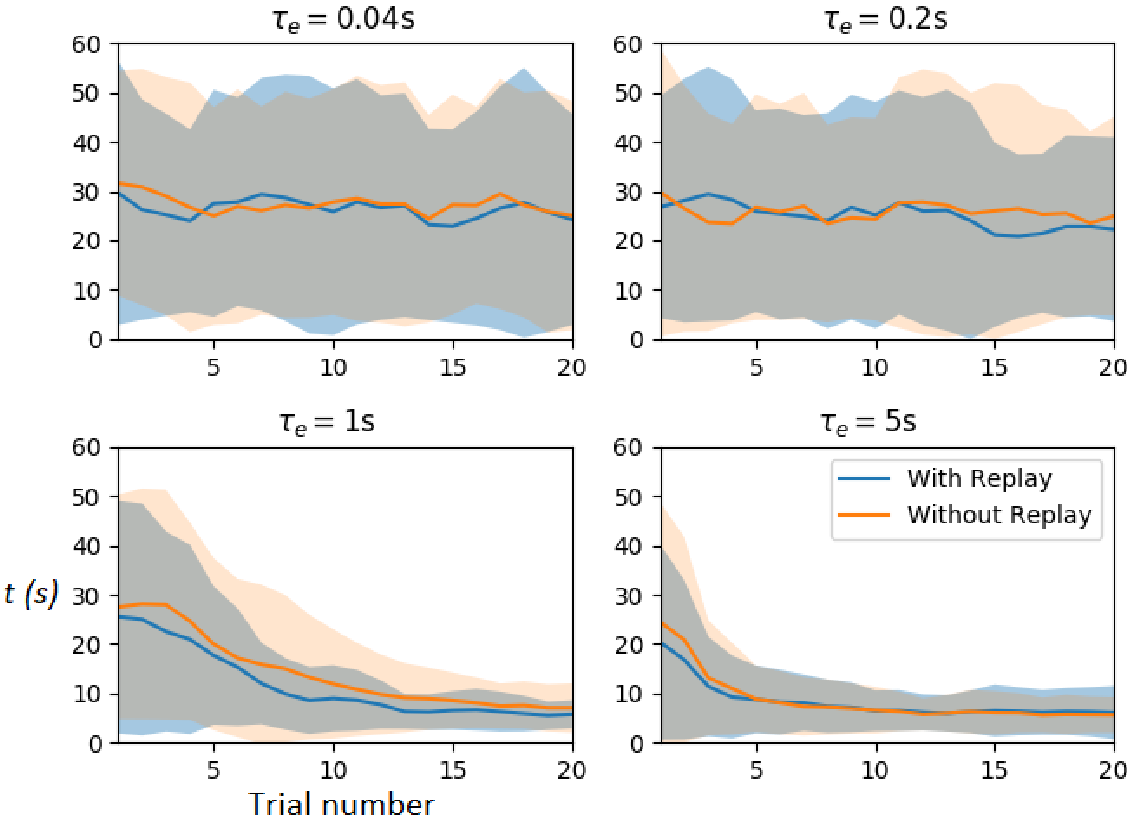


Fig. 5.5 Comparison of performance for when the learning rate is small. Here $\eta = 0.001$ across all figures. Solid lines represent the trial averages taken across 40 independent experiments, whilst shaded areas denote one standard deviation.

Comparison of Best Cases

Figure 5.6 compares the results for the best cases with and without reverse replays. There is no statistical significance across all trials, despite the apparently large deviations in the final few trials of the non-replay case ($p - value > 0.05$ for all 20 trials in a Wilcoxon Signed-Rank Test). What is most striking is the difference in parameters for the best cases. With reverse replays the parameters are $\tau_e = 0.04s$, $\eta = 1$, whereas without reverse replays they are $\tau_e = 1s$, $\eta = 0.01$. More is discussed regarding this in Section 5.4.

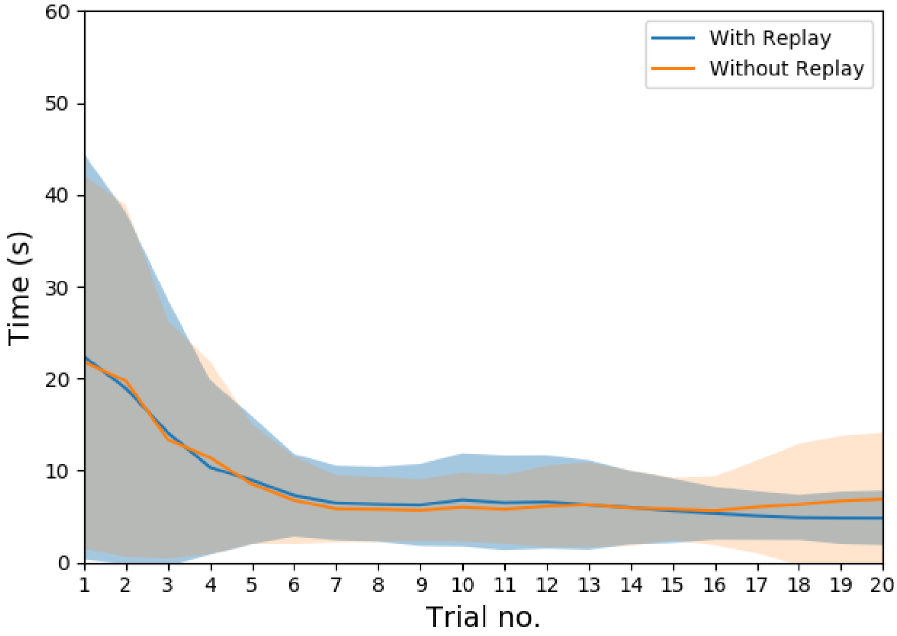


Fig. 5.6 Comparing the best cases with and without reverse replays. With reverse replays the parameters are $\tau_e = 0.04s$, $\eta = 1$. Without reverse replays the parameters are $\tau_e = 1s$, $\eta = 0.01$.

Performance Across Parameter Space

It is perhaps worth comparing performance across various parameters of τ_e and η. Figure 5.7 displays the average performance over the last 10 trials, comparing again with replays versus without replays. There are perhaps two noticeable observations to make here. Firstly, when the eligibility trace time constant is small, employing reverse replays shows considerable improvements in performance over the non-replay case across the various values of learning rates. Learning still exists in the non-replay case, however, it is noticeably diminished compared with the replay case. Secondly, although this marked improvement in performance vanishes for larger eligibility trace time constants, reverse replays do not at the very least hinder performance.

5.3.3 Preliminary Results for the Modified Learning Rule

Preliminary results show that the modified learning rule, which removes the eligibility trace from the synapses and onto the action cells, has the potential to learn shortcut trajectories towards the goal (Figure 5.8). For comparison, the weight population vectors after reward retrieval for a curved path are shown in the standard replay case and the non-replay case. No-

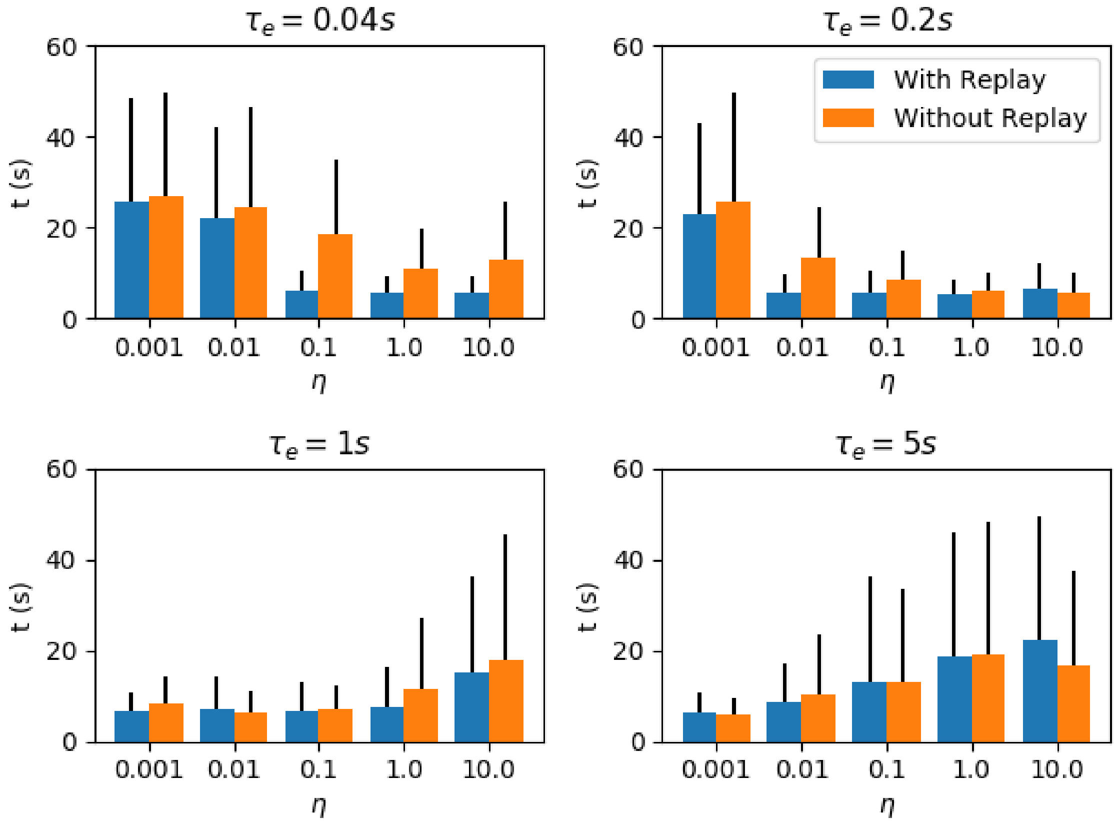


Fig. 5.7 Comparing average performance across a range of values for τ_e and η. Bars show the average time taken to reach the goal, averaged over the last 10 trials, with error bars showing one standard deviation.

tice that the primary difference between the standard replay and non-replay is the magnitude of the final weight vectors, in which replay has greater weight magnitudes.

Two results for the modified learning rule are shown, each using different eligibility trace time constants. In the first, the same time constant as in the standard replay and non-replay cases is used ($\tau_e = 1s$), whilst in the second a larger time constant is used ($\tau_e = 10s$). Since the modified learning rule takes a "weighted" average in updating the weights, the smaller the time constant the more weight is placed on the most recent actions. This is clear when comparing the weight population vectors for $\tau_e = 1s$ and $\tau_e = 10s$, where in the first instance the most recent actions (north-west heading, Figure 5.8C) has more weight, whereas in the second there appears to be more equal weighting placed on all actions across the trajectory (north heading, Figure 5.8D).

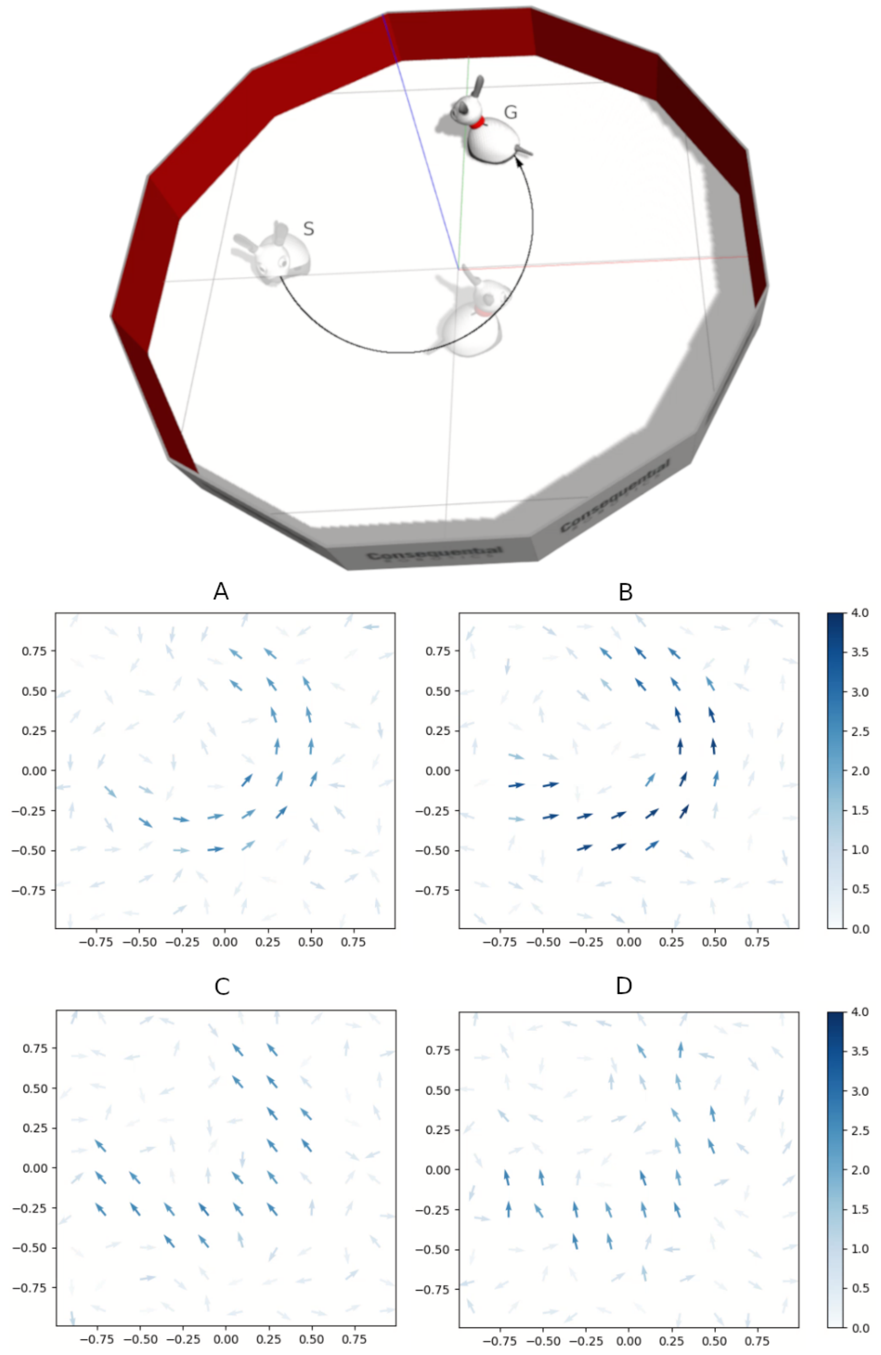


Fig. 5.8 Population weight vectors after reward retrieval in the non-replay, replay, and modified learning rule cases. Top figure shows the path taken by MiRo, where S represents the starting location and G the goal location. Top plots show weight population vectors for the non-replay case (A) and standard replay case (B) with $\tau_e = 1s$; $\eta = 0.1$. Bottom plots show weight population vectors for the modified learning rule with $\tau_e = 1s$; $\eta = 0.1$ (C) and $\tau_e = 10s$; $\eta = 0.01$ (D).

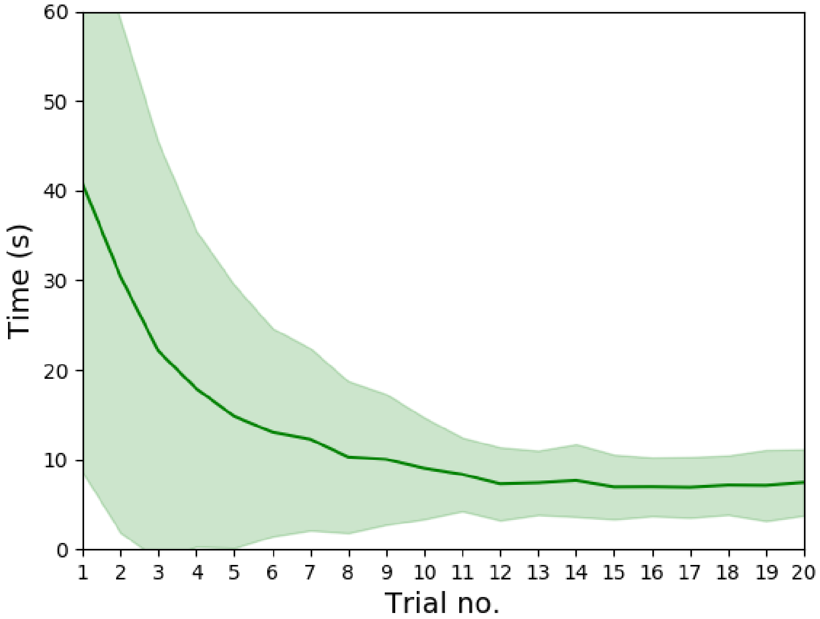


Fig. 5.9 Performance results for the modified learning rule using a post-synapse only eligibility trace time constant. Parameters used here are $\tau_e = 5s$ and $\eta = 0.001$.

The full results of performance for the modified learning rule using parameters $\tau_e = 5s$ and $\eta = 0.001$ is shown in Figure 5.9. Notice that learning is still achievable despite the causality of pre- and post-synaptic activities being removed from the eligibility trace.

5.4 Discussing the Model and Results

Hippocampal reverse replay has long been implicated in reinforcement learning (Foster and Wilson, 2006), but how the dynamics of hippocampal replay produce behavioural changes, and why hippocampal replay could be important in learning, are still ongoing questions. By embodying first a hippocampal-striatal inspired model (Vasilaki, Frémaux, et al., 2009) into a simulated MiRo robot, and then augmenting it with a model of hippocampal replay (Whelan, Prescott, and Vasilaki, in press), we have been able to examine the link between hippocampal replay and behavioural changes in a spatial navigation task.

In the three-factor, synaptic eligibility trace hypothesis, the time constants for the traces have been argued to be on the order of a few seconds, necessary for learning over behavioural time scales (Gerstner et al., 2018). However, results here indicate that the reinstatement of synaptic eligibility traces during reverse replays suggests it is not necessary for synaptic eligibility trace time constants to be on the order of seconds – a few milliseconds is sufficient. Yet the synaptic eligibility trace is still required here for storing the history; it just does not matter how much of the eligibility trace is stored – it is only important that enough is stored for effective reinstatement during a reverse replay. It has also been argued that neuronal, as opposed to synaptic, eligibility traces could be sufficient for storing a memory trace, as in the two-compartmental neuron model of (Brea et al., 2016). Intrinsic plasticity in this model is not unlike a neuronal eligibility trace, storing the memory trace within the place cells for reinstatement at the end of a rewarding episode.

Whilst it could be the case that reverse replays speed up learning by allowing for additional learning time, the results shown here provide some, but not strong, support for this. Experimental evidence does show however that disruption of hippocampal ripples during awake states, when reverse replays occur, does disrupt but not completely diminish spatial learning in rats (Jadhav et al., 2012). Whilst the longer eligibility trace time constants in this model ($\tau_e = 1s, 5s$) do not show diminished performance without reverse replays, the smaller time constants ($\tau_e = 0.04s, 0.2s$) do. Hence, these results support the view that reverse replays enhance, rather than provide entirely, the mechanism for learning. Beyond reverse replays however, forward replays have been known to occur on multiple occasions for up to 10 hours post-exploration (Giri et al., 2019), which could be more important for memory consolidation than awake reverse replays (Girardeau et al., 2009; Ego-Stengel and Wilson, 2010).

In the case of the best versus best case comparison (Figure 5.6), it is clear why a sufficiently large, but not too large, eligibility trace time constant for the non-replay case gives best performance – it must store a suitable amount of the trajectory history for learning. If the eligibility trace time constant were too small, it would not store enough of the history, whereas too large and it stores unnecessary trajectories that go too far back in time. It is interesting however to see that replays perform better with smaller eligibility trace time constants. One possible reason for this, as shown in Figures 5.8A and 5.8B, is that reverse replays increase the magnitude of weight changes. Again, if the time constant were too large, the increased learning of reverse replays would learn unnecessary trajectories going too far back in time. This can be seen through looking at Figures 5.8A and 5.8B, and noticing that learning reduces near the start of the trajectory in the non-replay case, but is strong in the replay case. Having a small eligibility trace time constant can therefore reduce this effect.

In the modified learning rule, whilst the preliminary results show a possible mechanism for learning shortcut routes, its overall performance appears slightly worse than the original learning rule. One explanation for this worse performance is that at the goal location, place cells would activate during reverse replay events despite them not taking part in the recent behavioural trajectory; notably place cells that were just in front of MiRo's position. Since the modified learning rule no longer associates behavioural causalities between place and action cells, these place cells would then learn the incorrect actions. It is interesting to note, however, that rewards have been found to shift the location of place fields for place cells (Tsuneyuki Kobayashi et al., 1997; Kobayashi et al., 2003). Furthermore, place cells tend to cluster their place fields in higher densities around rewarded sites, and away from the unrewarded sites (Hollup et al., 2001). So whilst the place fields remain static in this model, it could be that by shifting the place fields towards the reward locations would counter the problem of incorrect learning. It is a subtle difference, but rather than learning the place-action association by pairing the action to the place, this association could be achieved by pairing the place to the action; moving the place fields towards the location where the action is an appropriate one. It is important to note, however, that these are preliminary results, and it still requires further testing across the parameters.

5.4.1 Towards Sampling Efficiency

Reinforcement learning often requires there to be a large number of environmental interactions, in order to build a model of state-transition values for instance (Sutton and Barto, 2018). Yet hippocampal replay may provide help in this sampling problem, by allowing slight modifications of recent state-actions, or new unexplored paths entirely, to be replayed for rewarding experiences. This would relinquish the need to physically explore all state-action pairs. A noticeable example of this is in a model of hippocampal replay by Molter, Sato, and Yamaguchi (2007). They demonstrated that replays can pass throughout a region of the environment without the agent ever having traversed that portion of the environment, so long as the correct state-transitions are encoded into the place cell network's weights. If reverse replays were to replay not only the recently experienced trajectory, but also all other trajectories leading towards the goal, this could well speed up the learning of paths towards goal locations.

5.4.2 Hippocampal-Striatal Neuroanatomy

Recall from the review of hippocampal neuroanatomy in Chapter 2 that both the CA1 region and the Subiculum project to the striatum. How might the model presented here align with these known anatomical details? It is clear that the recurrency of the hippocampal network in this model could best be described as an approximation of region CA3, which is a region known for its strong associational connections (Ishizuka, Weber, and Amaral, 1990). And whilst there is little evidence that CA3 itself projects to the striatum, it does so through via region CA1 (Li et al., 1994; Amaral and Lavenex, 2007). But why might there be this additional layer? Whilst the functional properties for the hippocampal subregions are not yet completely known, but one hypothesis for region CA1 is that it combines the information in CA3, which must be separated for generating proper associations, into an efficient cue for retrieval in the neocortex (Rolls, 2010). It seems reasonable then that the efficiently coded memory in CA1 could be associated in striatal synapses. However, the specific functional properties of hippocampal subregions are still mostly unknown, despite there being some speculations on this (Hasselmo, 2011; Cherubini and Miles, 2015).

5.4.3 Action Selection in the Basal Ganglia

In this model, there are two sets of competing behaviours during the exploratory stage – the memory guided behaviour of the hippocampus and the correlated random walk behaviour – which are heuristically selected for based on the signal strength of the hippocampal output: If the hippocampal output does not express strongly for a particular action, the correlated random walk behaviour is implemented instead. An interesting comparison with the basal ganglia, and its input structure the striatum, could be made here, since these structures have for some time been speculated to play a role in action selection (Mink, 1996; Grillner et al., 2005; Prescott, González, et al., 2006; Redgrave, Vautrelle, et al., 2017). A basic interpretation of this action selection mechanism is that the basal ganglia receives a variety of candidate motor behaviours, each of which are perhaps mutually incompatible, but from which the basal ganglia must select one (or more) of these behaviours for expressing (Gurney, Prescott, and Redgrave, 2001a; Gurney, Prescott, and Redgrave, 2001b). Since the selection of an action in our model is determined from the striatal action cell outputs, it appears likely that this selection would occur within the basal ganglia.

But perhaps more interesting is that in the synaptic learning rule presented here, the difference between the action selected, y_i, and the hippocampal output, $\tilde{y}_i$, is used to update synaptic strengths. One interpretation for this could be that this difference behaves as an error signal, signalling to the hippocampal-striatal synapses how “good”, or how “close”, their predictions were in generating behaviours that led towards rewards. But how might this be implemented in the basal ganglia? Whilst the striatum acts as the input structure to the basal ganglia, neuroanatomical evidence shows that the basal ganglia sub-regions loop back on one another (Gurney, Prescott, and Redgrave, 2001a), and that in particular the striatum sends inhibitory signals to the substantia nigra (SN), which in turn projects back both excitatory and inhibitory signals via dopamine (D1 and D2 receptors respectively) to the striatum (Gerfen et al., 1990; Harsing Jr and Zigmond, 1997). There is therefore a potential mechanism for appropriate feedback to the hippocampal-striatal synapses in order to provide this error signalling, and an exploration of this error signal hypothesis could be a potentially interesting research endeavour.

5.4.4 Limitations of the Experiment

The experiments run here showed that the model, both with and without replays, is effective in its ability to learn appropriate actions that lead to the reward. However, testing has been restricted to the open arena and for a single reward location. It has not been tested for more complex environments where there might be obstacles or walls for instance, nor for cases in which the reward location might change. The model in its current form may be effective in the first case, but only for the standard learning rule, since causality between place and action would allow exact paths to be learned. This is true both for the non-replay and replay case. The modified learning rule however would likely fail for more complex environments, since it partially abandons direct causality in favour of average causality. That is, it takes an average over the recently chosen actions and uses that to learn the place-action associations. It is clear then that this form of place-action association would not account for the direct place-action associations needed to learn obstacle avoidance.

For cases in which the reward location might change, it may be possible to adopt the action selection mechanism described above. For example, if MiRo has not found the reward after a certain period of searching, a new behaviour overrides the memory-guided behaviour of the hippocampus – a random walk perhaps, or some other less random behaviour. The learning rule, via its error signal, could then update the hippocampal-striatal synaptic weights if or when MiRo finds the reward in a new location.

Chapter 6

Conclusions and Future Work

Whilst the literature on hippocampal reverse replay has speculated that it plays an important role in biological reinforcement learning, further experimental support is still required. Yet robots offer a unique testing opportunity for computational models, as they allow the realisation of those models for generating real-world behaviours. The work presented in this thesis has therefore asked whether, if a computational model of hippocampal reverse replay were to be embodied in a simulated robot, hippocampal reverse replays contribute to the reinforcement learning process. Specifically, this work has looked to test whether hippocampal reverse replays can improve learning in a classic reinforcement learning paradigm – the Morris water maze task.

The thesis began by first reviewing what is known regarding hippocampal neuroanatomy, showing that an important projection exists from the hippocampus to the striatum. Equally important is the finding that the CA3 region of the hippocampus shows significant recurrent connectivity. A hippocampal network based on this recurrent structure was then developed and embodied on a simulated version of the MiRo robot. Each of the cells in the hippocampal network represents place cells, encoding for specific regions of an environment and bidirectionally connected such that the connectivity represents a map of the environment. Via an implementation of short-term plasticity and intrinsic plasticity, this network was capable of reproducing fast reverse replays of the most recent spatiotemporal trajectories. With a model of hippocampal reverse replay at the ready, the next and final step was to couple it with a hippocampal-striatal model for testing in a reinforcement learning task, such that the striatal component comprised an ensemble of action cells.

6.1 Contributions of the Thesis

There are three contributions made in this thesis. The first is the minor addition to the literature review on hippocampal neuroanatomy (Chapter 2). That literature review was largely a condensed version of that given by Amaral and Lavenex (2007), but with one important and necessary extension: a discussion on the existence of projections from the hippocampus to the striatum. The model developed in Chapter 5 used this to assume projections from the hippocampal network to a striatal network.

The second is in the embodiment of a hippocampal reverse replay model in a simulated robot. Previous modelling attempts had not proven their applicability in real-time, robotics settings. Chapter 4 developed the first such model that performs in real-time for producing reverse replays of place cell activities. Robotic embodiment requires some necessary alterations to existing models of hippocampal reverse replay. Due to this, two very recent models of hippocampal reverse replay have been consolidated, showing the potentiality for both of their hypotheses in combination to be plausible.

The final, and most significant, contribution is in applying the reverse replay model for a robotic reinforcement learning task. Chapter 5 showed that when reverse replays are employed in a hippocampal-striatal network model, learning speed and robustness can be improved. This adds support to the hypothesis that hippocampal reverse replays are an important component in biological reinforcement learning, which is the primary motivation behind the work in this thesis.

6.2 Insights for Neuroscience and Robotics

How the biological nervous system generates the ability to control the body is a research problem for the neurosciences, and a more mathematically rigorous approach to hypothesising about the function of the nervous system is through computational modelling. By embodying a computational model of the brain in a robot, one can test that model in the real-world, and check whether it produces the hypothesised behaviour. But equally in the field of robotics, creating control systems for robotic bodies poses the same problems that the biological nervous system has seemingly solved. The embodiment of computational neuroscience models could therefore offer insights to both the neuroscience and robotics fields.

The study presented in this thesis has gone some way to support a current hypothesis made in the neuroscience literature: That hippocampal reverse replays support biological reinforcement learning. By augmenting a network model used for reinforcement learning with hippocampal reverse replays, the work developed here showed that hippocampal reverse replays enhance the reinforcement learning process in a simulated Morris water maze task. This existence proof of hippocampal reverse replay adds further support to the hypothesis stated above.

An additional hypothesis for biological reinforcement learning is that synaptic eligibility traces, combined with three-factor learning rules, might be necessary for learning over behavioural timescales (Vasilaki, Frémaux, et al., 2009; Gerstner et al., 2018). The time constants in these synaptic eligibility traces must be in the range of a few seconds, in order to capture the necessary behavioural history. The results here, however, have shown that it is not necessary to have large time constants when reverse replays are employed. Rather, reverse replays reinstate the behavioural history, by replaying both the place cells and the action cells, during which time synapses can undergo modifications according to the three-factor learning rule. Although, even with very small eligibility trace time constants, the model without reverse replays was still capable of learning to some extent, but it was markedly worse in performance than when reverse replays were employed. These results are similar to experimental findings, showing that disruption of awake hippocampal replay events (in the form of sharp-wave ripples) diminishes but does not eliminate goal-directed learning. This suggests that hippocampal reverse replays may enhance, rather than provide entirely, the necessary mechanism for learning over behavioural timescales, and that there is still a role for synaptic eligibility traces in learning.

There are particular challenges in reinforcement learning from a computer science perspective too, that this work could help advance. Whilst replaying states and actions is not new in reinforcement learning, doing so with continuous states and actions is. Yet having continuous states/actions is critical for robotic applications of reinforcement learning. The network model developed here then offers a method for replaying states/actions using neural and synaptic traces, which provides a method for replaying states/actions that are continuous. Whilst the main goal of this thesis was to help answer the neuroscience question on the role of hippocampal reverse replay, there is therefore some useful insights for robotic reinforcement learning.

However, this work has also presented new questions and research problems, which leaves open space for future work. We now turn to this in the next and final section of this thesis.

6.3 Scope for Future Research

In order to generate cell activity that was place specific in this model, the global x-y coordinates for the robot had to be used. But this is a biologically unrealistic property of place cell emergence. Questions remain therefore around how other models of the hippocampus, ones attempting to understand the emergence of place cells, grid cells, head direction cells, etc. (Sheynikhovich et al., 2009; Jauffret, Cuperlier, and Gaussier, 2015), could be consolidated with hippocampal replay models. Augmenting the hippocampal replay model developed here onto other hippocampal models of spatial navigation would be necessary step towards a more complete hippocampal model. In particular, it may be most desirable to embody these models in the physical MiRo robot, rather than the simulated MiRo. There are a number of challenges to overcome with this, beyond the usual engineering type difficulties. But the most prominent challenge relates to the point made here, regarding the generation of place cell activities. Although it could be simplified by using an external system for measuring MiRo's precise coordinate position, and using the same place cell computations (i.e. the normal distributions), it would be preferable, and more complete, to generate these using only MiRo's own sensing system. Employing the model in the physical MiRo robot would therefore provide more compelling evidence for the sufficiency of the model, displaying in particular that it can integrate with other hippocampal-inspired spatial navigation network models (see discussion from Section 3.3).

How the model performs in more complex environments, or in instances where the reward location changes, is still an open question, but it is worth examining here. The action selection mechanism discussed in Section 5.4.3 lends itself towards multiple types of behavioural implementations. This offers additional advantages beyond the original model of Vasilaki, Frémaux, et al. (2009), in which all actions were determined from the hippocampal network only. Recall that there is, in effect, an error signal used in the learning rule, which computes the difference between the action selected and the proposed action of the hippocampal network. This error signal is then used to update the hippocampal network if, and only if, the behaviour led to a non-zero reward. The hippocampal network is therefore used only for memory-guided behaviour. Should the memory-guided behaviour fail to lead to a reward, the action selection mechanism can select instead a new behavioural trajectory, allowing the hippocampal network to update if a new behaviour leads to a reward. This action selection mechanism could easily be expanded then, allowing for many different behaviours – obstacle avoidance in complex environments, for instance. A computational model of action selection in the basal ganglia (Gurney, Prescott, and Redgrave, 2001a; Gurney, Prescott, and Redgrave, 2001b) has previously been implemented in a robotic foraging task (Girard et al., 2003), which demonstrated six different behaviours in the action selection mechanism. However, the model did not include memory guided behaviours, and it could be interesting to extend it with the hippocampal model developed here for incorporating memory guided behaviours.

Finally, there is, of course, more to hippocampal replays than the reverse kind. Evidence suggests that forward replays, both during awake and sleep states, have perhaps a more important role in mnemonic functions (Girardeau et al., 2009; Ego-Stengel and Wilson, 2010). And forward replays have potentially a role in planning too (Ólafsdóttir, Bush, and Barry, 2018). It is not immediately clear how the model of replay presented here could support forward replays, particularly in the case where the replays occur some time after exploration. However, the results of the hippocampal replay model when intrinsic excitability is removed (Section 4.2.3) may provide a clue towards implementations. Those results showed that, since the environmental map is encoded in the network connectivity, replays propagated throughout all possible paths in the environment from the goal location. To initiate forward replays, therefore, one would need the trajectory encoded into the connectivity, and a cue (at the start location) for initiating the replay. Indeed, most other models of hippocampal replay (see Molter, Sato, and Yamaguchi, 2007; Haga and Fukai, 2018 for instance) store trajectory sequences in the connectivity. This leads therefore to the question of how the neural maps are generated, for which algorithms informally termed *Neural-SLAM* attempt to solve (see for instance Milford, Wyeth, and Prasser, 2004 for an example Neural-SLAM algorithm). Embedding hippocampal replay into these Neural-SLAM models could help better understand how replay could operate in both the forward and reverse directions, whilst the Neural-SLAM models could benefit from the enhanced learning capabilities of replays.

References

Amaral, David G, C. L. Dolorfo, and P. Alvarez-Royo (1991). “Organization of CA1 projections to the subiculum: A PHA-L analysis in the rat”. In: *Hippocampus* 1.4, pp. 415–435.

Amaral, David G, R. Insausti, and W. M. Cowan (1984). “The commissural connections of the monkey hippocampal formation”. In: *Journal of Comparative Neurology* 224.3, pp. 307–336.

Amaral, David G and P. Lavenex (2007). “Hippocampal Neuroanatomy”. In: *The Hippocampus Book*. Ed. by Per Andersen et al. Oxford: Oxford University Press. Chap. 3, pp. 37–114.

Amaral, David G and Menno P Witter (1989). “The three-dimensional organization of the hippocampal formation: A review of anatomical data”. In: *Neuroscience* 31.3, pp. 571–591.

Ambrose, R Ellen, Brad E Pfeiffer, and David J Foster (2016). “Reverse replay of hippocampal place cells is uniquely modulated by changing reward”. In: *Neuron* 91.5, pp. 1124–1136.

Andersen, P., T. V. P. Bliss, and K. K. Skrede (Aug. 1971). “Lamellar organization of hippocampal excitatory pathways”. In: *Experimental Brain Research* 13.2, pp. 222–238.

Ariav, Gal, Alon Polsky, and Jackie Schiller (2003). “Submillisecond precision of the input-output transformation function mediated by fast sodium dendritic spikes in basal dendrites of CA1 pyramidal neurons”. In: *Journal of Neuroscience* 23.21, pp. 7750–7758.

Aubin, Lise, Mehdi Khamassi, and Benoıt Girard (2018). “Prioritized sweeping neural DynaQ with multiple predecessors, and hippocampal replays”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 16–27.

Baisden, Ronald H, Michael L Woodruff, and Donald B Hoover (1984). “Cholinergic and non-cholinergic septo-hippocampal projections: a double-label horseradish peroxidase-acetylcholinesterase study in the rabbit”. In: *Brain research* 290.1, pp. 146–151.

Berger, T. W., S. Semple-Rowland, and J. L. Bassett (1981). “Hippocampal polymorph neurons are the cells of origin for ipsilateral association and commissural afferents to the dentate gyrus”. In: *Brain Research* 224.1, pp. 329–336.

Bi, Guo-qiang and Mu-ming Poo (1998). “Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type”. In: *Journal of neuroscience* 18.24, pp. 10464–10472.

Blackstad, T. W. (1956). “Commissural connections of the hippocampal region in the rat, with special reference to their mode of termination”. In: *Journal of Comparative Neurology* 105.3, pp. 417–537.

Boss, Barbara D, G. M. Peterson, and W. M. Cowan (1985). “On the number of neurons in the dentate gyrus of the rat”. In: *Brain Research* 338.1, pp. 144–150.

Boss, Barbara D, Kris Turlejski, et al. (1987). "On the numbers of neurons on fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats". In: *Brain research* 406.1-2, pp. 280–287.

Braun, Erin Kendall, G Elliott Wimmer, and Daphna Shohamy (2018). "Retroactive and graded prioritization of memory by reward". In: *Nature communications* 9.1, p. 4886.

Brea, Johanni et al. (2016). "Prospective coding by spiking neurons". In: *PLoS computational biology* 12.6.

Burwell, R. D. (2000). "The Parahippocampal Region: Corticocortical Connectivity". In: *Annals of the New York Academy of Sciences* 911.1, pp. 25–42.

Burwell, R. D. and David G Amaral (1998). "Perirhinal and postrhinal cortices of the rat: Interconnectivity and connections with the entorhinal cortex". In: *Journal of Comparative Neurology* 391.3, pp. 293–321.

Buzsáki, György (1989). "Two-stage model of memory trace formation: a role for "noisy" brain states". In: *Neuroscience* 31.3, pp. 551–570.

Buzsáki, György and David Tingley (2018). "Space and time: The hippocampus as a sequence generator". In: *Trends in cognitive sciences* 22.10, pp. 853–869.

Byrne, Patrick, Suzanna Becker, and Neil Burgess (2007). "Remembering the past and imagining the future: a neural model of spatial memory and imagery." In: *Psychological review* 114.2, p. 340.

Caballero-Bleda, Maria and Menno P Witter (1993). "Regional and laminar organization of projections from the presubiculum and parasubiculum to the entorhinal cortex: an anterograde tracing study in the rat". In: *Journal of Comparative Neurology* 328.1, pp. 115–129.

Cazé, Romain et al. (2018). "Hippocampal replays under the scrutiny of reinforcement learning models". In: *Journal of neurophysiology* 120.6, pp. 2877–2896.

Chapra, Steven C, Raymond P Canale, et al. (2010). *Numerical methods for engineers*. Boston: McGraw-Hill Higher Education,

Chenkov, Nikolay, Henning Sprekeler, and Richard Kempter (2017). "Memory replay in balanced recurrent networks". In: *PLoS computational biology* 13.1, e1005359.

Cherubini, Enrico and Richard Miles Miles (2015). "The CA3 region of the hippocampus: how is it? What is it for? How does it do it?" In: *Frontiers in cellular neuroscience* 9, p. 19.

Consequential Robotics (2019). *Documentation for the MiRo-E Robot*. URL: http://labs.consequentialrobotics.com/miro-e/docs/.

Cutsuridis, Vassilis and Michael E Hasselmo (2011). "Spatial memory sequence encoding and replay during modeled theta and ripple oscillations". In: *Cognitive Computation* 3.4, pp. 554–574.

Cutsuridis, Vassilis and Jiannis Taxidis (2013). "Deciphering the role of CA1 inhibitory circuits in sharp wave-ripple complexes". In: *Frontiers in systems neuroscience* 7, p. 13.

Dayan, Peter and Laurence F Abbott (2001). "Theoretical neuroscience: computational and mathematical modeling of neural systems". In:

Deshmukh, Sachin S and James J Knierim (2012). "Hippocampus". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 3.2, pp. 231–251.

Diba, Kamran and György Buzsáki (2007). "Forward and reverse hippocampal place-cell sequences during ripples". In: *Nature neuroscience* 10.10, p. 1241.

Dolleman-Van der Weel, Margriet J and Menno P Witter (2000). "Nucleus reuniens thalami innervates γ aminobutyric acid positive cells in hippocampal field CA1 of the rat". In: *Neuroscience letters* 278.3, pp. 145–148.

Dolorfo, C. L. and David G Amaral (1998b). “Entorhinal cortex of the rat: Organization of intrinsic connections”. In: *Journal of Comparative Neurology* 398.1, pp. 49–82.

— (1998a). “Entorhinal cortex of the rat: Topographic organization of the cells of origin of the perforant path projection to the dentate gyrus”. In: *Journal of Comparative Neurology* 398.1, pp. 25–48.

Ego-Stengel, Valérie and Matthew A Wilson (2010). “Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat”. In: *Hippocampus* 20.1, pp. 1–10.

Esposito, Umberto, Michele Giugliano, and Eleni Vasilaki (2015). “Adaptation of short-term plasticity parameters via error-driven learning may explain the correlation between activity-dependent synaptic properties, connectivity motifs and target specificity”. In: *Frontiers in computational neuroscience* 8, p. 175.

Foster, David J (2017). “Replay comes of age”. In: *Annual review of neuroscience* 40, pp. 581–602.

Foster, David J and Matthew A Wilson (2006). “Reverse replay of behavioural sequences in hippocampal place cells during the awake state”. In: *Nature* 440.7084, p. 680.

Frémaux, Nicolas and Wulfram Gerstner (2016). “Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules”. In: *Frontiers in neural circuits* 9, p. 85.

Freund, Tamás F and G. Buzsáki (1996). “Interneurons of the hippocampus”. In: *Hippocampus* 6.4, pp. 347–470.

Gaarskjaer, Frank B (1978). “Organization of the mossy fiber system of the rat studied in extended hippocampi. II. Experimental analysis of fiber distribution with silver impregnation methods”. In: *Journal of Comparative Neurology* 178.1, pp. 73–88.

Gauy, Marcelo Matheus et al. (2018). “A hippocampal model for behavioral time acquisition and fast bidirectional replay of spatio-temporal memory sequences”. In: *Frontiers in neuroscience* 12, p. 961.

Gerfen, Charles R et al. (1990). “D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons”. In: *Science* 250.4986, pp. 1429–1432.

Gerstner, Wulfram et al. (2018). “Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules”. In: *Frontiers in neural circuits* 12.

Girard, Benoıt et al. (2003). “A basal ganglia inspired model of action selection evaluated in a robotic survival task”. In: *Journal of integrative neuroscience* 2.02, pp. 179–200.

Girardeau, Gabrielle et al. (2009). “Selective suppression of hippocampal ripples impairs spatial memory”. In: *Nature neuroscience* 12.10, p. 1222.

Giri, Bapun et al. (2019). “Hippocampal reactivation extends for several hours following novel experience”. In: *Journal of Neuroscience* 39.5, pp. 866–875.

Gloveli, T. et al. (1997). “Morphological and electrophysiological characterization of layer III cells of the medial entorhinal cortex of the rat”. In: *Neuroscience* 77.3, pp. 629–648.

Gomperts, Stephen N, Fabian Kloosterman, and Matthew A Wilson (2015). “VTA neurons coordinate with the hippocampal reactivation of spatial experience”. In: *Elife* 4, e05360.

Gottlieb, D. I. and W. M. Cowan (1973). “Autoradiographic studies of the commissural and ipsilateral association connections of the hippocampus and dentate gyrus of the rat”. In: *Journal of Comparative Neurology* 149.4, pp. 393–421.

Greene, JRT and S Totterdell (1997). “Morphology and distribution of electrophysiologically defined classes of pyramidal and nonpyramidal neurons in rat ventral subiculum in vitro”. In: *Journal of Comparative Neurology* 380.3, pp. 395–408.

Grillner, Sten et al. (2005). “Mechanisms for selection of basic motor programs–roles for the striatum and pallidum”. In: *Trends in neurosciences* 28.7, pp. 364–370.

Groen, Thomas van and J Michael Wyss (1990). “The connections of presubiculum and parasubiculum in the rat”. In: *Brain research* 518.1-2, pp. 227–243.

Groenewegen, HJ et al. (1987). “Organization of the projections from the subiculum to the ventral striatum in the rat. A study using anterograde transport of Phaseolus vulgaris leucoagglutinin”. In: *Neuroscience* 23.1, pp. 103–120.

Gurney, Kevin, Tony J Prescott, and Peter Redgrave (2001a). “A computational model of action selection in the basal ganglia. I. A new functional anatomy”. In: *Biological cybernetics* 84.6, pp. 401–410.

— (2001b). “A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour”. In: *Biological cybernetics* 84.6, pp. 411–423.

Hafting, Torkel et al. (2005). “Microstructure of a spatial map in the entorhinal cortex”. In: *Nature* 436.7052, p. 801.

Haga, Tatsuya and Tomoki Fukai (2018). “Recurrent network model for learning goal-directed sequences through reverse replay”. In: *Elife* 7, e34171.

Hamam, B. N., David G Amaral, and A. Alonso (2002). “Morphological and electrophysiological characteristics of layer V neurons of the rat lateral entorhinal cortex”. In: *Journal of Comparative Neurology* 451.1, pp. 45–61.

Hamam, B. N., T. E. Kennedy, et al. (2000). “Morphological and electrophysiological characteristics of layer V neurons of the rat medial entorhinal cortex”. In: *Journal of Comparative Neurology* 418.4, pp. 457–472.

Hargreaves, E. L. et al. (2005). “Major dissociation between medial and lateral entorhinal input to dorsal hippocampus”. In: *Science* 308.5729, pp. 1792–1794.

Harris, Elana et al. (2001). “Intrinsic connectivity of the rat subiculum: I. Dendritic morphology and patterns of axonal arborization by pyramidal neurons”. In: *Journal of Comparative Neurology* 435.4, pp. 490–505.

Harsing Jr, LG and MJ Zigmond (1997). “Influence of dopamine on GABA release in striatum: evidence for D1–D2 interactions and non-synaptic influences”. In: *Neuroscience* 77.2, pp. 419–429.

Hasselmo, Michael E (2011). *How we remember: brain mechanisms of episodic memory*. MIT press.

Hasselmo, Michael E, Eric Schnell, and Edi Barkai (1995). “Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3”. In: *Journal of Neuroscience* 15.7, pp. 5249–5262.

Hebb, Donald Olding (1949). *The organization of behavior: a neuropsychological theory*. New York: Wiley.

Herkenham, Miles (1978). “The connections of the nucleus reuniens thalami: Evidence for a direct thalamo-hippocampal pathway in the rat”. In: *Journal of Comparative Neurology* 177.4, pp. 589–609.

Hjorth-Simonsen, A. and B. Jeune (1972). “Origin and termination of the hippocampal perforant path in the rat studied by silver impregnation”. In: *The Journal of Comparative Neurology* 144 (2).

Hollup, Stig A et al. (2001). “Accumulation of hippocampal place fields at the goal location in an annular watermaze task”. In: *Journal of Neuroscience* 21.5, pp. 1635–1644.

Hopfield, John J (1982). “Neural networks and physical systems with emergent collective computational abilities”. In: *Proceedings of the national academy of sciences* 79.8, pp. 2554–2558.

Humphries, Mark D and Tony J Prescott (2010). “The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward.” In: *Progress in neurobiology* 90.4, pp. 385–417.

Hyun, Jung Ho et al. (2015). “Kv1. 2 mediates heterosynaptic modulation of direct cortical synaptic inputs in CA3 pyramidal cells”. In: *The Journal of physiology* 593.16, pp. 3617–3643.

Ikemoto, Satoshi, Chen Yang, and Aaron Tan (2015). “Basal ganglia circuit loops, dopamine and motivation: a review and enquiry”. In: *Behavioural brain research* 290, pp. 17–31.

Insausti, R., M. T. Herrero, and Menno P Witter (1997). “Entorhinal cortex of the rat: Cytoarchitectonic subdivisions and the origin and distribution of cortical efferents”. In: *Hippocampus* 7.2, pp. 146–183.

Ishizuka, Norio, W. M. Cowan, and David G Amaral (1995). “A quantitative analysis of the dendritic organization of pyramidal cells in the rat hippocampus”. In: *Journal of Comparative Neurology* 362.1, pp. 17–45.

Ishizuka, Norio, Janet Weber, and David G Amaral (1990). “Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat”. In: *Journal of comparative neurology* 295.4, pp. 580–623.

Jadhav, Shantanu P et al. (2012). “Awake hippocampal sharp-wave ripples support spatial memory”. In: *Science* 336.6087, pp. 1454–1458.

Jahnke, Sven, Marc Timme, and Raoul-Martin Memmesheimer (2015). “A unified dynamic model for learning, replay, and sharp-wave/ripples”. In: *Journal of Neuroscience* 35.49, pp. 16236–16258.

Jauffret, Adrien, Nicolas Cuperlier, and Philippe Gaussier (2015). “From grid cells and visual place cells to multimodal place cell: a new robotic architecture”. In: *Frontiers in neurorobotics* 9, p. 1.

Johnson, Adam and A David Redish (2005). “Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model”. In: *Neural Networks* 18.9, pp. 1163–1171.

Kaitz, Suzan S and Richard T Robertson (1981). “Thalamic connections with limbic cortex. II. Corticothalamic projections”. In: *Journal of Comparative Neurology* 195.3, pp. 527–545.

Kametani, Hideki and Hiroshi Kawamura (1990). “Alterations in acetylcholine release in the rat hippocampus during sleep-wakefulness detected by intracerebral dialysis”. In: *Life sciences* 47.5, pp. 421–426.

Karimpanal, Thommen George and Roland Bouffanais (2018). “Experience replay using transition sequences”. In: *Frontiers in Neurorobotics* 12, p. 32.

Kirk, Ian J and Neil McNaughton (1991). “Supramammillary cell firing and hippocampal rhythmical slow activity.” In: *Neuroreport* 2.11, pp. 723–725.

Kiss, J et al. (2000). “The supramammillo-hippocampal and supramammillo-septal glutamatergic/aspartatergic projections in the rat: a combined [3H] D-aspartate autoradiographic and immunohistochemical study”. In: *Neuroscience* 97.4, pp. 657–669.

Klink, R. and A. Alonso (1997). “Morphological characteristics of layer II projection neurons in the rat medial entorhinal cortex”. In: *Hippocampus* 7.5, pp. 571–583.

Kobayashi, Tsuneyuki et al. (1997). “Task-dependent representations in rat hippocampal place neurons”. In: *Journal of Neurophysiology* 78.2, pp. 597–613.

Kobayashi, T et al. (2003). “Contribution of hippocampal place cell activity to learning and formation of goal-directed navigation in rats”. In: *Neuroscience* 117.4, pp. 1025–1035.

Kober, Jens, J Andrew Bagnell, and Jan Peters (2013). “Reinforcement learning in robotics: A survey”. In: *The International Journal of Robotics Research* 32.11, pp. 1238–1274.

Köhler, C. (1985). “Intrinsic projections of the retrohippocampal region in the rat brain. I. The subicular complex”. In: *Journal of Comparative Neurology* 236.4, pp. 504–522.

Köhler, Christer, Victoria Chan-Palay, and Jang-Yen Wu (1984). “Septal neurons containing glutamic acid decarboxylase immunoreactivity project to the hippocampal region in the rat brain”. In: *Anatomy and embryology* 169.1, pp. 41–44.

Kuutti, Sampo et al. (2020). “A Survey of Deep Learning Applications to Autonomous Vehicle Control”. In: *IEEE Transactions on Intelligent Transportation Systems*.

Le Duigou, Caroline et al. (2014). “Recurrent synapses and circuits in the CA3 region of the hippocampus: an associative network”. In: *Frontiers in cellular neuroscience* 7, p. 262.

Lever, Colin et al. (2009). “Boundary vector cells in the subiculum of the hippocampal formation”. In: *Journal of Neuroscience* 29.31, pp. 9771–9777.

Li, X-G et al. (1994). “The hippocampal CA3 network: an in vivo intracellular labeling study”. In: *Journal of comparative neurology* 339.2, pp. 181–208.

Loughlin, SE, SL Foote, and R Grzanna (1986). “Efferent projections of nucleus locus coeruleus: morphologic subpopulations have different efferent targets”. In: *Neuroscience* 18.2, pp. 307–319.

Maglóczky, Zsófia, László Acsády, and Tamás F Freund (1994). “Principal cells are the postsynaptic targets of supramammillary afferents in the hippocampus of the rat”. In: *Hippocampus* 4.3, pp. 322–334.

Mattar, Marcelo G and Nathaniel D Daw (2018). “Prioritized memory access explains planning and hippocampal replay”. In: *Nature Neuroscience* 21.11, p. 1609.

Meibach, Richard C and Allan Siegel (1977). “Efferent connections of the septal area in the rat: an analysis utilizing retrograde and anterograde transport methods”. In: *Brain research* 119.1, pp. 1–20.

Milford, Michael J, Gordon F Wyeth, and David Prasser (2004). “RatSLAM: a hippocampal model for simultaneous localization and mapping”. In: *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA’04. 2004*. Vol. 1. IEEE, pp. 403–408.

Mink, Jonathan W (1996). “The basal ganglia: focused selection and inhibition of competing motor programs”. In: *Progress in neurobiology* 50.4, pp. 381–425.

Mitchinson, Ben, M Pearson, et al. (2011). “Biomimetic robots as scientific models: a view from the whisker tip”. In: *Neuromorphic and brain-based robots*, pp. 23–57.

Mitchinson, Ben and Tony J Prescott (2016). “MIRO: a robot “Mammal” with a biomimetic brain-based control system”. In: *Conference on Biomimetic and Biohybrid Systems*. Springer, pp. 179–191.

Mnih, Volodymyr et al. (2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540, p. 529.

Molter, Colin, Naoyuki Sato, and Yoko Yamaguchi (2007). “Reactivation of behavioral activity during sharp waves: a computational model for two stage hippocampal dynamics”. In: *Hippocampus* 17.3, pp. 201–209.

Morris, Richard, Elizabeth Anderson, et al. (1986). “Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5”. In: *Nature* 319.6056, p. 774.

Morris, Richard, Paul Garrud, et al. (1982). “Place navigation impaired in rats with hippocampal lesions”. In: *Nature* 297.5868, p. 681.

Naber, Pieterke A, Fernando H Lopes da Silva, and Menno P Witter (2001). “Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum”. In: *Hippocampus* 11.2, pp. 99–104.

Nyakas, C et al. (1987). “Detailed projection patterns of septal and diagonal band efferents to the hippocampus in the rat with emphasis on innervation of CA1 and dentate gyrus”. In: *Brain research bulletin* 18.4, pp. 533–545.

O’Keefe, John (1976). “Place units in the hippocampus of the freely moving rat”. In: *Experimental neurology* 51.1, pp. 78–109.

O’Keefe, John and Jonathan Dostrovsky (1971a). “The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat.” In: *Brain research.*

— (1971b). “The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat.” In: *Brain research.*

O’keefe, John and Lynn Nadel (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

O’Keefe, John and Michael L Recce (1993). “Phase relationship between hippocampal place units and the EEG theta rhythm”. In: *Hippocampus* 3.3, pp. 317–330.

O’Mara, Shane M et al. (2001). “The subiculum: a review of form, physiology and function”. In: *Progress in neurobiology* 64.2, pp. 129–155.

Ólafsdóttir, H Freyja, Daniel Bush, and Caswell Barry (2018). “The role of hippocampal replay in memory and planning”. In: *Current Biology* 28.1, R37–R50.

Pang, Rich and Adrienne L Fairhall (2019). “Fast and flexible sequence induction in spiking neural networks via rapid excitability changes”. In: *eLife* 8, e44324.

Parikh, Vinay et al. (2007). “Prefrontal acetylcholine release controls cue detection on multiple timescales”. In: *Neuron* 56.1, pp. 141–154.

Pastalkova, Eva et al. (2008). “Internally generated cell assembly sequences in the rat hippocampus”. In: *Science* 321.5894, pp. 1322–1327.

Pennartz, CMA et al. (2004). “The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples”. In: *Journal of Neuroscience* 24.29, pp. 6446–6456.

Pikkarainen, Maria et al. (1999). “Projections from the lateral, basal, and accessory basal nuclei of the amygdala to the hippocampal formation in rat”. In: *Journal of Comparative Neurology* 403.2, pp. 229–260.

Pitkänen, Asla et al. (2000). “Reciprocal connections between the amygdala and the hippocampal formation, perirhinal cortex, and postrhinal cortex in rat: a review”. In: *Annals of the new York Academy of Sciences* 911.1, pp. 369–391.

Prescott, Tony J, Daniel Camilleri, et al. (2019). “Memory and mental time travel in humans and social robots”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1771.

Prescott, Tony J, Fernando M Montes González, et al. (2006). “A robot model of the basal ganglia: behavior and intrinsic processing”. In: *Neural networks* 19.1, pp. 31–61.

Prida, Liset Menendez de la et al. (2006). “The subiculum comes of age”. In: *Hippocampus* 16.11, pp. 916–923.

Pyapali, Gowri K et al. (1998). “Dendritic properties of hippocampal CA1 pyramidal neurons in the rat: intracellular staining in vivo and in vitro”. In: *Journal of Comparative Neurology* 391.3, pp. 335–352.

Rawlins, JNP, J Feldon, and JA Gray (1979). “Septo-hippocampal connections and the hippocampal theta rhythm”. In: *Experimental Brain Research* 37.1, pp. 49–63.

Redgrave, Peter, Tony J Prescott, and Kevin Gurney (1999). “The basal ganglia: a vertebrate solution to the selection problem?” In: *Neuroscience* 89.4, pp. 1009–1023.

Redgrave, Peter, N Vautrelle, et al. (2017). “Phasic Dopamine Signaling in Action Selection and Reinforcement Learning”. In: *Handbook of Behavioral Neuroscience*. Vol. 24. Elsevier, pp. 707–723.

Richmond, Paul et al. (2011). “Democratic population decisions result in robust policy-gradient learning: a parametric study with GPU simulations”. In: *PLoS one* 6.5, e18539.

Robertson, Richard T and Suzan S Kaitz (1981). “Thalamic connections with limbic cortex. I. Thalamocortical projections”. In: *Journal of Comparative Neurology* 195.3, pp. 501–525.

Rolls, Edmund T (2010). “A computational theory of episodic memory formation in the hippocampus”. In: *Behavioural brain research* 215.2, pp. 180–196.

Saravanan, Varun et al. (2015). “Transition between encoding and consolidation/replay dynamics via cholinergic modulation of CAN current: a modeling study”. In: *Hippocampus* 25.9, pp. 1052–1070.

Scatton, Bernard et al. (1980). “Origin of dopaminergic innervation of the rat hippocampal formation”. In: *Neuroscience letters* 18.2, pp. 125–131.

Schultz, Wolfram (1998). “Predictive reward signal of dopamine neurons”. In: *Journal of neurophysiology* 80.1, pp. 1–27.

Schwerdtfeger, Walter K (1979). “Direct efferent and afferent connections of the hippocampus with the neocortex in the marmoset monkey”. In: *American Journal of Anatomy* 156.1, pp. 77–82.

Scoville, William B. and Brenda Milner (1957). “Loss of recent memory after bilateral hippocampal lesions”. In: *Journal of neurology, neurosurgery, and psychiatry* 20.1, p. 11.

Shahan, Timothy A (2010). “Conditioned reinforcement and response strength”. In: *Journal of the Experimental Analysis of Behavior* 93.2, pp. 269–289.

Sheynikhovich, Denis et al. (2009). “Is there a geometric module for spatial orientation? Insights from a rodent navigation model.” In: *Psychological review* 116.3, p. 540.

Skaggs, William E and Bruce L McNaughton (1996). “Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience”. In: *Science* 271.5257, pp. 1870–1873.

Sloviter, R. S. and T. Lømo (2012). “Updating the Lamellar Hypothesis of Hippocampal Organization”. In: *Frontiers in Neural Circuits* 6.102.

Spruston, Nelson, Joachim Lübke, and Michael Frotscher (1997). “Interneurons in the stratum lucidum of the rat hippocampus: an anatomical and electrophysiological characterization”. In: *Journal of Comparative Neurology* 385.3, pp. 427–440.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Swanson, LW and W. M. Cowan (1977). “An autoradiographic study of the organization of the efferet connections of the hippocampal formation in the rat”. In: *Journal of Comparative Neurology* 172.1, pp. 49–84.

— (1975). “Hippocampo-hypothalamic connections: origin in subicular cortex, not ammon’s horn”. In: *Science* 189.4199, pp. 303–304.

Tamamaki, Nobuaki, Koutarou Abe, and Yoshiaki Nojyo (1987). “Columnar organization in the subiculum formed by axon branches originating from single CA1 pyramidal neurons in the rat hippocampus”. In: *Brain research* 412.1, pp. 156–160.

Taube, Jeffrey S (1998). “Head direction cells and the neurophysiological basis for a sense of direction”. In: *Progress in neurobiology* 55.3, pp. 225–256.

Taxidis, Jiannis et al. (2012). “Modeling sharp wave-ripple complexes through a CA3-CA1 network model with chemical synapses”. In: *Hippocampus* 22.5, pp. 995–1017.

Trouche, Stéphanie et al. (2019). "A Hippocampus-Accumbens Tripartite Neuronal Motif Guides Appetitive Memory in Space". In: *Cell*.

Tsodyks, Misha, Klaus Pawelzik, and Henry Markram (1998). "Neural networks with dynamic synapses". In: *Neural computation* 10.4, pp. 821–835.

Vandecasteele, Marie et al. (2014). "Optogenetic activation of septal cholinergic neurons suppresses sharp wave ripples and enhances theta oscillations in the hippocampus". In: *Proceedings of the National Academy of Sciences* 111.37, pp. 13535–13540.

Vanderwolf, Case H (1969). "Hippocampal electrical activity and voluntary movement in the rat". In: *Electroencephalography and clinical neurophysiology* 26.4, pp. 407–418.

Vasilaki, Eleni, Nicolas Frémaux, et al. (2009). "Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail". In: *PLoS computational biology* 5.12, e1000586.

Vasilaki, Eleni and Michele Giugliano (2014). "Emergence of connectivity motifs in networks of model neurons with short-and long-term plastic synapses". In: *PloS one* 9.1.

Verschure, Paul (2013). "Formal minds and biological brains ii: from the mirage of intelligence to a science and engineering of consciousness". In:

Vertes, Robert P, William J Fortin, and Alison M Crane (1999). "Projections of the median raphe nucleus in the rat". In: *Journal of Comparative Neurology* 407.4, pp. 555–582.

Verwer, Ronald WH et al. (1997). "Collateral projections from the rat hippocampal formation to the lateral and medial prefrontal cortex". In: *Hippocampus* 7.4, pp. 397–402.

Vouros, Avgoustinos et al. (2018). "A generalised framework for detailed classification of swimming paths inside the Morris Water Maze". In: *Scientific reports* 8.1, pp. 1–15.

Wainer, Bruce H et al. (1985). "Cholinergic and non-cholinergic septohippocampal pathways". In: *Neuroscience letters* 54.1, pp. 45–52.

Webb, Barbara (2001). "Can robots make good models of biological behaviour?" In: *Behavioral and Brain Sciences* 24.6, pp. 1033–1050.

West, M. J., L. Slomianka, and H. J. G. Gundersen (1991). "Unbiased stereological estimation of the total number of neurons in the subdivisions of the rat hippocampus using the optical fractionator". In: *The Anatomical Record* 231.4, pp. 482–497.

Whelan, Matthew T, Tony J Prescott, and Eleni Vasilaki (in press). "Fast Reverse Replays of Recent Spatiotemporal Trajectories in a Robotic Hippocampal Model". In: *Biomimetic and Biohybrid Systems. Living Machines 2020. Lecture Notes in Computer Science*. Springer.

Whelan, Matthew T, Eleni Vasilaki, and Tony J Prescott (2019). "Robots that Imagine – Can Hippocampal Replay Be Utilized for Robotic Mnemonics?" In: *Biomimetic and Biohybrid Systems. Living Machines 2019. Lecture Notes in Computer Science*. Springer, pp. 277–286.

Wilson, Matthew A and Bruce L McNaughton (1994). "Reactivation of hippocampal ensemble memories during sleep". In: *Science* 265.5172, pp. 676–679.

Witter, Menno P (1993). "Organization of the entorhinal—hippocampal system: A review of current anatomical data". In: *Hippocampus* 3.S1, pp. 33–44.

Witter, Menno P, Arjan W Griffioen, et al. (1988). "Entorhinal projections to the hippocampal CA1 region in the rat: an underestimated pathway". In: *Neuroscience letters* 85.2, pp. 193–198.

Witter, Menno P and Henk J Groenewegen (1990). "The subiculum: cytoarchitectonically a simple structure, but hodologically complex". In: *Progress in brain research*. Vol. 83. Elsevier, pp. 47–58.

Wyass, J Michael and Thomas Van Groen (1992). “Connections between the retrosplenial cortex and the hippocampal formation in the rat: a review”. In: *Hippocampus* 2.1, pp. 1–11.

Zhang, Wei and David J Linden (2003). “The other side of the engram: experience-driven changes in neuronal intrinsic excitability”. In: *Nature Reviews Neuroscience* 4.11, pp. 885–900.

Zhu, Henry et al. (2020). “The Ingredients of Real-World Robotic Reinforcement Learning”. In: *arXiv preprint arXiv:2004.12570*.

Appendices

A.1 Effects of Inhibition on Reverse Replays

To show that removing the inhibitory term in the equation for place cell dynamics, an example plot of a replay event is shown in Figures A.1-A.3. Notice in Figure A.3 particularly that there is little difference on the effects of either the trajectory or the replay events in either case.

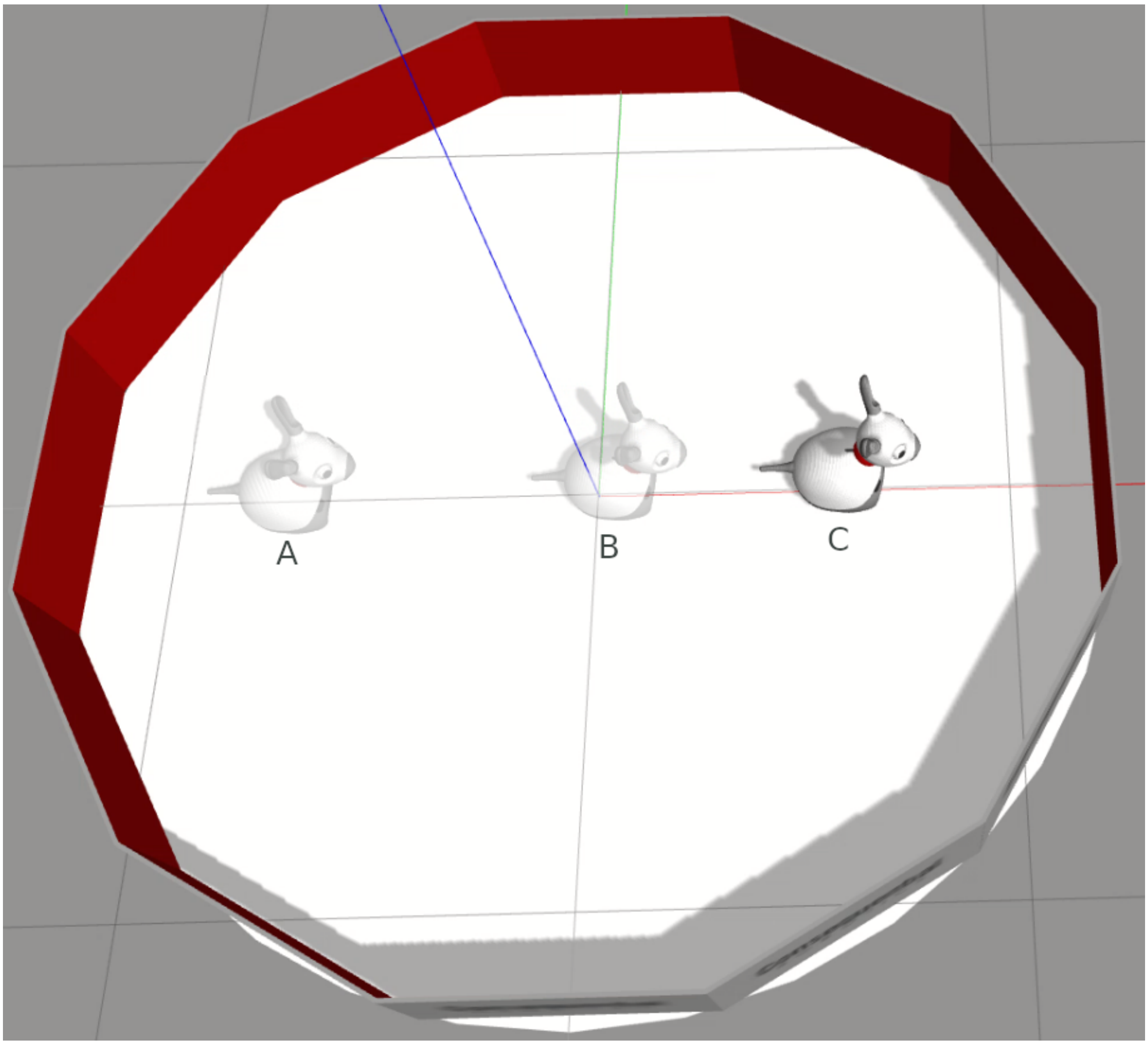


Fig. A.1 The example trajectory used to show there is no effect on reverse replays with or without the inhibitory term. MiRo begins in position A, passes through position B and ends in position C.

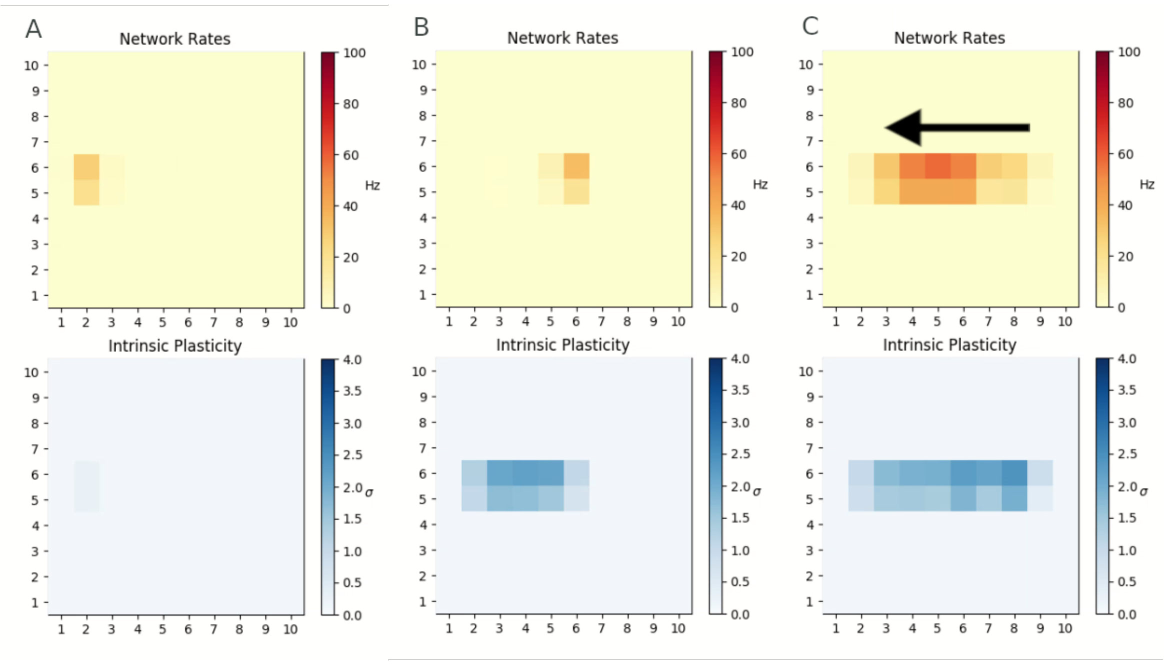


Fig. A.2 Plots of the place cell rates and intrinsic plasticities as MiRo passes through the points marked A, B and C in Figure A.1. Note that the activity in C is that of the reverse replay event, with the arrow indicating the temporal ordering of firing of the cells.

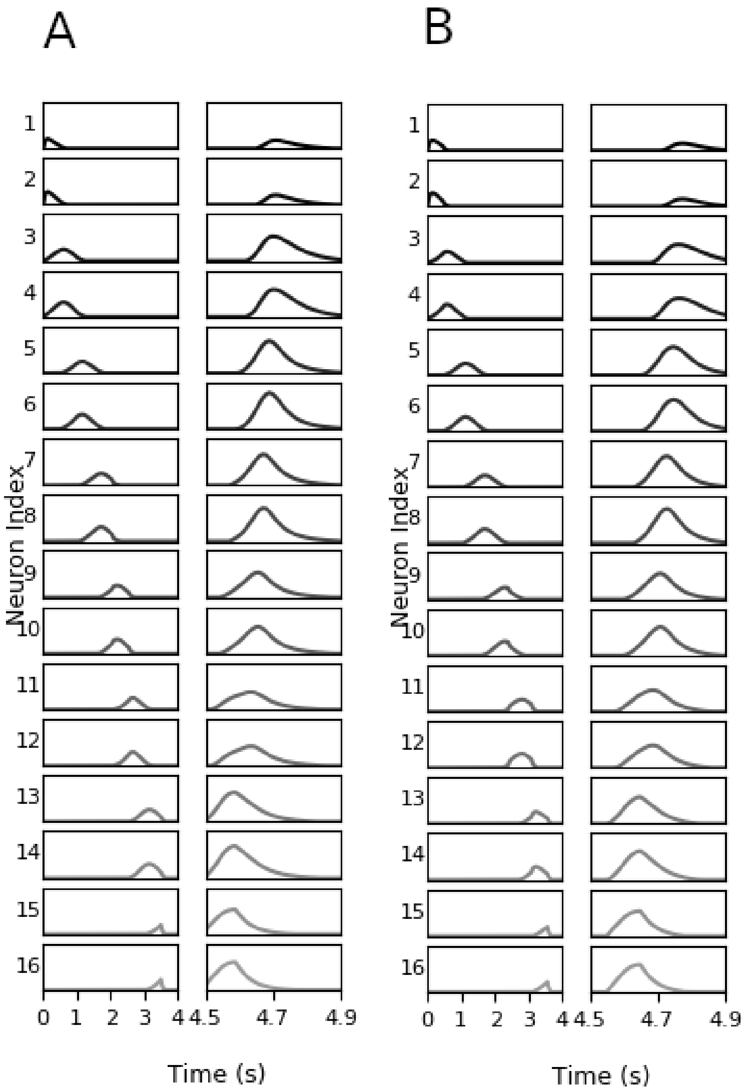


Fig. A.3 Line plots of the temporal ordering for the place cells that fired during the trajectory in Figure A.1. Plot (a) gives the rates without inhibition, whilst plot (b) is with inhibition. Left hand side plots in each is the activity during the trajectory, whilst the right hand side plots are for during a reverse replay.

A.2 Full Simulation Results

The experiments with and without replay in Chapter 5 were run over a wide range of values for τ_e and η. Plots of the average times to reward retrieval, averaged over 40 independent experiments, for all parameters are shown over the four figures shown here. As before, solid lines indicate the averages whilst shaded regions show one standard deviation.

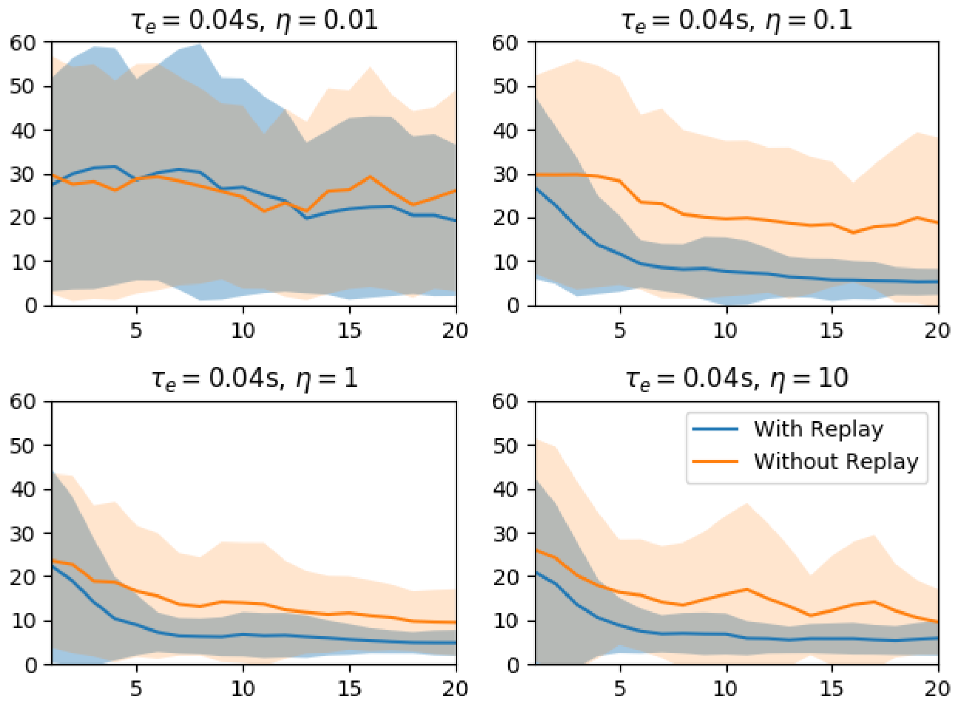


Fig. A.4 Full results for $\tau_e = 0.04s$

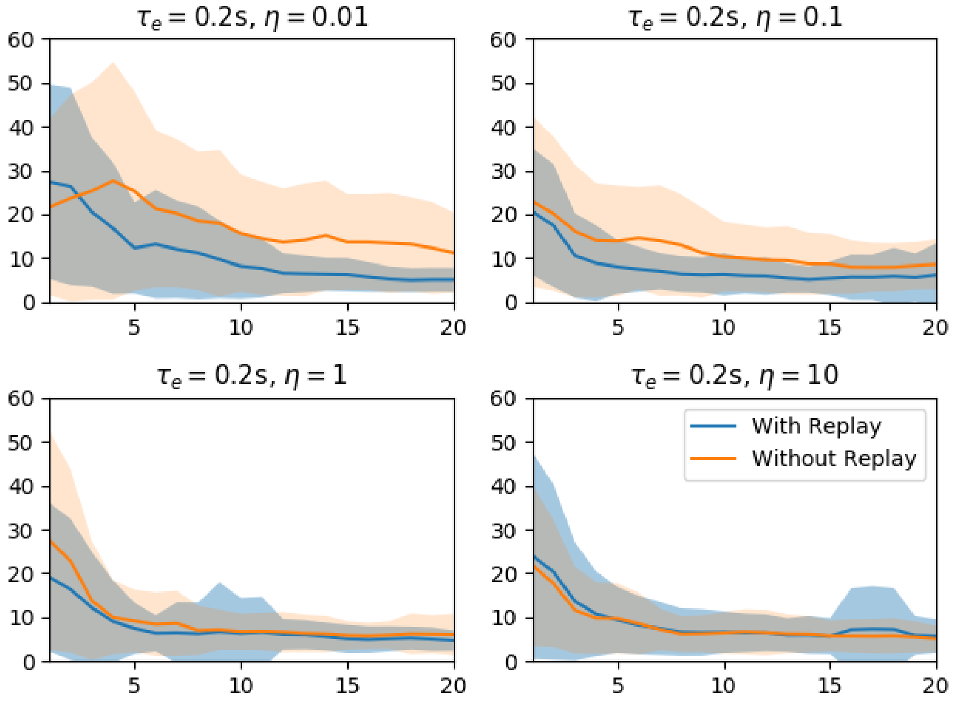


Fig. A.5 Full results for $\tau_e = 0.2s$

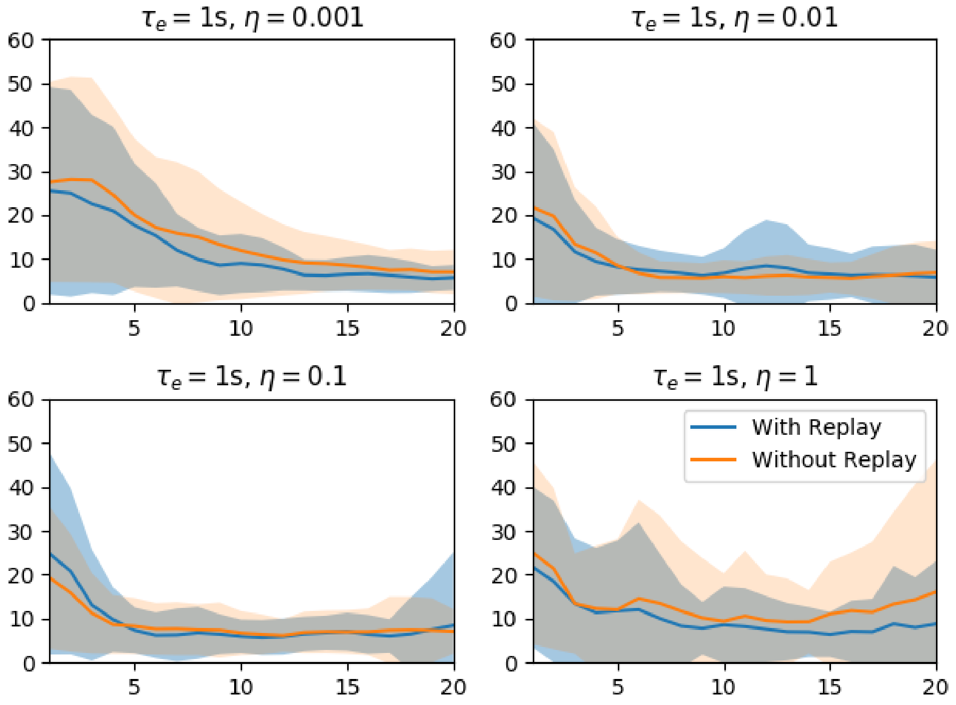


Fig. A.6 Full results for $\tau_e = 1s$

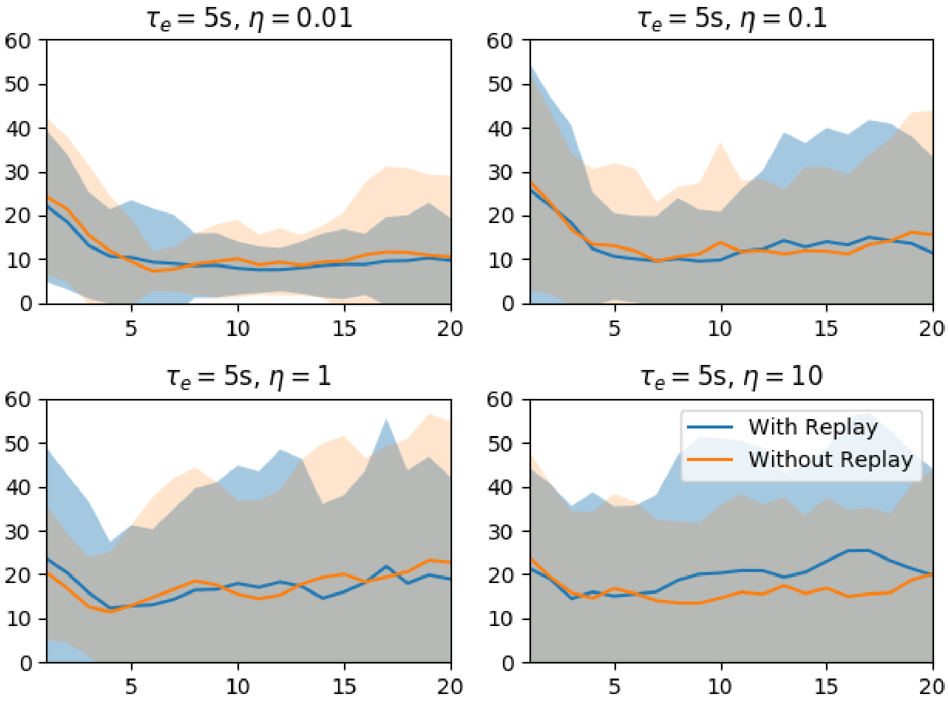


Fig. A.7 Full results for $\tau_e = 5s$

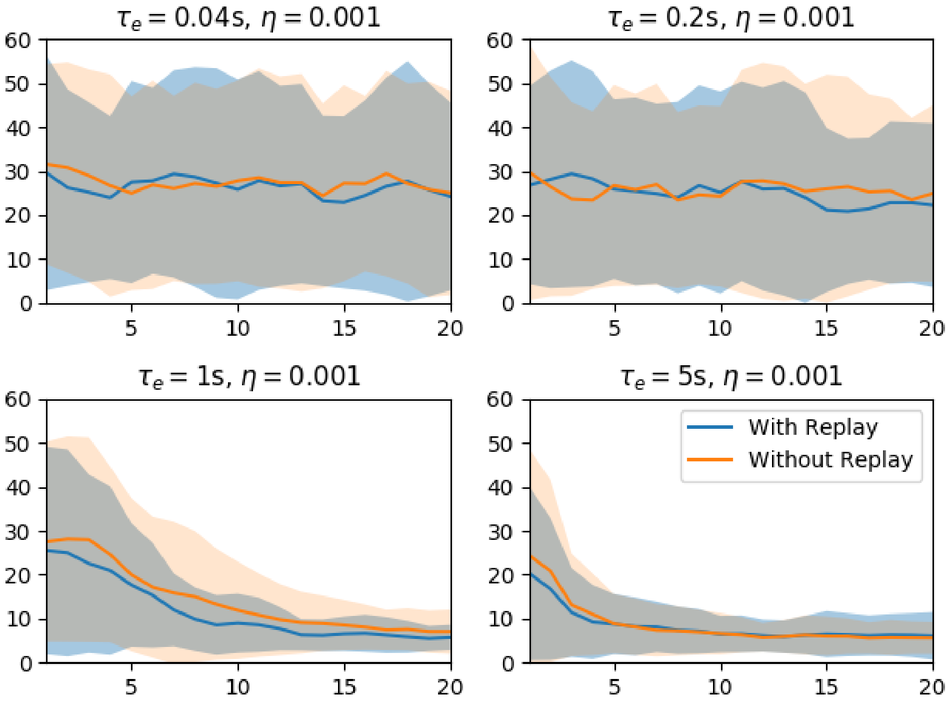


Fig. A.8 Full results for $\eta = 0.001$