LEARNING AND ACTING WITH PREDICTIVE COGNITIVE MAPS

by

ARTHUR WILLIAM JULIANI

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2020

DISSERTATION APPROVAL PAGE

Student: Arthur William Juliani

Title: Learning and Acting with Predictive Cognitive Maps

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Margaret Sereno	Chairperson
Dasa Zeithamova	Core Member
Thien Nguyen	Core Member
Richard Taylor	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2020

DISSERTATION ABSTRACT

Arthur William Juliani

Doctor of Philosophy

Department of Psychology

December 2020

Title: Learning and Acting with Predictive Cognitive Maps

Humans and other mammals possess two remarkable abilities: the capacity to store and retrieve a seemingly boundless series of episodic memories, and the capacity to quickly make sense of and navigate their changing environments. The latter has been described as a cognitive map, and along with the capacity to store and retrieve narrative memories, has been largely localized to the medial temporal lobe. Recent theorists have suggested that these two capacities are both aspects of a single unified system of 'experience construction.' In such a system, complex high-dimensional sensory experiences represented in the cortex are indexed by a low-dimensional representation within the medial temporal lobe. The dynamics of this representation then allow for the generation of coherent sequences of activation which correspond to coherent narrative experiences, as well as coherent trajectories through the environment, supporting both memory and navigation.

Such a theoretical perspective bears a strong resemblance to a recent class of deep neural networks called generative temporal models. In this work we explore this connection by introducing a series of increasingly complex generative temporal models, and analyzing each of their properties. We find that these models are able to learn representations which

bear a strong resemblance to known representations within the medial temporal lobe, such as place and time cells. Furthermore, we demonstrate that these representations are useful for rapidly learning to perform downstream goal-directed navigation tasks using biologically plausible reinforcement learning rules. We also examine the ways in which these models can be extended to display adaptation to changes in the structure or content of the environment, a key property of the cognitive map. Finally, we compare the behavior of artificial agents utilizing these learned representations to those of humans in a complex virtual navigation task. In doing so, we find evidence that humans utilize a hybrid behavioral strategy, and that such a strategy can be modeled by artificial agents utilizing a learned place cell like representation.

CURRICULUM VITAE

NAME OF AUTHOR: Arthur William Juliani

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene OR

North Carolina State University, Raleigh NC

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2020, University of Oregon

Master of Science, Psychology, 2015, University of Oregon

Bachelor of Arts, Psychology, 2013, North Carolina State University

AREAS OF SPECIAL INTEREST:

Cognitive Neuroscience

Machine Learning

PROFESSIONAL EXPERIENCE:

Senior Research Engineer, Unity Technologies, 2017-2020

Graduate Teaching Fellow, University of Oregon, 2014-2016

Data Science Intern, Duke University, 2013

GRANTS, AWARDS, AND HONORS:

Nvidia GPU Grant, Nvidia Corporation, 2016

PUBLICATIONS:

Juliani, A., Khalifa, A., Berges, V. P., Harper, J., Teng, E., Henry, H., Crespi, A., Togelius, J., & Lange, D. (2019). Obstacle tower: A generalization challenge in vision, control, and planning. International Joint Conferences on Artificial Intelligence 2019.

Juliani, A. W., Yaconelli, J. P., & Sereno, M. E. (2019). Learning to Integrate Egocentric and Allocentric Information using a Goal-directed Reward Signal. Journal of Vision, 19(10), 162-162.

Taylor, R. P., Juliani, A. W., Bies, A. J., Boydston, C., Spehar, B., & Sereno, M. E. (2018). The implications of fractal fluency for biophilic architecture. Journal of biourbanism, 6, 23-40.

Juliani, A. W., Bies, A. J., Boydston, C. R., Taylor, R. P., & Sereno, M. E. (2016). Navigation performance in virtual environments varies with fractal dimension of landscape. Journal of environmental psychology, 47, 155-165.
Juliani, A., Bies, A., Boydston, C., Taylor, R., & Sereno, M. (2016). Spatial localization accuracy varies with the fractal dimension of the environment. Journal of Vision, 16(12), 1370-1370.

Juliani, A., Leidheiser, W., McLaughlin, A., Allaire, J., & Gandy, M. (2013, September). Cognitive Ability Predicts Older Adult Performance in a Complex Task but is Moderated by Social Interaction. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 1740-1744).

ACKNOWLEDGMENTS

I want to acknowledge the support of my entire committee, and the faculty of the Department of Psychology as a whole. In particular, I wish to express gratitude to Margaret Sereno for her role as an advisor and source of constant support as my interests and career goals continued to develop throughout my time as a graduate student. I also wish to acknowledge the support of my professional colleagues, whose support it possible for me to complete this work.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The "above" is what is "on the ceiling," the "below" is what is "on the floor," the "behind" is what is "at the door." All these wheres are discovered and circumspectly interpreted on the paths and ways of everyday associations, they are not ascertained and catalogued by the observational measurement of space.

-Martin Heidegger, *Being and Time*

Humans and other mammals can quickly become familiar with and skillfully navigate new spaces. This is thought to be possible thanks to the existence of a mental representation of the space which we quickly generate and update unconsciously. Over half a century ago this idea was made more concrete with the proposal of a cognitive map of space in mammals (Tolman, 1948). This 'map' was demonstrated in rodents as one which is quickly learned from experience, conforms to the unique structure of a space, and is used by the animal to navigate that space. The following decades saw the discovery of place cells in the hippocampus, leading researchers to focus on this area as the site of the cognitive map (O'Keefe, 1976; O'Keefe & Nadel, 1978; Morris, Garrud, Rawlins, & O'Keefe, 1982).

Subsequent to the discovery of place cells was the discovery of a series of other spatially selective cells in the nearby regions of the hippocampus, collectively part of the medial temporal lobe. Most notable among these was the discovery of grid cells (Hafting, Fyhn, Molden, Moser, & Moser, 2005), which explicitly encode spatial information that corresponds to the position of an animal within an environment. Since then, there has

been the discovery of a variety of spatial-information-encoding cells and sub-regions within the hippocampal formation. These have been shown to encode a variety of different signals ranging from animal head orientation (Taube, Muller, & Ranck, 1990) to environment boundaries (Lever, Burton, Jeewajee, O'Keefe, & Burgess, 2009).

This spatial role of the hippocampus can be contrasted with the alternative perspective that the hippocampus is primarily involved in the formation, consolidation, and recall of episodic memories in animals, particularly in humans (Tulving & Markowitsch, 1998). Early lesion studies confirmed the essential role the hippocampus plays in ensuring that narrative experience enters long-term memory. Patients with hippocampal damage show a severely degraded ability to create new memories of personal experiences, a condition referred to as anterograde amnesia (Scoville & Milner, 1957; Aggleton & Brown, 1999). It has also been shown that patients with similar hippocampal damage are also unable to imagine new experiences with the same level of coherency as individuals without such damage (Hassabis, Kumaran, Vann, & Maguire, 2007), suggesting that the region is more generally involved in the construction of coherent narrative experiences (Hassabis & Maguire, 2009).

This encoding and decoding of coherent experiences in the hippocampus has been studied in much greater depth in rodents than in humans. This work has led to the discovery of replay, a phenomena characterized by trajectories of place cells corresponding to an environment spontaneously reactivating when the animal is at rest after having experienced that environment (Louie & Wilson, 2001; Foster & Wilson, 2006). These replay events have been shown to take place both during sleep and waking states, as well as to proceed in the "forward" and "reverse" directions. Studies have also found the existence of so-called preplay events, which take place prior to the animal experiencing a certain environment (Dragoi & Tonegawa, 2011, 2013). The value of these events to the animal has been theorized to be in their ability to both aid in the consolidation of memories as well as to support planning future behavior (Pezzulo, van der Meer, Lansink, & Pennartz, 2014).

Functional accounts of replay and preplay in the hippocampus point to a unified in-

terpretation of the role of the medial temporal lobe. Rather than performing both spatial navigation as well as memory storage and retrieval independently, the hippocampus can be interpreted as an experience construction system, as proposed by Hassabis and Maguire (2009). In this theory, the role of the hippocampus is to generate coherent sequences of activation which correspond to extended narrative experiences, or episodic memories. The fundamental building block in this system is the neural representation contained within the hippocampus. This representation has been proposed to serve as an index into a cortical state in the related hippocampal index theory (Teyler & DiScenna, 1986).

These indices take the form of place cell representations when the state space of interest is spatial (O'Keefe, 1976), and take other forms such as time cell representations (Eichenbaum, 2014), or event cell representations (Sun, Yang, Martin, & Tonegawa, 2020), when there are other relevant aspects of the environment required to form meaningful indices of experience. This hippocampal representation can then serve as the basis for a state space upon which behaviorally motivated learning can take place. In many cases this learning involves spatial navigation toward physical locations, and as such, the entire system appears to be a spatial navigation one. In cases where the behaviorally salient environment representation is non-spatial, then the state space induced within the hippocampus bears non-spatial properties (Behrens et al., 2018).

This interpretation of the medial temporal lobe as an experience construction system, one which indexes cortical experiences and learns their temporal dynamics, bears a strong resemblance to a recent class of neural networks referred to as generative temporal models. Like the proposed role of the MTL, these models also infer latent states from high-dimensional sensory streams, and learn to spontaneously generate coherent sequences of these latent states, which can then be decoded into high-dimensional sensory information. These models are also often then used to then guide goal-directed behavioral learning in artificial agents (Ha & Schmidhuber, 2018). The goal of this dissertation is to further clarify this connection, and explore its limit through the description and empirical evaluation

of a series of increasingly complex generative temporal models.

In the following text of the introduction, each of the findings discussed above will be expanded upon to provide a fuller picture of the current neurobiological and computational understanding of the hippocampal formation, and its role in the creation and support of cognitive maps. We will then introduce the main theme of this work, a class of neural networks referred to as generative temporal models, and discuss their connection with the medial temporal lobe and its cognitive mapping abilities.

The body of this text will then turn to the introduction and analysis of a series of increasingly complex generative temporal models which capture various aspects of the construction system of the medial temporal lobe. The second chapter will introduce a simple generative temporal model, and demonstrate the ability for this model to develop place and time-like cells within its latent representation. We will also demonstrate the ability for such a model to perform replay, and analyze the hidden representations of the model, showing that they display temporal community structure, a key aspect of the hippocampal representation (Schapiro, Turk-Browne, Norman, & Botvinick, 2016).

The third chapter of this work will then turn to utilizing the learned latent states of a generative temporal model for the purpose of goal driven navigation. We will explicitly utilize known reinforcement learning algorithms which have been connected with reward learning in the brain, specifically actor-critic and successor representations (Niv, 2009; Stachenfeld, Botvinick, & Gershman, 2017). Building on these methods, we introduce a novel reinforcement learning algorithm which learns more rapidly than previous related methods. Here we will also demonstrate the usefulness of the replay capabilities of a generative temporal model in guiding goal-directed learning.

In the fourth chapter we turn to the problem of content generalization, the ability to learn representations which are invariant to non-structural changes in sensory stimuli within an environment. Here we introduce a more complex generative temporal model which utilizes multiple latent states, as well as a storage and lookup mechanism for enabling episodic

memory. We demonstrate that this model is able to learn allocentric representations, in the form of place-like cells, directly from egocentric observations. We then validate this method on both allocentric 2D environments as well as egocentric 3D environments with various topographies.

In chapter five we then turn to the question of context generalization, the ability to learn representations which adapt to changes in the structure of the environment. Here we explore a number of approaches for augmenting a generative temporal model with a contextual representation. Along with two latent representations learned using a supervised learning signal, we introduce an additional model which learn an implicit contextual representation in an entirely unsupervised fashion. We draw a connection between this representation and the parahippocampal gyrus.

Finally, in chapter six we present a set of experiments in a novel realistic 3D virtual environment conducted both with human participants and with artificial agents. In both cases, the entity interacting with the environment is tasked with performing a goal-directed navigation task toward a hidden goal location within the environment. We use this task to test for the effect of environment complexity on human performance. In addition, a set of environment-change conditions are used to examine where human's behavior in this task can be classified on the spectrum between model-based and model-free decision making strategies. We find evidence for a hybrid strategy. We then demonstrate that an artificial agent using a latent state space from a generative temporal model learns a policy with a similar set of adaptation characteristics to that of humans performing the task.

I.1 Neuroscientific Evidence for Cognitive Maps

The hippocampal formation is a system of brain regions within the medial temporal lobe, containing the hippocampus, entorhinal cortex, and subiculum, among other connected regions. It has historically been implicated in two broad categories of cognitive function, the development of and access to episodic memories (Tulving & Markowitsch, 1998; Aggle-

ton & Brown, 1999), and the representation of a spatial cognitive map (O'Keefe & Nadel, 1978; Behrens et al., 2018). These functions were discovered in independent contexts, and originally existed as distinct lines of research. Part of the limbic system, the formation is also densely connected to other important areas such as the prefrontal cortex (Preston & Eichenbaum, 2013), implicating it in the process of high-level decision making (Tanji & Hoshi, 2001). This section will describe the lines of research around these two broad interpretations, and the empirical evidence for each, both from behavioral and neural data.

Early research into spatial learning in rodents suggested than rather than simply learning stimulus-response mappings, some animals are able to develop abstract representations of their environments, and use them for navigation (Tolman, 1948). In a set of classic experiments, Tolman showed that rodents were able to quickly take never before visited paths to regions of space associated with a known reward, suggesting that the rodents had developed an abstract representation of the environment they were able to utilize in the task. These abstract representations were referred to as a "cognitive map," because of their apparently spatial nature, and their specific application to navigation in the case of rodents.

Early research into the role of the hippocampal formation in mammalian cognition made clear the potential contribution of the brain region to the formation of this cognitive map (O'Keefe & Nadel, 1978). This was supported by the discovery of cells within the hippocampus which were robustly selective to an animal occupying a specific position in space (O'Keefe, 1976). The idea that this selectivity could be used to support a general-purpose map of an animal's location within the world, and thus be used for the selection of intelligent behavior has been built upon and continuously developed throughout the proceeding decades (for a review, see Behrens et al., 2018).

This development has been grounded in the gradual discovery of populations of cells within the hippocampal formation which are selective to different aspects of the environment within which an animal finds itself within. Most studied among these have been the place and grid cells (O'Keefe, 1976; Hafting et al., 2005), with a wealth of additional cell

types having been discovered as well (Solstad, Boccara, Kropff, Moser, & Moser, 2008; Behrens et al., 2018). Evidence from lesion studies suggest that the spatial information represented in these regions is critical for performing navigation in animals (Morris et al., 1982). This has led to a large amount of theoretical work attempting to provide computational models of both how these representations are learned, as well as how they could be used to aid in active navigation and memory for animals (Hasselmo, 2009; Erdem & Hasselmo, 2012; Bush, Barry, Manson, & Burgess, 2015).

I.1.1 Place, Grid, and Other Spatial Cells

Early evidence for the existence of a cognitive map in mammals came from experiments conducted in the 1970s by O'Keefe and collaborators (O'Keefe, 1976; O'Keefe & Nadel, 1978). This early work was conducted on rodents as they moved around in a small enclosed maze. During this movement recordings of cellular activation were collected via electrodes from the CA1 region of the hippocampus. Hundreds of cells in this region were monitored, and it was discovered that a large number of them preferentially responded to specific spatial locations within the maze.

Further experimentation suggested that while some of these activation patterns were the result of incidental features of the environment, a non-trivial number of them displayed robust activation despite various manipulations of the sensory and motor experience of the animal. This suggested that these cells in some way coded for an abstract notion of the "place" the animal found itself within. This sense of place was semi-invariant to incidental features of the environment such as lighting conditions. It was also found that there was no direct connection between the position of the animal within space and the position within the CA1 region of the cell which preferentially fired for that region of space. To the researchers at the time, the given responsiveness of a place cell seemed arbitrarily related to the spatial properties of the region of its affinity.

In subsequent decades, follow-up work was conducted to more rigorously determine

the firing properties of place cells (Muller, Kubie, & Ranck, 1987; Muller & Kubie, 1987). Using a video monitoring system, Muller and Kubie were able to characterize the statistical properties of the place cells, and their impact from changes in the environment. Most compelling was the discovery that the place fields were able to quickly remap their location of preference when a cue card serving as the primary landmark in the environment was rotated. The complete removal of the cue card resulted in only minor shifts in place field firing, suggesting that they were supported by more complex perceptual anchors than just the cue card. More recently research has been conducted which provides evidence for the existence of similar place-specific cell populations in the human hippocampus as well (Ekstrom et al., 2003).

In the years following the discovery of place cells, there remained an open question regarding how it was that the semi-invariant and non-uniform representation of the place cells was generated and sustained during navigation. It was hypothesized that there must be a more consistent underlying representation of space (possibly developed from pure ego-motion cues) that serves as a foundation for the more environment-specific place cell representation (O'Keefe & Nadel, 1978). This representation was finally discovered in rodents in the mid-2000s in a region not of the hippocampus proper, but in the entorhinal cortex, specifically this region was the medial entorhinal cortex (MEC) (Hafting et al., 2005). This population of cells with this highly uniform spatial firing pattern became known as the grid cells, named for their triangular tiled pattern of activation.

Unlike the place cell populations, within which each cell responded preferentially to just a single region of space, grid cells respond with periodic firing that resulted from the spatial position of the animal. As such, each cell displayed a "grid" of activation for a given environment, each with a uniquely offset phase. The specific periodicity and scale of these activation patterns were found to vary in a predictable manner across the region in the entorhinal cortex, with the spatial scale increasing with distance from the dorsal end of the entorhinal cortex. Most importantly this grid representation develops after the animal

has been placed into a new environment. They then remain invariant to manipulations of sensory features of the environment such as lighting or other visual cues. The fast and stable representation is critical for supporting a useful navigation-oriented representation of space. It was subsequently found that there are additional populations of grid cells deeper in the entorhinal cortex whose firing patterns are dependent on head-cell firing, supporting a bridge between purely head direction selective cells and position selective only grid cells (Sargolini et al., 2006).

More recent fMRI work has provided evidence that humans possess an analogous region of grid cells in their entorhinal cortex (Doeller, Barry, & Burgess, 2010) as well, suggesting that the region may be shared by most mammals, and not specific to rodents. In the study by Doeller and collaborators, participants were placed into an fMRI machine and asked to perform a foraging task in a virtual environment. The BOLD signal was then measured and analyzed in the entorhinal region as a function of the direction and speed of the participant's movement through the virtual space. The finding that this signal corresponded to the expected firing pattern from grid cell recordings from rodents provided the evidence for a similar system. This pattern of activity was one which synced with the expected six-fold symmetry found in grid cell firing patterns. A similar activation pattern was found in later work in which human participants were given an imagined navigation task (Horner, Bisby, Zotow, Bush, & Burgess, 2016). Due to the lack of spatial resolution of fMRI, it is difficult to determine the exact structure of activation at the cellular level in this region in humans, making it unclear whether there is simply an analogically similar pattern of activation or whether humans indeed possess individual cells with grid-like firing profile as rodents do.

In addition to place and grid cells, an array of other spatially selective cells have been discovered within the hippocampal formation, including border and head-direction cells (for a review of additional spatially selective cell types, see Behrens et al., 2018). Evidence for these were discovered using largely similar methods to those originally used by O'Keefe

years earlier (O'Keefe, 1976), with single-unit recording from rodents within an artificially constructed environment primarily being the method of choice.

Head-direction cells were identified in the early 1990s, and as their name suggests, they consist of a population of cells in the subiculum of rodents which preferentially responded to the animal's head facing a specific direction in space (Taube et al., 1990). Similar to place cells, these cells were found to be robust to other environmental stimuli which were non-essential for determining the primary feature of activation: the direction of the animal's head. Each head-direction cell fires rapidly when the head is oriented in a specific direction, and maintains a low baseline level of firing otherwise. Also similar to place cells, but unlike grid cells, there is no topographical organization of the cells within the brain region that corresponds to firing preference in head direction space.

Another cell population with specific spatial firing features in the hippocampal formation are the border cells, which preferentially respond to the animal's proximity to a boundary in the environment, with greater activation as the animal gets closer to the preferred boundary for the cell (Solstad et al., 2008). This cell type was later generalized into a "Boundary Vector Cell," found in the subiculum (Lever et al., 2009). These boundary vector cells responded to proximity to border regardless of the animal's orientation or head direction, and maintained firing even when the animal was not necessarily in proximity to the boundary. This suggests that the cells could be used to compute distance to a given boundary, rather than simply providing a binary signal reflecting the presence or absence of a proximal boundary. It has been hypothesized that these boundary cells may be used to determine the limits of a given environment for the animal for the purpose of aligning grid cell responses.

Taken together the cell types discussed above seem sufficient for an understanding of the hippocampal formation as a purely spatial mapping system. Indeed, as will be discussed in Section I.2, a large amount of theoretical work has been done to demonstrate the sufficiency of these cell types for navigation. In more recent years however, more sophis-

ticated recording techniques and experimental designs have shed doubt on the concept of the hippocampus exclusively as a representation system for space.

I.1.2 Time, Event, and Other Non-spatial Cells

The picture of the hippocampus as a spatial cognitive map has been complicated in recent years by a variety of findings showing that in addition to cells which fire based on spatial features of the environment (such as place, grid, and border cells), there are additional cells which fire regularly according to non-spatial aspects of the environment. One of the more prominent of these is a class of cells which fire based on the elapsed time within a specific task, referred to as "time cells" (for a review, see Eichenbaum, 2014). Early evidence for this was put forward by Pastalkova and colleagues who showed that activation patterns in rodent hippocampus reflect internally generated sequences which corresponded to delay in the task rather than spatial position or other physical stimuli (Pastalkova, Itskov, Amarasingham, & Buzsáki, 2008). The existence of cells with this firing profile suggest that the activation patterns of the hippocampus reflects more than just a spatial selectivity, and points to a more general organizing principle behind these representations.

Subsequent research showed similar results in the case where the animal was stationary as well (MacDonald, Lepage, Eden, & Eichenbaum, 2011), and were isolated to be specifically anchored around task-specific temporal delays. Their work consisted of examining activation patterns in CA1 of the rodent hippocampus. The task involved the animal moving through a circular line maze. At the beginning of the maze, the animal was presented with one of two colored objects. In the next phase the animal remained in a fixed position in the maze for ten seconds. In the final phase the animal was then presented with an odor at the end of the maze. As expected, the researchers found place cells which were sensitive to the animal's spatial location within the maze. In addition, they also found cells which were sensitive to the temporal delay in the second phase of the task. Importantly, these patterns could not be explained by the animal's position, rotation, or velocity. When the

delay was extended in the second phase, the cells "remapped," with a different set of cells now corresponding to time cells for the task.

Surprisingly, the same cells which display an activation profile consistent with time cells sometimes also display place cell like activation in the work of MacDonald et al. (2011), suggesting that the simplistic narrative of place cells supporting spatial representation only is at best missing critical aspects related to temporal coding. If the hippocampus does not provide a spatial cognitive map, then what could be a more appropriate alternative? The evidence provided above suggests that the hippocampus represents experiences in both a spatial and temporal manner, but specifically one which is environment specific, in which neither a spatial or temporal component is dominant, and rather the needs of the task and environment are captured in the place cells.

Indeed, recent work looking at human hippocampal activation using fMRI shows that a spatio-temporal signal rather than a spatial or temporal one provided the best fit for the representation in the region (Deuker, Bellmund, Schröder, & Doeller, 2016). This was done using a task where participants navigated a virtual environment in which the spatial and temporal distances between objects in the environment were manipulated. The researchers referred to the joint representation learned as an "event map" to capture it's more abstracted nature.

More recently work in rodents has demonstrated the existence of specific cells which do not respond to either temporal or spatial properties of a task per-se, but rather to a more complex and general relationship between the animal and its environment (Sun et al., 2020). In this work, Sun and colleagues recorded from the rodent hippocampus while the animals performed a navigational task around a series of circular tracks. The shape and length of these tracks differed, such that there was no specific temporal or spatial correspondence between turning the left corner of one track and turning the left corner of another. Despite this, cells within hippocampus reliably responded to such events as turning a specific corner, suggesting that these cells encoded a notion of "event" rather than time

or place.

Taken together, this evidence suggests that cells in the hippocampus are sensitive to the task and environmental structure rather than to exclusively the spatial or temporal properties of the environment itself. As such, cells in the hippocampus are remapped as necessary to provide a state space which contains a coherent representation of the structure of any given task. We can then interpret the development of spatially and temporally specific cells as just one instantiation of a more general representation of the abstract state structure of a task or environment. The development and utilization of these abstracted state representations will be a major focus of the subsequent chapters of this work.

I.1.3 Replay, Preplay, and Structured Temporal Sequences

The construction system hypothesis states that the hippocampus serves to support both the consolidation and recall of past memories, but also the imagination of novel experiences (Hassabis & Maguire, 2009). One key element of both cases is their temporally extended and coherent nature. As such, rather than studying place cell activity in isolation at a given moment, we would expect it to follow a temporally extended pattern of activation, or at least a representation which is amenable to temporal extension. Such evidence has indeed been discovered. Early evidence suggested that concurrent pairs of place cells that were activated beforehand in a reward-driven task were reactivated together at a frequency greater than chance when in slow wave sleep (Wilson & McNaughton, 1994). Due to lack of technical sophistication in the recording equipment at the time, only correlations between pairs of cells could be shown in the rodents studied.

Subsequent research was able to not only verify this early finding, but extend it to long sequences of place cell firing patterns which lasted over dozens of seconds (A. K. Lee & Wilson, 2002). In their work, Lee and colleagues showed that place cell firing corresponding to a full trajectory of the animal through a linear maze were reactivated during short wave sleep. Significantly, these patterns of activation were not present during sleep

before the animal had experienced the linear maze, meaning that they were the result of experience running through the maze. Concurrent work also demonstrated that that these "replay" events also took place during REM sleep in rodents (Louie & Wilson, 2001). The explanations provided for these replay events during sleep is one of memory consolidation, with one prominent theory suggesting that replay of the event allows it to be bound to neocortical areas for long-term retention (Marr, Willshaw, & McNaughton, 1991; Nyberg, Habib, McIntosh, & Tulving, 2000).

In addition to being found in sleeping animals, replay has also been demonstrated in awake animals after some level of exposure to an environment (Foster & Wilson, 2006). Unlike earlier studies on sleep-based replay, in which the sequence of place cell activation was in the same order both during the experience and during replay, Foster and Wilson found that awake replay took place in the reverse order. Because of this, the phenomenon was aptly named reverse-replay. Similar to earlier experiments done on sleeping rodents, they utilized a linear maze in which the animal ran back and forth. Reverse-replay activation took place while the animal was at rest at an end of the maze, with the place cells firing from the unit correspond to the animal's current location backwards.

The phenomenon of replay and reverse replay has been further generalized by studies showing that awake animals experience both types of replay, dependent on their position within an environment (Diba & Buzsáki, 2007). Diba and Buzsaki found that once an animal had experienced a linear maze, the characteristic reverse replay was detectable while the animal was at rest at the end of the maze after running through it. In addition to this, a forward replay of the place cell sequence was found when the animal was at the beginning of the maze. These findings suggest that the replay activity is in relationship to the animal's context within the environment, and that replay activity during the awake state emanates outward from the animal along the previously experienced trajectory.

The results presented above all described the replay of events which span the order of a few seconds when originally experienced, and correspond to only one to two meter long

trajectories. The experiences of animals (including humans) in the wild however typically involve much longer sequences of movement, often spanning up to miles in the cases of long-distance runners. Davidson and colleagues were able to show that replay occurs in rodents at scales an order of magnitude larger than earlier work (Davidson, Kloosterman, & Wilson, 2009). While replay trajectories have typically been observed to be confined to the duration of a sharp wave ripple event in the hippocampus, they found that these extended replay events took place over the course of multiple such ripples, with each ripple corresponding to a sub-section of the full trajectory being replayed. This decomposition of long sequences into shorter sub-sequences provides a mechanism for the potential temporal abstraction of movements into a coarser temporal and spatial scale.

The picture of awake replay has been made richer by results showing that waking replay events need not be linked to the environment the animal is currently within (Karlsson & Frank, 2009). In a set of experiments conducted by Karlsson and Frank, it was shown that replay events corresponding to a previously experienced environment took place while an animal was resting in a second environment. These replay events were referred to as "remote replay," because the animal is no longer in spatial (or temporal) proximity to the original environment which they corresponded to. They found that the activation of replay events for the two environments were independent of one another, with a separate association process responsible for the local activity in the current environment enabling the remote replay.

Everyday subjective experience of recalling memories suggests that a kind of replay should be evident in humans as well. Similar to difficulties in showing the existence of grid cells however, the lack of spatial (EEG) or temporal (fMRI) resolution makes it difficult to isolate individual replay events in the human hippocampus. Nevertheless, research has been able to demonstrate that it is possible to decode the identity of individual experience trajectories in humans from the human hippocampus during replay at a frequency greater than chance using fMRI (Chadwick, Hassabis, Weiskopf, & Maguire, 2010). Similar work

has also isolated the specific contribution of the hippocampal region to reconstruction of episodic memories, specifically when a temporally structured experience is being recalled (Lehn et al., 2009).

There has been debate over the functional role of each of these kinds of replays in their various contexts (for a reivew, see Foster, 2017), with the predominant understanding revolving around memory consolidation. This is particularly true for the role of replay during sleep, where there is a history of evidence for brain-wide synaptic changes that would support such a mechanism (Bliss & Collingridge, 1993). Replay during the awake state is more complicated, as it could be seen to potentially interfere with the present experience, unless it is significantly delayed from the animal taking any action. One proposed theory is that the reverse replays experienced after movement through a maze serve to quickly propagate backward state information. In the case where the final state is rewarding for the animal, a kind of value iteration process may take place where the value at the final state is propagated backward to earlier states (Foster, Morris, & Dayan, 2000).

In the case of forward replay prior to animal action, the activation has been hypothesized to serve a planning-like function. Additional evidence for this comes from the fact that when there is more than one possible path available to the animal, multiple forward replays will correspond to different possible paths (Pfeiffer & Foster, 2013). Combined with a mechanism for evaluating future states, this would serve as a simple tree-search like planning method. This predictive perspective on replay will be expanded upon in the following section.

It is of value to discuss the hypothesized existence of one additional form of place cell sequence activation in animals, "preplay." Like replay, preplay involves the activation of a sequence of place cells which correspond to the movement through a physical environment. Unlike replay however, which is conditioned on the animal actually having moved through that space before, preplay takes place before the animal has experienced the environment (Dragoi & Tonegawa, 2011, 2013). Dragoi and Tonegawa propose that the existence of

preplay shows that place cell sequences are already connected together prior to experience, and the experience simply binds this sequence to a set of actual stimuli.

The existence of preplay is somewhat controversial however. There have been unsuccessful attempts to replicate the findings of Dragoi and Tonegawa using highly-sensitive recording equipment and more rigorous statistical techniques, calling into question the original finding (Silva, Feng, & Foster, 2015). The argument has also been made that the existence of coherent activation sequences of place cells before experience would render their activation as replays afterwards incomprehensible as reflection of any sort of memory or learning (Foster, 2017). The argument for experience-dependence in place cell sequence activation opens up the possibility for a kind of spectrum of reactivation between experience-less preplay and the replay of only trajectories explicitly experienced by the animal. Indeed, evidence for this comes from experiments showing that rodents experienced replay events for never-taken trajectories along a maze (Gupta, van der Meer, Touretzky, & Redish, 2010).

More recent work has shown that preplay events take place not only sequentially, but often in a cyclic fashion (Kay et al., 2020), with theta wave activity rapidly transitioning between encoding multiple sets of possible future trajectories in rodents. Such activation patterns would make possible much more efficient exploration of future possible behaviors in any given environment, since rather than serially imagining trajectories, they could be explored in near-parallel, in a method with similarities to how modern implementations of the monte-carlo tree search algorithm functions (Silver et al., 2016).

The structure of these temporal sequences has also been the object of study for many. Given the apparently arbitrary nature of remapping, the question might naturally arise as to why specific sequences of place cells seem to activate together at all, specifically within conditions of seemingly pure preplay. A review of some of these ideas was recently presented by Dragoi (2020). One recent insightful finding has been the discovery that as a rule place cell sequences seem to be made up of repeating motifs, consisting mostly of three

units (Liu, Sibille, & Dragoi, 2018). These motifs form a kind of grammar regarding the activation of place cell sequences, with larger sequences being made up of these motifs, but the structure within the motif being largely immutable.

With all of these computational possibilities opened up by the wealth of cell types and their temporal dynamics, we can now turn to the attempts made thus far to derive and validate concrete computational principles by which all of these make possible a cognitive map.

I.2 Computational Theories of Mammalian Navigation

Taken together, it would seem that boundary vector, head-direction, and grid cells provide a relatively extensive representation of the spatial situation of an animal at a given time. Unlike place cells, whose firing patterns are environment-specific, these three populations are all relatively invariant to disruptions in the environment, suggesting that when combined with environment specific perceptual cues, they could serve to provide a foundation for the computation of the local-specific activation patterns found in place cells. Since the discovery of these cell types, a large body of theoretical work has been undertaken to attempt to provide biologically valid computational models for how the development and maintenance of these representations could take place (for examples, see Samsonovich & McNaughton, 1997; Burgess, Barry, & O'keefe, 2007; Hasselmo, 2009; Erdem & Hasselmo, 2012; Bush et al., 2015). While not exhaustive, this section will provide a survey of some of the more influential of such theories and models, and how they have been used to reason about the mechanisms by which animals with these representations would be able to efficiently navigate the environments they find themselves within.

I.2.1 Path Integration, Attractors, and Other Early Models

One potential solution to the problem of spatial navigation in animals is path integration (Mittelstaedt & Mittelstaedt, 1980) (for a review, see McNaughton, Battaglia, Jensen,

Moser, & Moser, 2006). In path integration an animal utilizes self-motion cues to keep track of its location relative to a global starting position. This ability has also been referred to as "dead reckoning," with the implication being the ability of an animal or person to "reckon" about their location in the absence of external sensory cues. Beyond being a skill that appears intuitive for humans familiar with navigating their worlds, the system of representation provided by the hippocampal formation provides all of the components necessary for such a path integration system. More specifically, internal vestibular information generated by the animal's movement can drive the firing of head-direction cells, providing an oriented path signal with direction and velocity of movement (McNaughton, Chen, & Markus, 1991). Given a starting position, this incremental signal from head-direction cells can then be integrated together to produce a representation of the animal's updated location after movement.

The idea that such a system could be used for the maintenance of a cognitive map with place cell like activity was expanded upon in the late 1990s by Samsonovich and McNaughton, in a model which was able to mimic the recurrent structure of the hippocampus using an attractor network to generate units with place cell like activation patterns (Samsonovich & McNaughton, 1997). This model was referred to as a map-based path integrator, which consisted of an attractor network with activation layers imposed onto a 2D plane, thus providing an activation pattern with similar periodic firing as that of the grid cells. This representation was fed by a set of sensory inputs, which activated head and motion detectors, then feeding into the series of attractor maps, referred to as "charts," This process ultimately produces activation patterns in the final layer of cells which are reminiscent of place cell firing profiles. While this model predated the discovery of grid cells, the use of multiple 2D charts with semi-periodic firing foreshadowed the eventual discovery and incorporation of grid cells into subsequent models.

With the discovery of grid cells, models of navigation and path integration followed which explicitly incorporated this population of cells into the model. These fell into two

broad categories: models which utilized the attractor dynamics as described in (Samsonovich & McNaughton, 1997), such as (McNaughton et al., 2006), and models which instead utilized an oscillatory interference pattern to generate grid and place cell representations (Burgess et al., 2007; Hasselmo, 2009; Erdem & Hasselmo, 2012; Bush et al., 2015). In all cases the spatial structure of the representation chosen for these models was of importance. In order for the finite capacity of a fixed population of cells to represent a large and varied amount of space, a looping representational space has typically been employed. In the case of a set of cells with a 1D representation such as the head-direction cells, a ring is the representation of choice, with the torus being the 2D extension used to model populations of grid cells.

This latter set of models use as a foundation the theta-wave oscillations within the hippocampal formation. The offset phases of these signals in different populations of cells can then be used to generate an interference pattern. If projected onto a 2D plane, this interference pattern reflects the periodic firing pattern found in grid cells (Burgess et al., 2007). Place cell firing then corresponds to the conjunction of specific spatial information with a grid cell firing pattern. This basic interference model has been extended to simulate the storage and retrieval of experienced trajectories by an animal (Hasselmo, 2009), as well as goal-directed navigation in rodents (Erdem & Hasselmo, 2012).

In the work of Erdem and Hasselmo (2012), the interference model of place cell development and firing is combined with a basic model of the prefrontal cortex in which specific place cells, assumed to correspond to specific spatial states of the animal, are correlated with a goal provided by the prefrontal cortex. This model then uses a basic form of linear lookahead to plan out a path to the desired goal. While not explicitly mentioned in the original work, this lookahead corresponds to a form of model-based reinforcement learning (Sutton & Barto, 2018) in which the animal is able to scan its model of the environment for rewarding states by probing neighboring place cells adjacent to the currently active cell.

I.2.2 Vector Navigation, Neural Networks, and Other Later Theories

Aside from being used for path integration, the system of spatial representation cell populations in the hippocampal formation also have the potential to enable the inverse functionality: vector navigation. In contrast to path integration in which a starting location and movement information is used to predict final location, in vector navigation a desired path vector is computed between a starting point and a goal location (Bush et al., 2015). Once computed, an animal could then follow this vector within this space in order to navigate to a desired location in physical space. Bush and colleagues show in their set of simulations that the problem of vector navigation can be reduced to finding a straight line (or plane in the 2D case) between the starting and goal position within the grid representation when the phases of all different scales of grid cell representation are aligned. They discuss a variety of possible biologically grounded models which might accomplish this, including models which introduce intermediate cell populations such as distance cells (Fiete, Burak, & Brookings, 2008) and vector cells (Climer, Newman, & Hasselmo, 2013), and the model of (Erdem & Hasselmo, 2012), providing a consistent unified framework for the consideration of grid cell models of spatial navigation.

Most recently the development and downstream utilization of grid cells has been modeled in more ecologically valid conditions using Deep Neural Networks (DNNs) and more realistic virtual environment (Banino et al., 2018; Cueva & Wei, 2018). Concurrent modeling studies both showed that a neural network containing a Recurrent Neural Network (RNN) layer (Williams & Peng, 1990; Hochreiter & Schmidhuber, 1997), when trained to perform a dead-reckoning task develops representations in the RNN layer that are strikingly similar to those of rodent grid cells. In both cases the information provided to these networks consists of the starting position and angular velocity at a series of time-steps during the simulation of rodent movement within an enclosed environment. The networks were trained to predict the absolute position of the animal within the environment in x and y coordinates, as well as the head direction of the animal. Unlike previous studies, there was

no special structure imposed on the networks which a-priori biased them toward grid-like representations. Despite this, consistent grid-cell and border-cell like activation patterns were found in a large number of the neurons within the RNN in both studies.

The RNN network was chosen because of its recurrent connections, which can be thought of as modeling the highly recurrent structure within the hippocampus and entorhinal cortex. One crucial element in both sets of experiments is that unlike traditional neural networks which are trained as deterministic function approximators, the addition of noise to either the inputs or the activation of the hidden units was required for the formation of the grid cell representation. This is hypothesized to mirror the stochastic noise inherent in biological neural systems.

While Cueva and Wei (2018) only showed that a supervised learning procedure could produce grid-like cells in a neural network, Banino and colleagues went further and demonstrated that this representation could then be used in a goal-directed navigation task, suggesting that the networks learned in an unsupervised way to perform vector navigation similar to the more formal system and simulations described by (Bush et al., 2015). Banino et al. showed that an artificial agent trained with the addition of the learned grid-cell representation performed significantly better on a series of navigational tasks designed to mimic those found in the traditional rodent navigation literature, such as the Morris Water Maze (Morris et al., 1982).

This work provided the first end-to-end model of learned grid-cell activity as well as ecologically valid application of this representation in a virtual 3D environment. One main point of interest is that these spatial representations and vector-navigation ability came about without strong explicit priors on the structure of the system, or the loss function used. This suggests that the system within the hippocampus may be an instance of a more general mechanism for representing state spaces and producing efficient means of navigating them.

I.2.3 Prospective and Successor Models

The picture of the spatial representation in the hippocampus presented in the preceding section suggests that space is represented with respect to the physical make-up of the environment via bottom-up information from sensory systems, and the animal then uses this goal agnostic representation elsewhere in the brain in order to determine the optimal path to take. While the existence of time and event cells described in the preceding section has disrupted this notion, one could still imagine that when the hippocampus represents space, it does so in a straightforward Euclidean fashion. The earliest evidence from place cell firing however disagrees with this picture. Rather than providing an even representation of space, place cells are known to fire in biased ways with respect to the structure of the environment (O'Keefe, 1976). This biased firing represents a warping of the represented space that is specific to the bodily possibilities of movement available to the animal. At the end of (Muller & Kubie, 1987), the authors mention that their recordings show that the place cell system may be used to encode a forward-looking and action-oriented representation of space for the animal, which they refer to as "Kinematics."

These findings and others have led some theorists to propose that place cells are better modeled by a successor representation of the environment rather than a purely geometric one (Stachenfeld et al., 2017). In this model the activation of any particular place cell would be reflective of the exponentially discounted sum of future states reachable by the animal from its current state. Using a temporal difference update rule, this representation could be developed quickly both online and offline as an animal moves around the space or is at rest. This would also mean that the place cell firing patterns are inherently predictive of future animal behavior, rather than simply descriptive of the space itself. Such a representation also easily allows for the incorporation of time cells, whose firing patterns are naturally reflective of the temporal structure of the task at hand. In that case the intervals of animal immobility within the experiments described above correspond to distinct durations of time, interpretable as unique states.

Interpreting place field sensitivity from the perspective of a successor representation allows for a new perspective on a number of findings in the literature. Early findings that in an open 2D circular maze place fields are uniform Gaussian (Muller et al., 1987) naturally falls out of a successor representation, since the animal is equally likely to move in any adjacent position when in a given position. Findings that in 1D linear mazes separate sets of place cells fire for each direction along the maze (Foster & Wilson, 2006) are explainable when the place cells encode for a discounted sum of future states, and the animal only turns around at the end of the maze. Furthermore, the skewed nature of the receptive fields of place cells as described by (Mehta, Quirk, & Wilson, 2000) can be explained as corresponding to a prospective representation of the animal's position, rather than a geometric one. Similarly, the successor representation also helps explain the role that boundaries play in shaping place cell firing patterns, which naturally skew away from boundaries (Stachenfeld et al., 2017).

This move away from a strictly Euclidean metric representation of physical space to a representation of an abstracted state space based on future reachability opens the door to interpreting the representational nature of the hippocampus in a manner divorced from notions of physical space and time as well. If what is being represented is a non-physical state space which the animal can simulate itself "moving through," then it should be the case that this representation is utilized for tasks which are not physical in nature, but involve only abstract relations. Evidence for this kind of representation has been presented in humans (Schapiro et al., 2016; Garvert, Dolan, & Behrens, 2017). Schapiro et al. showed participants a series of images during fMRI scanning. Unbeknownst to the participants, the ordering of the image presentation was based on a predetermined graph structure, where each image corresponded to a node in the graph, and the presentation order was based on a random walk through the graph. The graph was explicitly broken into separate sub-graphs, with only sparse connections between the sub-graphs, and dense connections within them. The learned representation in the hippocampus after being exposed to sequences of

images from the graph was closer for images drawn from the same sub-region of the graph, suggesting that their temporal proximity helped shape the nature of the representation. The authors suggested the possibility of a successor-like representation being at play, but did not explicitly test this hypothesis (Schapiro et al., 2016).

Garvert et al. (2017) conducted similar work nearly concurrently, but used a graph structure that was not explicitly broken into sub-regions. During scanning, participants were asked to provide orientation judgments of the images presented. The researchers found that the representational similarity of the different images reflected their proximity on the graph. In particular, the representation was found to be best modeled by a successor representation as described in (Stachenfeld et al., 2017), in which images were represented as similar to those that were likely to be presented in the future based on the structure of the graph. Importantly, a Euclidean space representation based purely on the actual distance between items on the graph was found to be a worse fit for the data than a successor representation that considered the structure of the graph, and the policy used to walk it. This suggested that the representation encoded in the hippocampus for the task was explicitly future, and "action" oriented.

The successor representation bears a strong similarity to the Temporal Context Model (TCM) introduced to explain recency and contiguity in the domain of episodic memory (M. W. Howard & Kahana, 2002). In that model, a distributed vector representation is used to describe all the items presented to a participant, along with a separate vector being used to represent the temporal context. During learning the context vector is updated based on each newly presented item along with the previous context. The model has been applied to describe both canonical findings in the memory literature as well as to model the development of place cells in animals (M. W. Howard, Fotedar, Datey, & Hasselmo, 2005), with similar predictions as those of (Stachenfeld et al., 2017). It is perhaps unsurprising that these models show similar predictions, as it has been shown by Gershman and collaborators that not only are these models similar, but can be shown to be equivalent under

certain circumstances (Gershman, Moore, Todd, Norman, & Sederberg, 2012). This connection opens the possibility of bridging results from decades of research into both episodic memory encoding and retrieval as well as spatial learning and navigation, and both concern representing trajectories of experience in a way that can be generalized to new situations, and ultimately used to guide future action.

I.2.4 Goal Signals and the Hippocampus

The world that humans and other animals find themselves in is not a neutral space. It is filled with salient locations associated with goals and rewards that impact the nature of our behavioral choices. These locations need not be physical, as anyone who has received a pleasant surprising email, phone call, or text message can attest to. There is evidence that this inherent salience of certain states of the world is reflected in the nature of the representations present within the hippocampus.

Consistent with evidence presented in the preceding section that the place cell representation of space is not uniform and Euclidean, research conducted by (Hollup, Molden, Donnett, Moser, & Moser, 2001) found that there were significantly more place cells with activation fields near the goal location in a water maze task as compared to any other region in the maze. This biased preference in activation occurred regardless of the actual position of the goal platform within the maze. The researchers suggested that this was because of a bias toward a behaviorally salient part of the environment. This biasing of activation suggests not only a representation which is geared toward an abstract state space representation, but also one which privileges the goal location in that state space.

Similar to the bias of place cell firing around a goal location is the bias of firing patterns in the replay of trajectories during rest in rodents. Pfeiffer and Foster conducted experiments showing that the trajectories in replay events are biased towards the goal location in a 2D foraging task (Pfeiffer & Foster, 2013). They showed that this bias in the kinds of trajectories replayed could not be explained by either frequency of visitation, or

by a simple function of the heading direction of the animal. Instead they reflected the likely future behavior the animal would engage in after rest, and preferentially ended at the goal location significantly more often than any other location in the maze. They furthermore demonstrated that the replay trajectories in many cases corresponded to non-experienced combinations of both start and end positions, suggesting the ability for the animal to generalize to novel goal locations and start positions.

The ability to replay novel trajectories was further demonstrated in work showing that rodents were able to preplay trajectories to a novel goal even when that goal location was never visited before by the animal (Ólafsdóttir, Barry, Saleem, Hassabis, & Spiers, 2015). Using a T-maze, researchers examined spontaneous trajectory activation patterns in CA1 of the rodent hippocampus. The animal was placed at the end of the maze, with a barrier preventing it from reaching the decision making fork in the maze. From there a rewarding object was placed in the right wing of the maze, and made visible to the animal. During a rest period following this presentation, recordings were made from CA1. These recordings showed significant preplay-like events for the wing of the maze containing the reward, but not for the wing without the reward. This suggests that the animal was able not only to integrate the presentation of the goal into a general representation of the maze, but also then perform a kind of planning corresponding to the future behavior of the animal, taking a path to that rewarding location. These results, along with those of (Pfeiffer & Foster, 2013) further complicate the distinction between the notions of replay and preplay, suggesting that both exist as instances of a more general phenomena of "playing" or "simulating" possible experiences within an abstracted state space. The extent to which these simulated trajectories do or do not correspond to actually experienced trajectories is the extent to which they might be referred to as either replay or preplay.

Pezzulo and colleagues propose to generalize this notion into phenomena which they refer to as internally generated sequences (IGS) (Pezzulo et al., 2014), which corresponds to any activation of cell populations in the hippocampus which reflect trajectories through

the represented state space that are not reflective of the actual position of the animal. They propose that these IGS events are goal-driven, and correspond to a model-based learning system. In this view, IGS events can either correspond to the updating of an environmental model, or the application of that model for forward-planning. The model they propose bears a strong resemblance to the MCTS algorithm (Silver et al., 2016), in which an explicit search procedure is combined with a learned value estimator. Such a view finds supporting evidence in research showing that these IGS events in the hippocampus are correlated with similar sequence events in the ventral striatum, which are known to relate to value estimation (Lansink, Goltstein, Lankelma, McNaughton, & Pennartz, 2009).

The existence of a bias toward goal locations and trajectories in the CA1 region of the hippocampus suggests that goal signals are indeed influential on the hippocampal formation as a whole. Consistent with this hypothesis have been results from a number of recent studies showing the ability of researchers to decode goal signals in the entorhinal cortex and subiculum (Spiers & Maguire, 2007; L. Howard et al., 2014; Chadwick, Jolly, Amos, Hassabis, & Spiers, 2015). Spiers and Maguire had human participants engage in a goal-directed navigation task in a virtual environment (Spiers & Maguire, 2007). While conducting this task, fMRI was used to explore whether goal proximity could be decoded from the brain. They found that both mPFC and entorhinal cortex signals were significantly correlated with goal distance. Similar work using a set of snapshots from a video of city navigation was used to demonstrate that a human's Euclidean distance from a goal could be decoded from the entorhinal cortex (L. Howard et al., 2014). Lastly, Chadwick and colleagues demonstrated that goal direction could be decoded from the entorhinal cortex in humans during a goal direction judgment task in a virtual environment (Chadwick et al., 2015). Interestingly the goal direction signal was allocentric in the entorhinal cortex, while a separate egocentric goal direction signal was found in the precuneus. Taken together these experiments provide evidence for the modulation of hippocampal formation activity by a goal signal. Chadwick and colleagues hypothesized that it was a population

of head direction cells responsible for both goal direction and head direction representation being decoded in their experiments. This suggests that there is a kind of negotiation between bottom-up sensory information and top-down goal or recall driven signals guiding the active representation in the hippocampus at any given time.

The "top-down" nature of goal-selective activation described above must correspond to some other brain region or regions. It has been hypothesized that the medial prefrontal cortex (mPFC) in particular is a potential generator of such a goal signal or specification (Poucet & Hok, 2017; Erdem & Hasselmo, 2012; Pezzulo et al., 2014). Such a theory states that the mPFC would generate relevant goal states based on the current state of the animal and relevant incoming sensory information, as mediated by the hippocampus. It would then engage in a goal specification, which would influence the hippocampal representation in the ways described above, thus inducing trajectory to or from the goal location within the abstract representation space provided by the hippocampal formation. This activation would then be passed to the ventral striatum, where explicit value estimations would be produced. Indeed, recent evidence suggests that the hippocampal formation serves a critical role in the stable functioning of such a model-based planning system in humans (Vikbladh et al., 2019). What remains to be explained is what computational principles may serve as the basis for the learning and application of such goal, state-space, and value estimation representations.

I.2.5 Policy Learning from Real and Imagined Experience

Thus far we have described a system of representations which allow for the generation of a state representation and goal signals, the simulation of an abstracted future-oriented state space. There are a number of methods which can then be used to obtain value estimates and optimal actions using these building blocks. These include TD-learning (temporal difference learning) methods which update a policy or value function every time step (Sutton & Barto, 1990), Dyna, which enables additional offline learning using a model of the en-

vironment (Russek, Momennejad, Botvinick, Gershman, & Daw, 2017), and tree-search planning methods (Daw, Niv, & Dayan, 2005). These have all been proposed at the higher level of behavior, rather than as specific suggestions concerning the nature of how the hippocampal state space representation is learned. In this specific domain there have been similar suggestions, such as the TD-update rule proposed as a means of potentially learning place cell activation patterns (Foster et al., 2000; Stachenfeld et al., 2017).

Evidence for the existence of internally generated hippocampal sequences during sleep and rest have opened the door to more sophisticated models of learning in these systems. In particular, the Dyna algorithm has been seen as a means of providing a unified explanation for replay during sleep and awake states (Johnson & Redish, 2005; Russek et al., 2017). At the simplest level, the reactivation of trajectory sequences can be interpreted as the brain performing learning on these sequences. In models where we are directly learning a value estimate, this reactivation would correspond to updating the value estimates for the states which are part of the reactivated trajectory. If we assume that the state representation is based on a successor representation as proposed above, then the update would be not to the value estimation, but rather to the state representation itself. In either case, learning from whole trajectories opens the possibility of applying more sophisticated learning rules than TD.

This model of learning from internally generated sequences can be further extended if we assume a non-uniform activation of the internally generated sequences. Indeed, there is a evidence for this, as described in Section I.2.4. As described above, sequences which lead to goals, either visited or known through observation are played more frequently than random (Pfeiffer & Foster, 2013; Ólafsdóttir et al., 2015). This has been hypothesized to correspond to a prioritized replay mechanism (Mattar & Daw, 2018), similar to what has been proposed in the artificial intelligence literature as a mechanism for increased efficiency and stability during learning (Moore & Atkeson, 1993; Mnih et al., 2015). In this system, experiences are selected for activation (and learning) based on the strength of an error signal

(referred to as gain), combined with an expected visitation signal (referred to as need). By biasing the activation of events and sequences using these factors, the canonical forward and reverse replay events naturally fall out.

Take for example the well-studied phenomenon of reverse replay which takes place when an animal reached a goal location. If this location is novel, then the value estimation error signal will be large for that state, and lead to replay, and subsequent updating using a TD-learning rule, thus decreasing the discrepancy between the estimated value of that state and the experienced value of the state. From this point the state with the next greatest discrepancy would then be the preceding state, then the state preceding that one. In this way a reverse replay backwards all the way along the trajectory would take place until the sequence reaches the starting position. Mattar and Daw (2018) also include a need term in their model to explain the forward play of sequences from the animal's current position when the animal is placed at the beginning of a maze. This need term gives high priority to states likely to be reached from the current state, such as those directly in front of the animal.

Beyond fitting behavioral data to computational models, these theories around offline prioritized replay as a mechanism for learning have begun to be tested empirically in humans (Momennejad, Otto, Daw, & Norman, 2018). Momennejad and collaborators have recently shown that replay events in humans during rest not only correspond to what would be expected by a prioritized replay mechanism, but that their reactivation is predictive of subsequent performance improvements on a two-step decision task. In those experiments, participants were exposed to a two-step decision task with certain transition and reward values. The structure of the task was then changed for participants in one condition, those participants had to re-learn the new optimal policy and value estimates for the task. The researchers used multi-voxel pattern analysis to decode the replay events of the participants during rest, finding that in the reevaluation condition there was significantly more replay. They also show that the replay that takes place is consistent with a prioritized reactivation

scheme similar to what is described above. Importantly, this reactivation is predictive of the extent of adaptation of the participants to the new task structure. While preliminary, these results give initial support to the computational theories of prioritized replay, offline learning, and the hippocampus as a cite of a successor representation.

I.3 Generative Temporal Models

The general framework of mapping sensory information into an abstracted state, representing that state in an predictive manner, combining the predictive representation with a goal signal to produce value estimates and candidate actions has been proposed as a possible model of high-level decision making in mammals (Russek et al., 2017; Pezzulo, Kemere, & Van Der Meer, 2017). In one such model, each of these processes can be roughly mapped onto the sensory cortices, medial temporal lobe, and ventral and dorsal striatum, respectively (Pezzulo et al., 2017). Indeed, there is evidence for sensory cortices learning compressed state representations, (Van Essen & Maunsell, 1983) corresponding to both the 'what' and 'where' or 'how' of the visual stream (Kravitz, Saleem, Baker, & Mishkin, 2011). There is evidence of the hippocampus performing pattern separation (Yassa & Stark, 2011), prospective representation learning (Stachenfeld et al., 2017; Garvert et al., 2017), and preplay of future movement (Pfeiffer & Foster, 2013). Finally, there is evidence of ventral striatum performing value estimation (Kable & Glimcher, 2007; Peters & Büchel, 2010), and guiding behavioral learning in the dorsal striatum (O'Doherty et al., 2004).

In addition to evidence of each of these distinct brain regions acting separately, there is evidence for the relevant sets of connections to support the hypothesized goal-oriented model of Pezzulo et al. (2017). These connections include between sensory cortex and the MTL (Ji & Wilson, 2007; Kravitz et al., 2011), and between the hippocampus and the striatum (Lansink et al., 2009; Pennartz, Ito, Verschure, Battaglia, & Robbins, 2011). Taken together these interconnected systems provide one possible working framework for goal-directed learning and action necessary to support and utilize a cognitive map.

This theoretical model of a predictive agent described above shares a strong resemblance to a class of neural network models called generative temporal models (GTMs). Also more informally referred to as world models, in this work we will make use of both. These models are trained to predict future sequences of observations within an environment, given a past sequence of observations and actions. In doing so, these models learn the underlying structure of the environment they are trained in. These GTMs can then be used to guide reinforcement learning and goal-driven decision making in the environment which they were trained.

A central thesis of this work is that the connection between GTMs and the "experience construction system" of the medial temporal lobe in humans is more than just a superficial one. Here we propose that a certain class of GTMs can serve as a useful model for an array of neural and behavioral findings associated with the hippocampus. The purpose of this work is to demonstrate this with a series of informative examples, and provide insight into how this theoretical model can be further extended with future work.

A main tenet of the construction systems hypothesis is that the hippocampus serves to support the generation of coherent narrative experiences, whether they be of remembered events in the past, or imagined events in the future. A more computationally specific way of framing this hypothesis is that the hippocampus is responsible for maintaining a generative model of semantically coherent trajectories through an abstract state space which is correlated with cortex states, and their accompanying experiential properties. In this sense, the hippocampal formation is indeed a very specific kind of generative temporal model, one capable of storing and making sense of, and allowing for the recall of, millions of experiences that an animal may encounter in their life.

We believe that GTMs can generalize a number of the theoretical findings discussed above in this introduction. Rather than starting from pre-existing computational building blocks, this work seeks to demonstrate that cells with firing properties such as place and time cells, as well as phenomena such as replay/preplay can naturally derive from unsu-

pervised learning of GTMs for the purpose of goal-driven navigation. In the course of this dissertation, we will be demonstrating this with increasingly realistic stimuli and environments.

This section of the introduction seeks to further make clear this connection, as well as to provide preliminary context and definitions for the generative temporal models and components that will make up the majority of the work discussed in the subsequent chapters.

I.3.1 Basics of Generative Temporal Models

At the core of many recent instantiations of the generative temporal model framework is a variational autoencoder (VAE) (Kingma & Welling, 2013). VAEs are a class of stochastic generative models which learn compact latent representations of data using a log-likelihood learning objective. Due to their theoretical grounding and connection with Bayesian learning, VAEs have recently emerged as computationally viable means of performing predictive coding (Friston & Kiebel, 2009), whereby the probability distribution over high-dimensional observation spaces can be tractably computed.

A VAE is composed of two tightly coupled neural networks, one which performs inference of the latent variable from an observation, and the other which performs generation of a predicted observation using a latent variable. These networks are sometimes also referred to as encoder and decoder networks, respectively. In a VAE, the sensory stream of observations o is sent through the inference network to be encoded into a latent distribution, from which a latent variable z is sampled. The specific nature of this distribution can vary, with most common instantiations using a gaussian distribution (Kingma & Welling, 2013; Higgins et al., 2016). As will be discussed below for its connection with the hippocampus, the gumbel-softmax distribution (Jang, Gu, & Poole, 2016), which induces sparsity in the representation is another choice for latent distribution within a VAE. Once a latent variable is sampled, the generative process takes place, whereby a predicted sensory perception o^* is decoded from z using the decoder network. See Figure 1 for a diagram of the variational

autoencoder network flow.

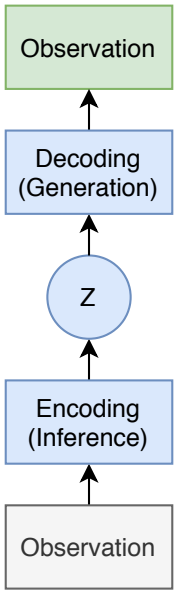

Figure 1: Diagram of a variational auto-encoder. Boxes correspond to deterministic variables. Circles correspond to stochastic variables. Grey corresponds to input variables. Green corresponds to output variables. Blue corresponds to network layers responsible for context information.

A VAE is trained end-to-end in order to both maximize prediction accuracy as well as to optimize a regularization term used to induce a smooth manifold in the latent space (Kingma & Welling, 2013). This regularization term often takes the form of a KL (Kullback-Leibler) divergence loss between the current distribution and some target prior distribution. In the case of a gaussian latent distribution, the prior is a normal distribution. In the case of a gumbel-softmax distribution, the prior is a uniform distribution. The weighting of these two loss terms results in a trade-off between accuracy and generality. The regularization term has been demonstrated to induce a disentangled latent representation in the gaussian case (Higgins et al., 2016). In Chapter II, we demonstrate that a similar phenomenon takes place when using a gumbel-softmax distribution.

By itself, a VAE is not sufficient to meet the criteria of a GTM, as there is no temporal component allowing for the generation of future states from a current state. By extending a VAE with a forward dynamics model however, it gains the ability to model the temporal

dynamics of an environment. A forward model is a function $s_{t+1} = f(s_t, a_t)$, where s is a state, and a is an action taken by an agent in the environment, and t is the current time-step of the environment simulation. A recent example of this simple but powerful idea was the "World Model" (Ha & Schmidhuber, 2018). This model combined a latent state representation from a VAE with a recurrent neural network, specifically implemented as a LSTM (Hochreiter & Schmidhuber, 1997). Rather than modeling the temporal dependencies between the high-dimensional sensory observations o, the World Model learns to model only the dependencies between the low-dimensional latent states z produced by the VAE, making the learning problem significantly more tractable. See Figure 2 for a diagram of the World Model. Furthermore, it makes possible planning within the latent space, since the learned low-dimensional latent states can be used as the basis for performing reinforcement learning.

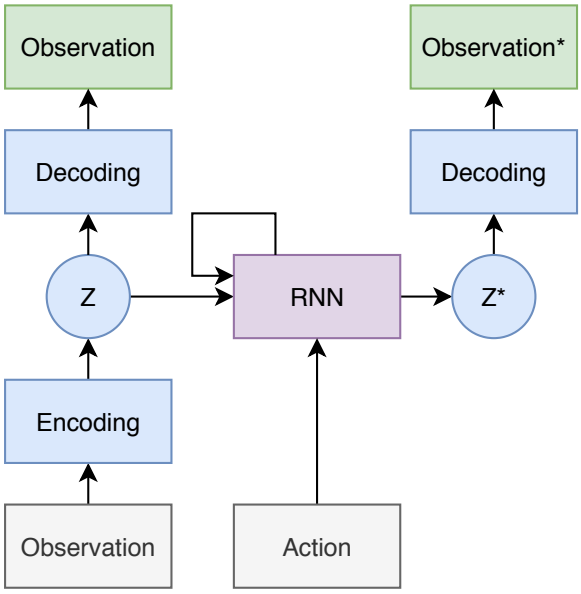

Figure 2: Diagram of a World Model. White corresponds to model input. Green corresponds to model output. Blue corresponds to content information. Purple corresponds to joint context and content information. Nodes marked with a $*$ correspond to values at the next time-step, or predictions of those values.

This basic formulation has been extended in the "Recurrent State Space Model" (RSSM), which augments the stochastic latent state z with an additional deterministic latent state h

kept within the RNN (Hafner et al., 2018). This results in both greater model representational capacity, but also the ability for the model to partition that capacity into representing stochastic and deterministic aspects of the environment independently. By utilizing both stochastic and deterministic latent states, RSSMs have been able to model the dynamics of complex control tasks from high-dimensional visual observations, and use the model to perform efficient reinforcement learning (Hafner, Lillicrap, Ba, & Norouzi, 2019). See Figure 3 for a diagram of the network flow of an RSSM. More recent models have extended this formulation in a hierarchical manner, to enable modeling of environment dynamics at multiple different temporal scales, enabling planning in environments with large state spaces (Kim, Ahn, & Bengio, 2019).

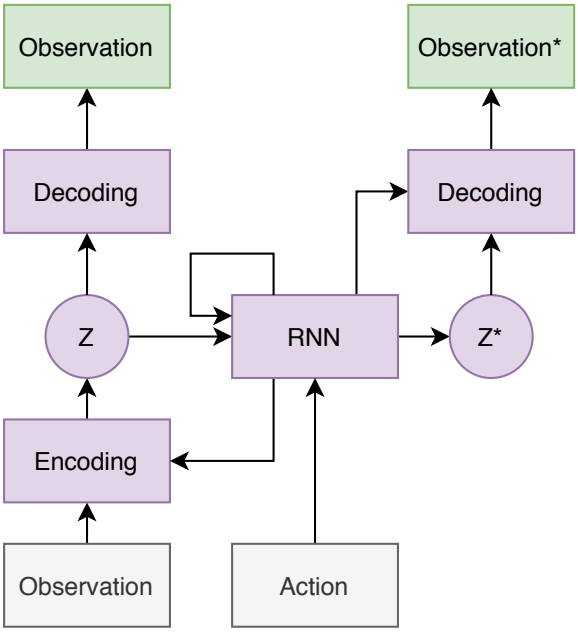

Figure 3: Diagram of a Recurrent State Space Model (RSSM). White corresponds to model input. Green corresponds to model output. Purple corresponds to joint context and content information.

I.3.2 Extending GTMs with Memory and Multiple Latent States

The generative temporal models described above are powerful methods for learning the dynamics of an environment and using them to plan goal-directed behaviors. They fall

short of one key aspect of the capabilities of mammals with cognitive maps however, and that is the ability to quickly learn from and make use of novel experiences. Humans and other mammals are able to remember a series of events that only needs to take place once, learning to navigate novel environments in a so-called "one-shot" manner. This is made possible due in part to the highly plastic nature of the recurrent connections within the hippocampus (Frank, Stanley, & Brown, 2004).

In addition to this plasticity, there is a critical separation between the content (objects within scene) and context (location) of the incoming sensory stream of information. Representations in the upstream LEC and MEC have been demonstrated to contain content and context information respectively (Hafting et al., 2005; Deshmukh & Knierim, 2011). This structured information allows for more intelligent storage and retrieval of latent states than is possible in a World Model or RSSM, where this information is entangled together.

Attempts to use more structured and fast-adapting methods have resulted in a new class of GTMs which are indeed able to capture many of the additional capabilities of the cognitive map that the simpler models were lacking. Key to these innovations has been the addition of various kinds of differentiable neural dictionaries (DND) used for additional storage within the network beyond what a recurrent neural network is capable of (Pritzel et al., 2017). These differentiable dictionaries are initiated at the beginning of an episode of experience for a virtual agent, and are then used to store and recall information during the episode. A simple example of this designed for 2D environments is the Generative Temporal Model with Spatial Memory (GTM-SM), which uses a VAE along with a hand-crafted DND (Fraccaro et al., 2018). In a GTM with a DND, a new memory is written at each time step in the form of a key-value pair, with the context variable serving as the key, and the content variable serving as the value. During recall, stored keys are compared to a query key, and used to determine which value to recall. See Figure 4 for a visual representation of the network flow of a GTM-SM.

The recently proposed "Model-Based Predictor" (MBP) model utilized recurrent VAE,

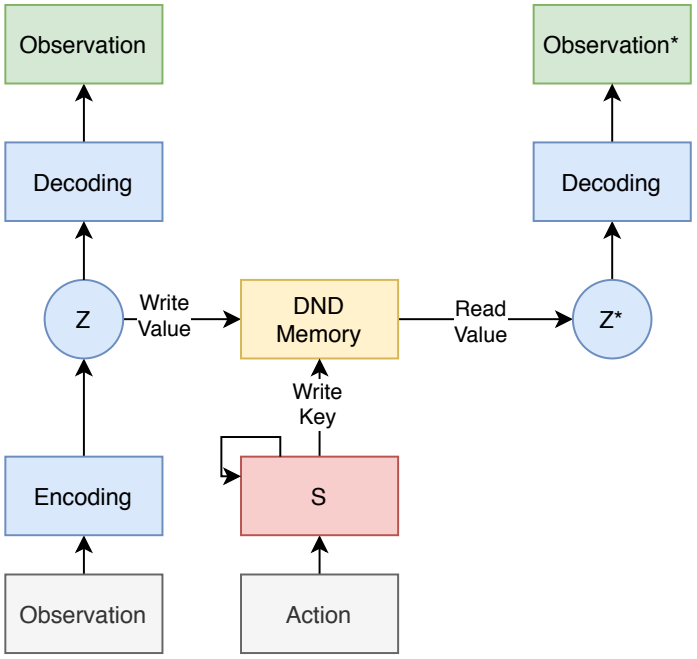

Figure 4: Diagram of the Generative Temporal Model with Spatial Memory (GTM-SM). Blue corresponds to content information. Yellow corresponds to a differentiable memory store. Red corresponds to context information. While corresponds to model inputs. Green corresponds to model predictions.

but additionally augmented with a differentiable memory module similar to a DND, but which uses a multi-headed query system in order to enable more complex learned storage and retrieval mechanisms (Graves et al., 2016). This model was furthermore trained end-to-end to not only perform memory recall, but also to perform goal-directed navigation tasks in a few-shot manner (Wayne et al., 2018). As such, the memory module acted as a learn-able dictionary look-up, where new experiences could be stored and retrieved as was demanded by the task.

Even more recently the "Tolman-Eichenbaum Machine" (TEM) has been proposed as a model of entorhinal and hippocampal representation learning (Whittington et al., 2019). This model similarly utilizes a VAE framework, but explicitly accounts for separate 'content' and 'context' input streams from the lateral and medial entorhinal cortices, respectively. Like MBP, TEM also uses a differentiable memory module to store and retrieve the bound representations. The resulting model demonstrates many predicted properties and

representations in the medial temporal lobe such as grid, border, and place cells, along with neurally consistent remapping.

We propose and examine a novel variant of the dictionary-based GTM called a Dual Stream World Model in Chapter IV of this work.

I.3.3 Hippocampal Index Theory and a Language Metaphor

Due to the success of dictionary-based generative temporal models, it is perhaps of value to examine the dictionary metaphor more closely, as it pertains to the medial temporal lobe. If the medial temporal lobe is a kind of dictionary, with keys and values, then it is for a language of narrative experiences, or episodic memories.

This notion of a dictionary is closely related to that of the hippocampal index theory (Teyler & DiScenna, 1986). According to this theory, the hippocampus quickly forms a low-dimensional representation corresponding to the higher-dimensional cortex states. This low-dimensional representation being an "index" for the higher-dimensional one. This index can be interpreted in the simplest context as the latent state of a variational auto-encoder, as discussed above. It can also be interpreted as a key of an entry in a dictionary, with the value of that entry corresponding to the higher-dimensional state.

Humans use and deploy a verbal and written language composed of words which we string together using a system of syntax and grammar. Each of these words has a corresponding meaning, and a specific place within any given sentence that the word must go in order to be semantically meaningful. Given a series of words in a sentence, there are only so many words that might end the sentence, for example. Consider a sentence like 'The cat sat on the ___.' Most people who have undergone traditional English education would implicitly want to end that sentence with 'mat.' Furthermore, English speakers also know what a 'mat' refers to in this context. Likewise, when we walk around our homes, and walk into a kitchen, we know to expect an oven, a refrigerator, and cabinets.

This concept of language consisting of meanings, words, and a grammar can be a useful

metaphor for understanding the role that the medial temporal lobe, and the hippocampus in particular plays in the mammalian brain. In this metaphor, we can think of place, time, and event cells as being "state cells," with each corresponding to the words of a language. These word tokens can be seen as equivalent to the indices of the hippocampal index theory. The temporal dynamics of the hippocampus, specifically of the CA1 and CA3 regions then correspond to the syntactic structure within which the language unfolds, and how one index follows another. The connection between the hippocampus and the cortex, mediated by the lateral and medial entorhinal cortex acts as the process of storing and looking up words in a dictionary, and associating words with their definitions.

The use of such a symbolic language is convenient for many reasons. It allows us to swap in simple tokens consisting of a few syllables for complex ideas and objects within the world. We then simply need to make use of a mapping between these high-dimensional meanings and the words. A similar problem arises in the domain of memory and goal-directed navigation. Our narrative experiences are filled with extremely high-dimensional perceptual, cognitive, and affective information. Rather than storing and learning the transitions between these high-dimensional variables which exist as states of the cortex, the hippocampus generates low-dimensional tokens in the form of sparsely-firing "state" cells. These then serve as an index or placeholder for the cortical activation, and corresponding phenomenal experience in the animal.

Because they are abstracted away from the cortex state itself, these hippocampal states can also serve to enable generalization when the content of the sensory perception changes, but the structural aspects of the environment remains the same. Such is the case if a room needs to be navigated after a new coat of paint on the walls, or a new pattern on the rugs. Such superficial changes should not change the state itself. Indeed, such changes to environments do not result in hippocampal remapping of place cells in experiments with rodents (Muller & Kubie, 1987).

In addition to the greater capacity for storing these low dimensional "state" cells comes

the related benefit of easier composability. Because these tokens are low-dimensional and re-usable, it is possible to generate sequences or motifs of them with comparative ease, compared to their high-dimensional cortical counterparts. These sequences can be seen as being akin to stock phrases in languages. These would map to sequences of known experiences, such as the experience of walking down a hallway, where a sequence of half a dozen place cells might always activate in the same order every time the hallway is traversed. In the same way, longer narrative experiences such as one's trip across town to run errands can be composed of sequences of these motifs without recourse to tying all of the underlying cortical states together. In this way memories can be quickly formed and stored in the hippocampus before the much longer-term process of transfer to long-term memory takes place.

The use of these simple sequenced tokens also allows for the creation of novel sequences of "state" cell activations. In the same way that humans learn to play with language to explore the linguistic possibilities, the processes of imagination and planning engage the 'language' of the medial temporal lobe and can allow for the exploration of novel sequences of events. Importantly however, these sequences are not, and cannot be arbitrary, as there is a syntax and grammar to this language. In the same way that some sentences don't make grammatical sense, some sequences of experiences don't make navigational sense. This ties directly into empirical research which has shown that spontaneous place cell activity follows motifs of groups of two or three units (Liu et al., 2018). These can be thought of as the basic phrases by which the language of the hippocampus is composed. The breakdown of this capacity is related to breakdowns in narrative coherence in patients with hippocampal damage (Hassabis et al., 2007).

While not perfect, we believe that this metaphor has the value of providing an interpretation for the success of recent GTMs such as MBP and TEM. Furthermore, it can help guide the development of novel models, such as those we will present here. First however, it will be of benefit to point out the properties of simpler generative temporal models, and

how even basic models can support the development of place-like cells.

CHAPTER II

THE HIPPOCAMPUS AS A GENERATIVE TEMPORAL MODEL

The preceding chapter surveyed the current state of our understanding of the hippocampus and its ability to support a flexible system of navigation which has been referred to as a cognitive map. It also introduced a powerful class of computational models referred to as GTMs which can match a number of the empirical findings of the cognitive maps in humans and other mammals with respect to navigational ability in both familiar and novel environments. We now turn to a concrete demonstration of the properties of GTMs and their relationship to the representations found in the hippocampal formation. Rather than starting with a complex GTM, we will begin our analysis from first principles, demonstrating basic properties of a simple GTM, and only later moving on to a more complex model. In this chapter, we will demonstrate that cells with firing patterns similar to those found in hippocampal place and time cells, which we have referred to as "state cells" can arise from a basic form of a GTM. The learned representations will then be shown to display properties of hippocampal representations in humans and other mammals, namely the temporal community structure (Schapiro et al., 2016).

II.1 Place and Time Cells in a GTM Latent State

As discussed above, the place cell was the first major spatially selective cell to be discovered in the hippocampus (O'Keefe, 1976), and provided the initial evidence that the hippocampus is an important brain region for those interested in understanding cognitive maps in mammals (O'Keefe & Nadel, 1978). We likewise begin our analysis of the connection between GTMs and cognitive maps in the same way, by looking at the conditions under which a simple GTM can be shown to develop units with place-cell like firing properties.

Key to the development of place-like cells in our model will be the use of the gumbel-softmax (GS) distribution to represent the latent space of the variational auto-encoder (Jang et al., 2016). This distribution was developed to allow for sampling from categorical distributions while maintaining differentiability, which is essential for solving certain tasks with neural networks trained using backpropagation. This representation has the effect of inducing sparsity on the representation being learned, due to the "softmax" operation. See Figure 5 for a visual representation of the distribution, and example samples from it. In the context of a model of the medial temporal lobe, this sparsity can be seen as being induced by the dentate gyrus within the hippocampus, a region through which a significant amount of incoming information passes, and which contains sparse connections to downstream hippocampal regions (Leutgeb, Leutgeb, Moser, & Moser, 2007).

The use of a gumbel-softmax latent distribution also has an important connection to clustering algorithms, where the size of the GS distribution determines the upper bound on the number of possible clusters, and each cluster emerges in a "soft" and probabilistic sense. This directly relates to a recently proposed theoretical model of hippocampal dynamics (Mok & Love, 2019). In the model proposed by Mok and Love, the hippocampus performs clustering on the sensory stream of inputs, and place cells develop as a special case of this in strictly spatial environmental contexts. Likewise, time cells emerge as the temporal case, where specific durations of time are clustered into groups, and these are

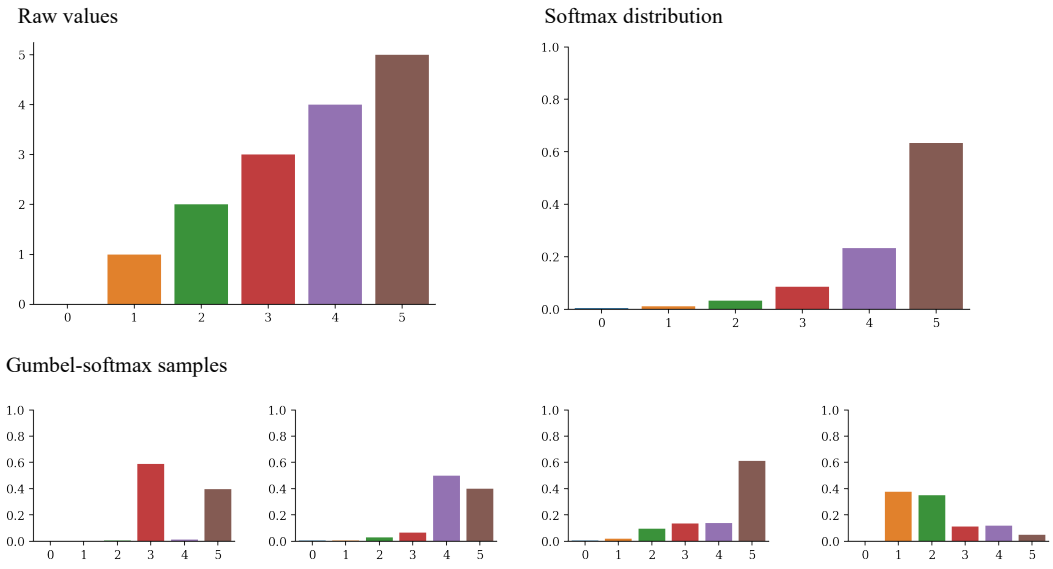

Figure 5: Explanation of Gumbel-Softmax distribution. Top left: hand-generated underlying values used to produce distribution. Top right: softmax distribution created from raw values. Bottom: four random samples from the gumbel-softmax distribution.

used as downstream variables. The model proposed below can be seen as taking similar computational inspiration as (Mok & Love, 2019), but demonstrating this principle within the context of an end-to-end differentiable neural network, which has both greater biological plausibility than the k-means clustering algorithm used in (Mok & Love, 2019), as well as allows for greater representational capacity.

As discussed in the introduction, there is not a clear delineation between place cells and other cells in the hippocampus which also display limited selectivity, such as time cells. Indeed, there is evidence for cells taking on either place or time like properties as the task and environment contingencies demand (MacDonald et al., 2011). Here we also demonstrate that the same computational principle which allows for the development of place cells also allows for the development of time cells in the case of the incoming information providing a temporal signal, as has been found in the LEC (Tsao et al., 2018).

II.1.1 Evaluation Methods

We begin by defining a simple two-dimensional environment within which an artificial agent might move and act. The observations available to the agent within this environment will be the x and y coordinates of the agent's position, as well as the time that has passed since the beginning of the episode. While neither quantity is available as raw sensory information to an animal directly, both are known to be represented in the entorhinal cortex as a result of integrating sensory information over time. Specifically, Euclidean position is decodable from the spatial information represented in MEC in the form of grid cells (Hafting et al., 2005), and time information can be decoded from the LEC in the form of ramping cells (Tsao et al., 2018). The actions available to the agent will be movement in the northern, eastern, southern, and western directions by one unit per time step. While a simplification of actual animal action, this can be seen as corresponding to a simplified version of the animal's head-direction system, which exists in the subiculum and provides a global orientation input to the hippocampus (Taube et al., 1990).

In this environment, the positions the agent can occupy are discrete (as such it falls into the category of virtual environments typically referred to as a "gridworld", a term we will use throughout this work), and the size of the environment is 12×12 units, with walls the agent cannot move onto taking up the outer rim of units. As a result, there are 10×10, or 100 movable positions the agent can occupy. The observations the agent receives is then a vector of length 3, corresponding to $< x, y, t >$. Likewise, the agent will produce actions as a one-hot vector $< n, n, n, n >$, where n is 1 in the position corresponding to the current movement-direction, and 0 elsewhere. See Figure 6 for a visual representation of the gridworld environment.

The model is trained in an offline fashion, with the data being first collected from a series of random walk trajectories with each initializing the position of the agent in a randomized location in the environment. The random walk is based on a policy whereby either a new action is taken with a uniform random probability, or with some probability

the previous action is repeated. Each random walk lasts 50 time-steps, and 1000 of these trajectories, each referred to as an "episode" were collected.

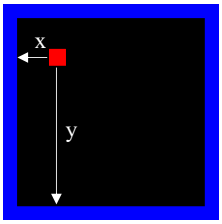

Figure 6: The simple two-dimensional "gridworld" environment. Agent is represented by a red square. Walls are represented by blue squares. Agent's observation of x and y coordinates represented as white arrows.

II.1.2 Modeling Methods

We then used the collected dataset to train a World Model (see Figure 2), as described by (Ha & Schmidhuber, 2018), with minor modifications. In the original implementation of the World Model, a gaussian distribution was used to represent the latent variable z. Here we compare this approach to two other candidate latent space types, a gumbel-softmax distribution, and a deterministic linear layer. The World Model can be broken into an inference and generation phase which alternate throughout each time step of an episode of training. Below are the explicit equations describing these phases.

The inference phase is governed by the following equations.

$$z_t \sim p(z_t|o_t) \tag{II.1}$$

$$h_{t+1} = f(h_t, a_t, z_t) \tag{II.2}$$

Where h_t corresponds to the hidden state of the recurrent neural network, and z_t refers to the inferred latent state. The sampling of z_t differs based on the distribution being used.

In the gaussian case, it is sampled as follows:

$$z_t = \mu(x_t) + \sigma(x_t) * \varepsilon \tag{II.3}$$

Where μ and σ are outputs from the encoder network, and ε is sampled from a normal distribution.

In the case of a gumbel-softmax (GS) distribution, z is sampled as follows:

$$z_t = \frac{\exp(\log(x_t) + g)}{\sum \exp(\log(x_t) + g)} \tag{II.4}$$

Where g is sampled from the gumbel distribution, which consists of a transformation of a uniform random sample between 0 and 1, u as follows: $g = -\log(\log(u))$.

Once a latent variable z_t has been sampled, the generation phase then proceeds as follows.

$$z_{t+1} \sim q(z_{t+1}|h_{t+1}) \tag{II.5}$$

$$o_t^q = f(z_t) \tag{II.6}$$

$$o_{t+1}^q = f(z_{t+1}) \tag{II.7}$$

We train the model using the same loss functions used in the original World Models paper, which include a reconstruction loss, a forward model loss, and a regularization loss.

$$L_O = \frac{1}{n}\sum_{n=1}^{N} |o_t^q - o_t|^2 \tag{II.8}$$

$$L_Z = D_{KL}(p(z_t|o_t)||q(z_t|h_t)) \tag{II.9}$$

$$L_{Total} = L_O + L_Z - \beta Hs \tag{II.10}$$

Where Hs is the regularization term which varies based on the distribution used, and β is the strength of the regularization. This regularization term is essential to the training of variational auto-encoders, as it enforces non-deterministic latent spaces, and has the effect of inducing disentangled representations in the latent space as a result (Higgins et al., 2016). In the case of the gaussian distribution, this is the KL divergence between the current distribution and a normal distribution. In the case of the gumbel-softmax distribution, this is the entropy of the distribution.

II.1.3 Results

First, we trained a model using only the $< x, y >$ components of the observation space. Using this dataset, we trained three separate models, each with a latent space size of 64, but each containing a different latent distribution type: gaussian, gumbel-softmax, and a deterministic linear layer. When comparing the learned latent spaces of these models, we find that only the GTM trained with the GS latent space learned a representation with place-like cells. This can be seen clearly in Figure 7, where the activation profile of units in the GS model show extremely high spatial selectivity, and little redundancy between units. In contrast, the spatial selectivity of the other models is non-coherent, and highly redundant.

We then compared the reconstruction error of the three models trained using different latent distributions, we find a significant difference between all three (ANOVA, $F(2, 3897) = 158.668, p < 0.0001$), with the gumbel-softmax model ($Mean = 0.010, Std = 0.017$) resulting in the lowest reconstruction error, followed by the gaussian model ($Mean = 0.021, Std =$

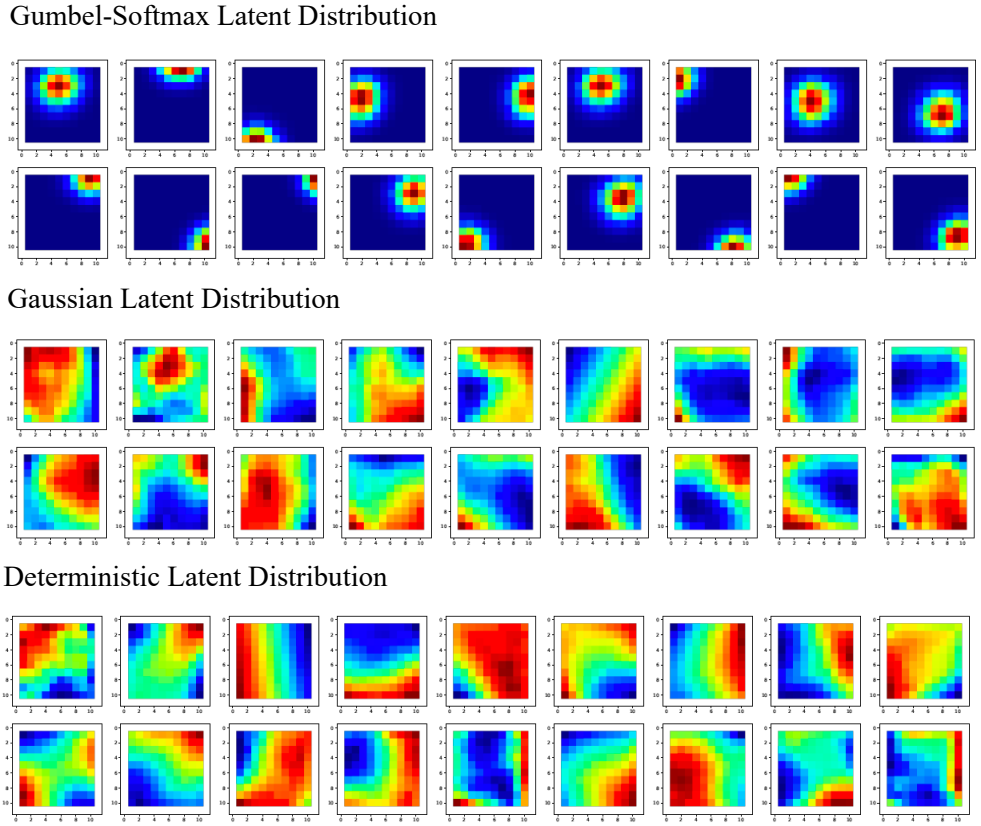

Figure 7: Representative activation patterns of the first 18 units in the latent variable z in world models trained using gumbel-softmax, gaussian, and deterministic latent distributions. Around each box are the walls of the environment which were not accessible to the agent.

0.024), followed by the model with a deterministic linear layer ($Mean = 0.027, Std = 0.030$). Pairwise comparisons result in highly significant differences between the three (all $p < 0.001$). These results are presented in Figure 8.

These results suggest that in addition to supporting the development of structured place-like cells, the gumbel-softmax distribution is also results in an auto-encoder with better reconstruction accuracy for spatial information than a gaussian distribution or deterministic linear layer.

In order to better understand the effect of the regularization term in the optimization process of the GTM with gumbel-softmax latent space, we trained a set of four additional models, each using a different value for β. We choose the following set of values, to

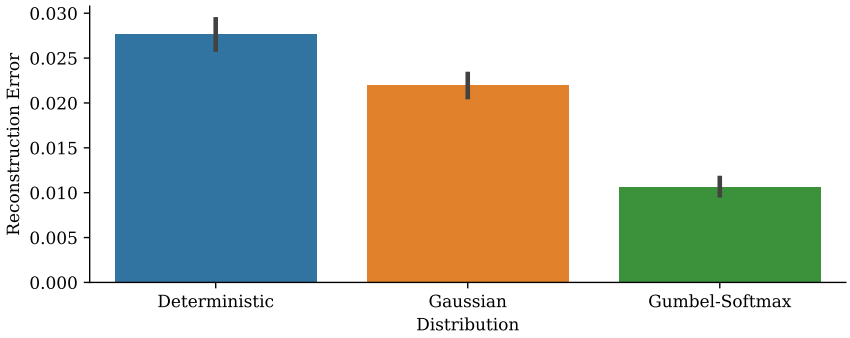

Figure 8: Reconstruction errors of three model types trained to auto-encode spatial observations. Error bars represent standard error.

provide a range of values with which to examine $\beta \sim < 0.0, 0.01, 0.05, 0.1 >$. As described in Higgins et al., in the case of a VAE with a gaussian latent space, there is a trade-off between reconstruction accuracy, and disentanglement which is governed by the magnitude of β. Here we seek to understand whether this holds also for a VAE gumbel-softmax distribution as well. Specifically, we are interested in the extent to which the development of cells with place-like coverage of an environment can be connected to the principle of disentanglement described in (Higgins et al., 2016).

As can be seen in Figure 9, there is indeed a large impact of the strength of the regularization term on the resulting latent space. In the case of a large value of β, each variable in the latent space learns to represent a large portion of the environment. As the regularization strength is decreased, each unit represents a smaller part of the space. However, when no regularization is applied the units learn indistinct, and largely redundant activation patterns, suggesting that the regularization term does indeed induce disentanglement. In this case, $\beta = 0.01$ corresponds to the most place cell like latent space.

We then trained another set of three models using the same latent distributions, but taking as input only the $< t >$ component of the observation space to examine whether time-like cells would emerge from each of the three models. We find that cells with an affinity for specific offsets from the start of the episode emerge in the latent space of the gumbel-softmax model, but not the other two. Instead, in the gaussian and deterministic cases, a

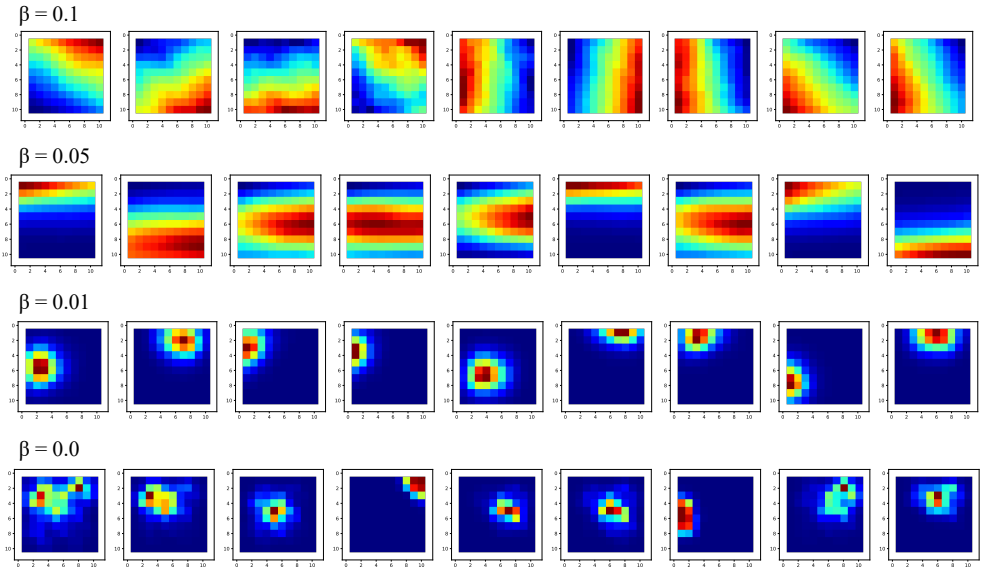

Figure 9: Example activation patterns for nine units of GTM with GS latent space models trained using different values of β for regularization loss.

single cell learns to represent duration as a scalar value, and the rest are not sensitive to the input. Example latent space activations are presented in Figure 10. We can understand this as the 1D case of the place-cell development described above. This process also generalizes to high-dimensional observational spaces, as will be demonstrated in subsequent chapters.

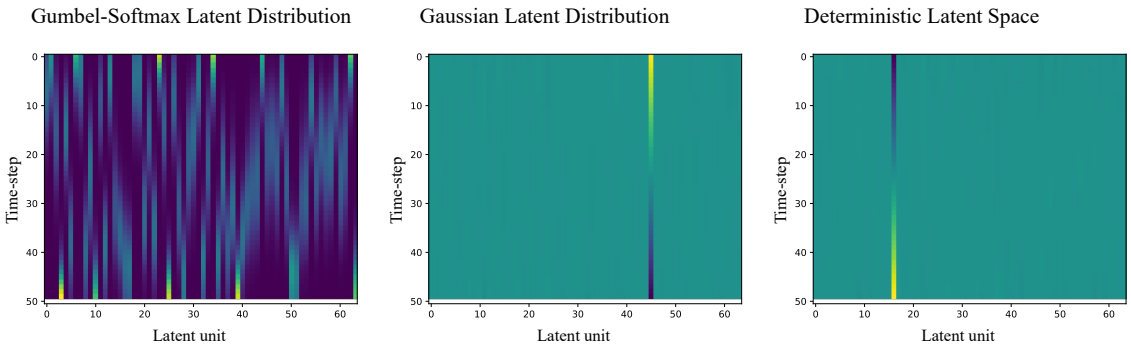

Figure 10: Representative activation patterns of the 64 units in the latent variable z by time-step in world models trained using gumbel-softmax, gaussian, and deterministic latent distributions.

We can furthermore compare the quality of the reconstructions in the temporal obser-

vation case. We find that like in the case with spatial observation data, there is a significant difference between the three models ($F(2,3897) = 70.885, p < 0.001$), with the gumbel-softmax model ($Mean = 0.004, Std = 0.009$) significantly outperforming both the gaussian ($Mean = 0.010, Std = 0.016$) and the deterministic ($Mean = 0.010, Std = 0.015$) models in terms of reconstruction quality ($p < 0.001$). See Figure 11 for a graphic representation of these results. These results might be surprising, since in all three cases the model simply needs to learn to return the same original input value. Due to the complex non-linear transformations that are part of the of the variational-autoencoder architecture however, this task is not entirely trivial.

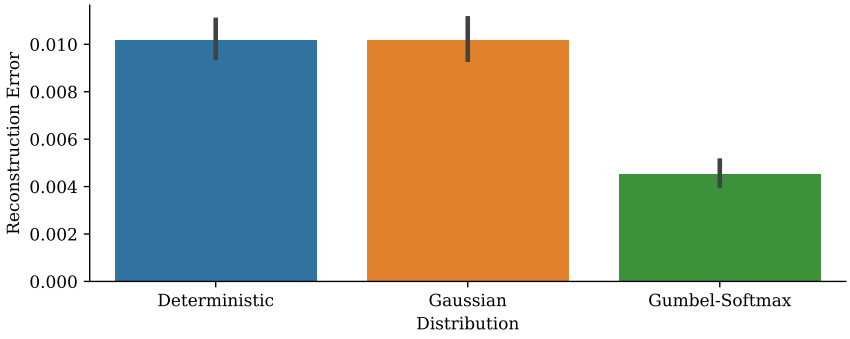

Figure 11: Reconstruction errors of three model types trained to auto-encode temporal observations. Error bars represent standard error.

Subsequent sections of this chapter will further explore the properties of a GTM trained using a gumbel-softmax distribution, with the next chapter exploring the usefulness of this representation for goal-driven navigation tasks.

II.2 Place-like Cells are Distributed based on Underlying Agent Behavior

So far, we have demonstrated that both place-like and time-like cells can come about within the latent space of a variational autoencoder with a gumbel-softmax distribution. To do so, we used a semi-random walk policy to collect the dataset used to train the model. While

such a policy is a reasonable proxy for animal foraging behavior (Viswanathan, Da Luz, Raposo, & Stanley, 2011), it does not capture the just as prevalent behavior of goal-directed navigation, or any biased movement through the space. It is known for example that in rodents performing a goal-directed navigation tasks, the place cells in the hippocampus cluster near the goal location (Hollup et al., 2001). This suggests a behavioral impact on the structure and placement of place cells.

Here we explore the extent to which different behavioral policies induce different place cell biases within the same environment in the generative temporal model introduced in the previous section. We find the behavioral bias of the agent corresponds to a bias in activation preference for the induced latent units as well, consistent with what is found in animals.

II.2.1 Evaluation Methods

In order to test for the influence of the behavioral policy on the distribution of place-like cells, we developed five separate behavioral policies, four of which each having a movement bias for the north, east, south, and west directions, respectively. The fifth policy was the same as described in the previous section. In each of the biased policies there is a 50% probability that the biased action will be taken at each time-step, and a 50% probability that an action will be selected with uniform random probability instead. Figure 12 shows the action probability distributions for each of the five policies. For each policy, we collect 1000 episodes of 50 time-steps each, and train each model using the same hyperparameters described in the previous section.

II.2.2 Results

We find significant differences in the biases of the latent spaces induced by sets of observations generated from agents with different biased behavioral policies. These results can be seen clearly visually in Figure 13. In the case of each of the biased policies, there is a greater number of place cells in the region more likely to be visited by the behavioral

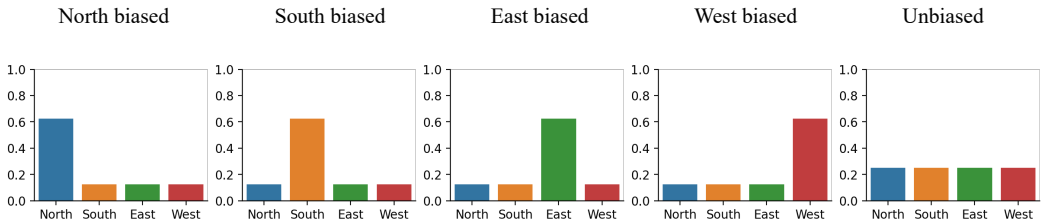

Figure 12: Action distributions for each of the five biased policies.

policy than anywhere else.

Statistical analysis reveals that there are indeed significant differences in the induced latent spaces of each of the different models. We analyze both the x and y bias in the models by taking the point in the environment of maximal sensitivity for each unit in the latent space, and performing ANOVA analysis to determine distributional differences. We find that there is a significant difference between the data types in both the x ($F(4,495) = 65.93, p < 0.001$) and y directions ($F(4,495) = 69.04, p < 0.001$).

As would be expected from the qualitative results presented in the figure, we find that in the case of biases with respect to the y axis, there are significant differences between the south and north policies from each other ($p < 0.001$), as well as between these policies and the other three ($p < 0.001$). There are no significant differences between the other three policies ($p > 0.5$). Likewise, when looking at biases with respect to the x axis, we find significant differences between the east and west policies ($p < 0.001$), as well as significant differences between each of these policies and the other three ($p < 0.001$), but no differences between the other three ($p > 0.5$).

These results suggest that there is indeed a clear bias in the preference of the units in the learned latent space of the generative temporal model. This preference is biased towards parts of the state space of the environment which are more frequently visited, and thus can benefit from greater representational capacity. Greater representational capacity then

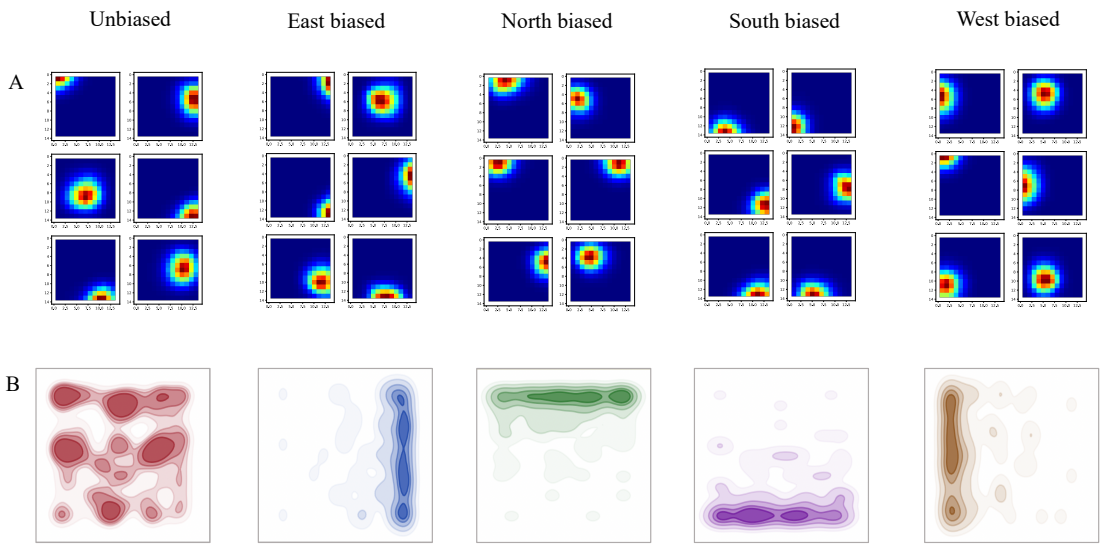

Figure 13: Activation patterns of latent units trained with a biased behavioral policy. A. Activation patterns of first six units of latent space of models trained with one of five different biased behavioral policies. B. Contour map of firing affinity for each of the five models.

corresponds to finer sensitivity to small (potentially behaviorally relevant) changes in that region of the environment. This provides one potential explanation for the similar biases seen in the formation of place in cells within the hippocampus of rodents (Hollup et al., 2001).

II.3 Internally Generated Sequences and Auto-regressive Models

In the previous sections we described how place and time cell representations can come about in generative temporal models trained to perform a simple prediction task with spatial and temporal observations. An additional property of this class of models is their ability to be used in an auto-regressive manner once trained. Concretely this means that rather than providing the z_t which was inferred from the current observation to the forward model, the z_t which the model generated at the previous time-step is used instead. If this process is continued, an entire trajectory of "imagined" observations can then be decoded from the

sequences of latent states z. This process is sometimes referred to as performing a "rollout" or "unrolling" the model, because of the recursive nature of the procedure, and these term will be used below to refer to this process.

In this section, we will demonstrate that this unrolling procedure, when performed on a fully-trained GTM, reliably produces coherent trajectories which can match the original sequences of observations fed into the model. This capability bears a strong resemblance to the phenomena of replay and preplay in the hippocampus (Foster, 2017). In both cases, sequences of latent states are spontaneously generated in a coherent trajectory in the absence of additional sensory input. Here we show that action sequences generated from the same policy used to infer the latent state can be used to "unroll" the model and generate coherent sequences of place-like cell activations that match those which would come about from exposing the model to the actual sequence of observations.

II.3.1 Evaluation Methods

In this section, we will use the same trained models from the previous section, three GTMs, each with a different latent space distribution. Instead of examining the representational quality of the z inferred from the observations, we will examine the quality of the predictions of the z generated by the forward model.

We will utilize the same 2D gridworld environment described above, but examine solely the $< x, y >$ component of the observation space. We will examine both the quantitative accuracy of each model being used in an auto-regressive manner to generate a trajectory of predicted observations, as well as perform a qualitative examination of the latent space representation during this rollout.

II.3.2 Results

We examine both the latent space representation during the rollout (z), as well as the resulting predicted observations (o^*). In both cases, we find that they track their target,

suggesting that the model is indeed capable of learning the transition dynamics of the environment. Figure 14 displays the unit activations for the inferred z and the z generated via the auto-regressive rollout. We find that these two largely match one another, suggesting that the same sequence of actions results in the same activation pattern, regardless of whether observations are being inferred directly, or the activation is the result solely of the learned recurrent dynamics of the neural network.

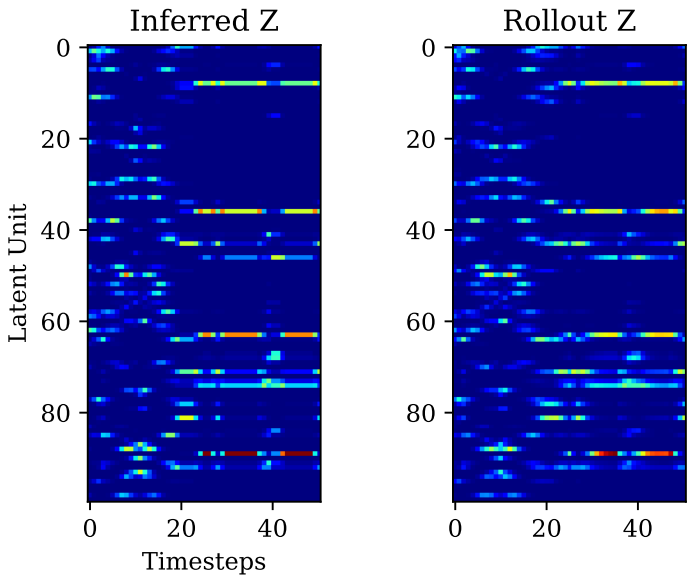

Figure 14: Inferred and generated latent variables during a single trajectory.

While a correspondence between the latent representations is useful to know, the value of main interest is the reconstruction quality of the observations of the trajectory from the latent space. In Figure 15 we compare the original observations in the trajectory to their reconstructions from the inferred z variables, as well as to the predictions of auto-regressive rollout. While there is some representational drift, we find that it is not catastrophic.

When we quantitatively compare the reconstruction errors of the three models, we find that there is a significant difference in their capacity to reconstruct the observations from the latent space induced during the auto-regressive rollout ($F(2,3897) = 126.197, p < 0.001$). We furthermore find that as was the case in reconstruction from the inferred latent space, the model trained using a gumbel-softmax latent space shows the lowest level of reconstruction

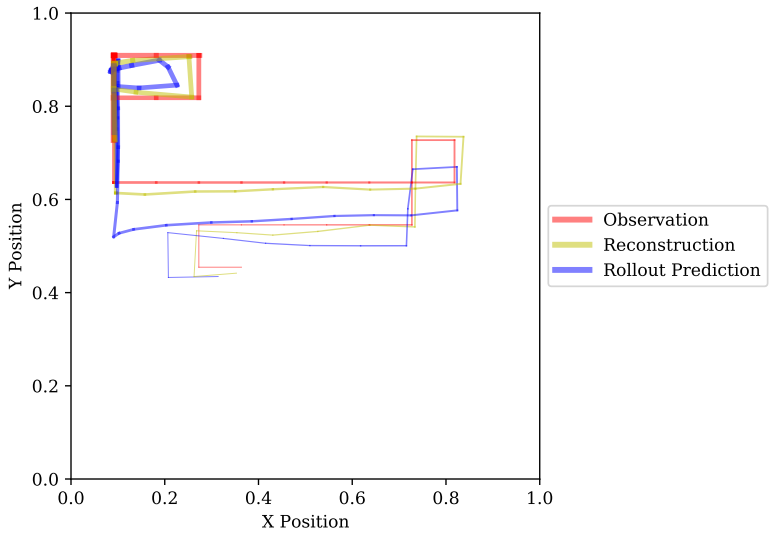

Figure 15: Comparison between ground-truth observations, their reconstructions from the inferred latent variable, and their reconstruction from the rollout of the generative model using a gumbel-softmax latent space.

error of the three when constructing from the rollout latent space as well ($p < 0.001$).

Altogether, this suggests that GTMs are capable of both learning a meaningful latent space, as well as learning a coherent forward model of the environment dynamics, which retains this coherence even when unrolled in an entirely auto-regressive manner. Later, in Chapter III, we will demonstrate the application of this model unrolling in improving the learning process during a goal-directed navigation task using the Dyna algorithm (Sutton, 1991).

II.4 Generative Temporal Models Learn Temporal Community Structure

Thus far we have demonstrated that a GTM using a gumbel-softmax latent space is capable of representations which bear similarities to both place and time cells in the hippocampus. Furthermore, we have demonstrated that models with these cell types are useful for creating coherent trajectories of experience entirely in the latent space, similar to the phenomena of

replay and preplay in the hippocampus (Foster, 2017).

Beyond the place-like appearance of these units, and the ability to generate trajectories of them, it is of interest to know whether these learned representations in and of themselves display other known properties of hippocampal representations. One property of interest is the temporal community structure which has been demonstrated in human hippocampal representations (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). This structure results in sensory perceptions which are temporally more likely to co-occur being represented more similarly within the hippocampus, regardless of the underlying sensory similarity of the observation itself. This can be thought of as a process of sensory decorrelation followed by temporal correlation.

Schapiro et al. demonstrated this phenomena in humans exposed to a series of fractal images drawn from a random walk on a graph. They demonstrated that the hippocampal representations of these stimuli were best captured by their temporal structure, rather than the properties of their visual appearance. This capability has been modeled using the successor representation (Stachenfeld et al., 2017), as well as simple feed-forward neural networks trained to perform a prediction task (Schapiro et al., 2013). Here we show that community structure comes about within a predictive model in the absence of any explicit successor learning, and in a model that is trained end-to-end to perform prediction from raw visual observations.

II.4.1 Evaluation Methods

In order to demonstrate learned community structure in the latent representations of GTMs, we utilize the same generative temporal model with a gumbel-softmax latent space described in the above section. We change however the environment being used. In the work of Schapiro et al. (2016) a series of fractal images were used as the stimuli, and rendered to the human participants according to a random walk along a graph structure. Here we use a similar series of fractal observation vectors as model input, and arrange them on a 2D

graph structure similar to the environment described earlier in this chapter. Instead of an open field however, the states in this environment are arranged along a ring structure. Fractal images were generated using the inverse-Fourier method, using a $\beta = 2.5$ (for details on this method, see Bies, Boydston, Taylor, & Sereno, 2016). See Figure 16 for an image of the graph structure along with examples of the fractal images used as stimuli to train the model.

As done in previous sections, we collect the dataset using 1000 semi-random walks through the environment of 50 steps each, and then separately train the model with the collected dataset.

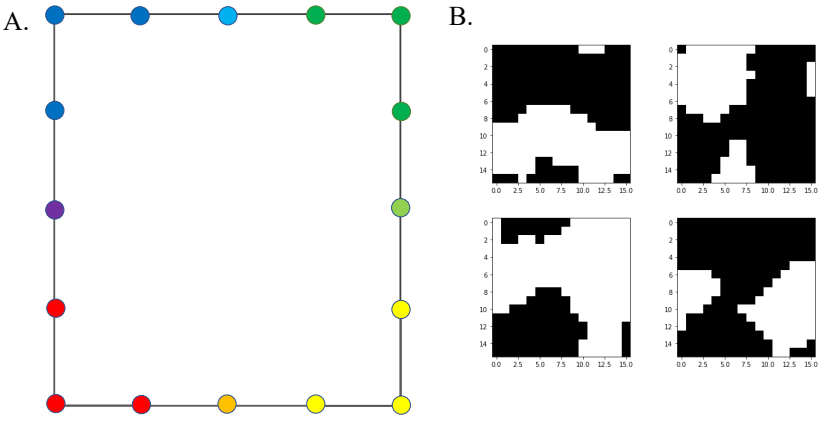

Figure 16: Diagram of a graph environment. A. Graph structure used for environment in community experiments. Nodes indicate states, and edges indicate connections between states made possible by agent action. B. Examples of fractal images used as observations in each node of the graph.

II.4.2 Results

We trained a GTM with a gumbel-softmax latent distribution for 5000 iterations, and find that it is able to perform the reconstruction and predictions tasks highly accurately, with low reconstruction and rollout losses ($Mean = 3.181, SE = 0.186$, $Mean = 7.999, SE = 0.475$). See Figure 17 for example reconstruction images of a random trajectory through the graph environment, and note that the reconstructed observations and predicted observations match

those of the true observations in structure.

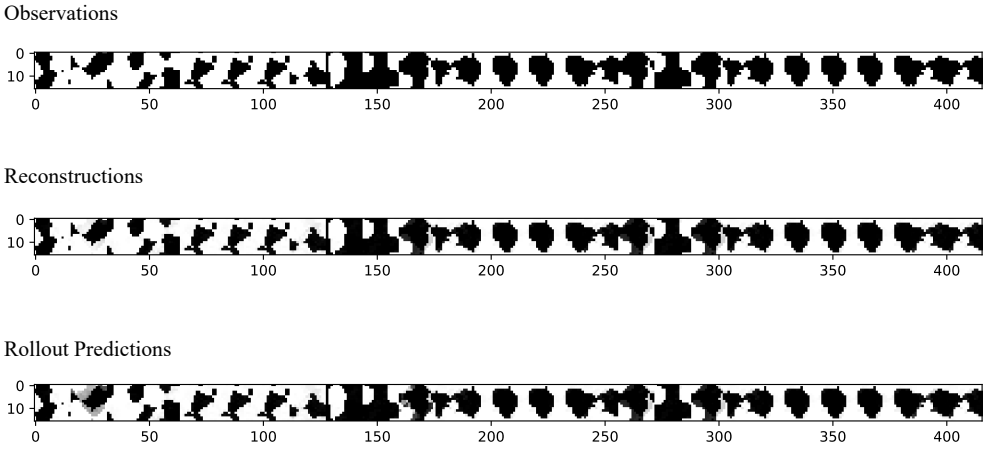

Figure 17: Fractal Rollout Examples. Comparison between ground truth fractal observations in trajectory, their reconstructions from inferred latent variable, and their reconstructions from latent variable generated as part of auto-regressive rollout.

Satisfied that our model is capable of generating coherent trajectories through this fractal graph state space, we can then turn our attention to the learned representations within this model. The first question of interest is what kinds of latent space representations have been learned from these non-visual observations. We find that in most cases the inferred z representation has learned to assign a single unit to each of the individual fractal images, resulting in a kind of 'place cell' representation, where each place is a single image. This kind of extremely sparse representation has a connection to the so-called "grandmother" cells found in the MTL (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). We can also interpret this process as pattern separation (Yassa & Stark, 2011), where each observation is encoded in a way orthogonal to the visual properties of the image. See Figure 18 for activations of each of the 16 units in the latent space.

We then turn our analysis to the learned latent representations. We perform multi-dimensional scaling on the inferred z, the generated z, and the hidden state of the recurrent

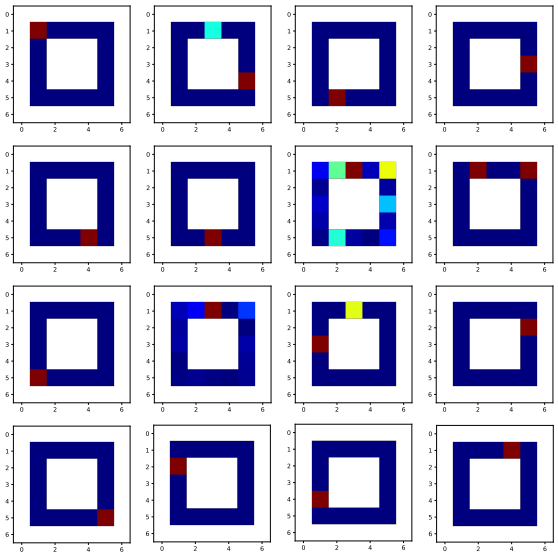

Figure 18: Latent space activations for each of the 16 units in the network.

network h. We find that while the two z representations are uncorrelated with the transition structure of the environment, the recurrent network hidden state h displays temporal community structure (Procrustes transformation results: $Error(z) = 0.862, Error(h) = 0.109$). The results of the multi-dimensional scaling procedure and Procrustes transformations are presented in Figure 19.

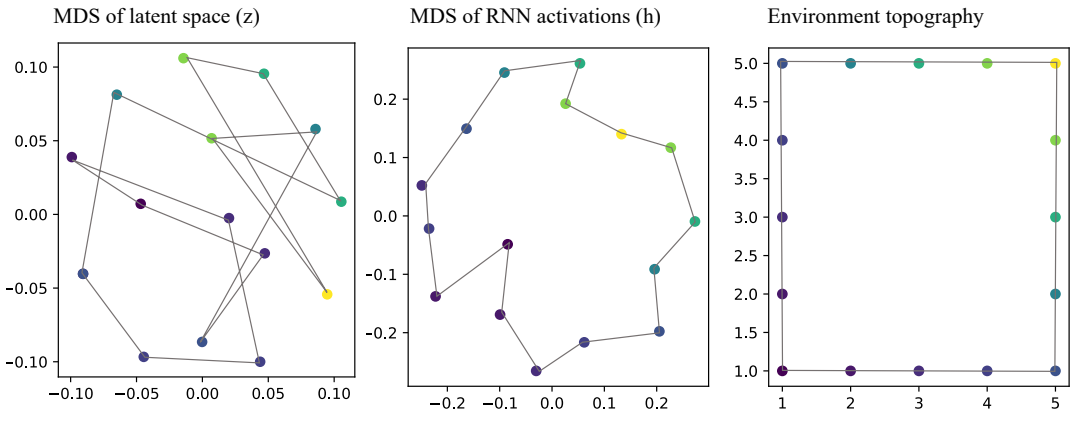

Figure 19: Multi-dimensional scaling of latent representations of learned model compared to true underlying topography of environment. The inferred latent space z shows no temporal community structure, while the hidden state of the forward model h does.

These results suggest a two-fold process in the generative temporal model. The first is that the inference from o to z involves a kind of pattern separation, where each stimuli is represented by a mutually exclusive set of representations, as seen in the activation profile of the units presented in Figure 18. Secondly, the process of computing the forward function z_{t+1} $p(z_{t+1}|a_t, h_t, z_t)$, involves a pattern completion process, whereby nearby states are represented more similarly in the h representation. Whereas this is demonstrated in (Schapiro et al., 2013) using a simple feed-forward artificial neural network, the latent space z was pre-discretized in their experiments, and only $z_{t+1} = f(z_t)$ was learned. Here we have modeled the same principle of learned temporal community, but in an end-to-end fashion, where the model receives as input the raw fractal stimuli.

II.5 Discussion

In this chapter we have demonstrated how a simple generative temporal model with a biologically-inspired latent distribution can capture a number of important properties of the hippocampal formation. These include the development of place-like and time-like cells with a behaviorally guided bias in distribution, the ability to generate long coherent latent trajectories in the absence of ongoing observational input, the presence of learned representations which reflects environmental structure, and both pattern separation and completion in the inference and generation processes respectively. Taken together, these findings suggest that GTMs with gumbel-softmax latent layers are a strong candidate model for some basic properties of hippocampal representation learning.

Our proposed model can be thought of as a kind of soft-clustering whereby observations are probabilistically grouped into states (which we refer to here and elsewhere as "state cells"), reflective of the number of units in the latent space. This bears a similarity to the recently proposed model of hippocampal representation by Mok and Love (2019). In both cases, the hippocampus can be thought of as learning a low-dimensional latent space for abstract representations of the environment an animal is in. In many cases this information

in temporal or spatial in nature (O'Keefe & Nadel, 1978), but it need not be, and can instead be information regarding other quantities.

There are of course many other models of place cell formation which have been proposed (Samsonovich & McNaughton, 1997; Erdem & Hasselmo, 2012; Whittington et al., 2019), but each of these make specific assumptions regarding the structure of the observation space, or of the environment itself. While the model proposed in this chapter is simple in comparison to previous ones, its simplicity reflects a lack of strong assumptions about the nature of the observations or the structure of the environment from which they are drawn in order for the model to operate.

The soft clustering of the gumbel-softmax latent space used in the model has the additional effect of inducing a semi-discrete state space. By discretizing the high-dimensional observations in an environment, they can then be used downstream for performing goal-driven navigation using reinforcement learning. In the following chapter, we will explore the efficacy of using the latent space z of a GTM as the state space when performing goal-directed navigation tasks using reinforcement learning.

CHAPTER III

LATENT STATES AND GOAL-DIRECTED NAVIGATION

In the previous chapter we demonstrated that generative temporal models which utilize a gumbel-softmax latent distribution can reproduce the existence of place and time cells and display temporal community structure in those representations. Since they are generative models, GTMs can also be used to generate "imagined" trajectories of experience, thus drawing a useful connection to the replay and preplay phenomena found in the place cells of the hippocampus (Foster, 2017), and an even more specific connection to the "internally generated sequences" model of Pezzulo et al. (2014). Once general structured representations like the ones described above are learned, the natural next question is to ask what downstream tasks these representations might be useful for.

In this chapter, we demonstrate that these learned latent representations are a strong candidate for providing the state space basis functions upon which value functions and policies for goal-direction action can be built. Reinforcement learning algorithms are a prime candidate for modeling such learning (Niv, 2009), and there are a number of reinforcement-learning based models of hippocampal-striatal learning. Here we focus on two specific algorithms of interest in the literature, the classic Actor-Critic algorithm, and the more recent Successor models of learning (O'Doherty et al., 2004; Stachenfeld et al., 2017). In both cases, we demonstrate that the learned latent space provided by the GTM can support the learning of optimal behavioral policies in a goal-driven navigation task more efficiently than other state spaces.

In addition to demonstrating the efficacy of the learned representations for basic reinforcement learning, we also demonstrate how this state space can be used in the context of fast-adaptation learning algorithms, where the goal location changes during the learning process. Rather than learning entirely from online experience, it is also possible to take advantage of our learned forward dynamics model to perform additional reinforcement learning updates using the Dyna algorithm (Sutton, 1991), one of the proposed models of hippocampal learning (Russek et al., 2017). We demonstrate that this results in faster learning compared to a fully online algorithm, and connect it to the replay phenomena using the model of internally generated sequence learning (Pezzulo et al., 2014).

III.1 State Cells for Actor-Critic Learning

In the previous chapter we demonstrated how place and time like cells, here referred to as "state cells" can naturally emerge from a specific kind of generative temporal model utilizing a gumbel-softmax latent distribution (GTM-GS). While the properties of these units are of interest in and of themselves, they are also of interest for their applicability to downstream tasks such as goal-directed navigation.

One key area to look for with respect to potential downstream tasks is the hippocampal-striatal axis, which is thought to be involved in memory-based decision making tasks (van der Meer, Johnson, Schmitzer-Torbert, & Redish, 2010). It is known for example that the hippocampus provides input to the striatum, and that during replay sequences place cell activations precede corresponding cell activations in ventral striatum (Lansink et al., 2009). Furthermore, there is evidence that different sub-regions of the striatum are specialized for different aspects of conditional learning, with ventral striatum involved in value estimation and dorsal striatum policy learning (O'Doherty et al., 2004). These two functions have been proposed to work together as part of an Actor-Critic learning system, a method derived from the reinforcement learning literature (Sutton & Barto, 2018). While the exact relationship between dorsal and ventral striatum has been the topic of some debate, re-

sulting in the actor-critic formulation being made more nuanced in recent years (Atallah, Lopez-Paniagua, Rudy, & O'Reilly, 2007; van der Meer et al., 2010), the underlying division, and usefulness for capturing the main empirical findings remains (Tessereau, O'Dea, Coombes, & Bast, 2020).

Here we demonstrate that the learned latent representations from the GTM model introduced in the previous chapter serves as a useful basis function for performing reinforcement learning using an actor-critic algorithm. We compare these to a set of alternative basis functions, which we will demonstrate are either hand-generated using additional knowledge of the state space and perform well, or result in slower or failed learning. In contrast, the learned state space from the GTM-GS model is generated in a task-agnostic fashion, and still results in good performance on downstream navigation tasks.

III.1.1 Methods

We utilize the same simple two-dimensional environment described in Chapter II, with the same observation space of $< x, y >$ spatial coordinates. We introduce now the additional concept of a goal within the environment. This goal can be located in any free location in the environment, and provides the agent a reward signal of $r = 1$ when the agent enters the same position as the goal. At this point, the episode is terminated, and the agent is returned to its starting position for the next episode. See Figure 20 for a diagram of this simple environment.

We are interested in understanding to what extent the learned representations of a GTM can be useful as a basis function for performing reinforcement learning. Rather than utilizing a complex Deep Neural Network (DNN) for our policy and value networks, we use simple linear functions which take the basis functions as input and compute $\pi(a|s)$ (the policy) and $V(s)$ (the value function) respectively. In these experiments we are interested in the quality of the learned representations for supporting reinforcement learning. The ability for a linear transformation to be sufficient to calculate an optimal policy and value function

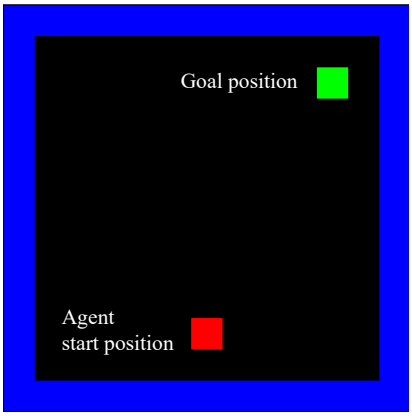

Figure 20: Diagram of two-dimensional reinforcement learning environment with single goal and single agent.

serve as a useful measure of this quantity (Bellemare et al., 2019). As such, we avoid using any deep or multi-layer neural networks in these experiments, to prevent the models from simply learning sufficient intermediate representations from a poor basis function.

Given the generative temporal models discussed above, we have a number of choices for potential basis functions which could enable reinforcement learning in an actor-critic context. While there are many options, given that we are here concerned only with linear function approximation, we only focus on basis functions which appear relevant to this context, and compare four such different functions.

The first is the raw observation space itself, $< x, y >$. While this representation is simple, it is not clear that it is amenable to linear function approximation. We include it here for completeness. The second is the canonical "one-hot" state encoding (i.e. $< 1, 0, 0... >$ for first state). Deriving this basis function requires knowledge of the total number of states in the environment, as well as a function for converting a given observation o into a state s. We know however that in the tabular case, which is what linear function learning reduces

to with one-hot observations, algorithms such as actor-critic and Q-learning are guaranteed to converge (Sutton & Barto, 2018). We derive the third and fourth state spaces from the GTM-GS model, using the distribution z, and a discretized sample from the distribution, respectively.

For each of these basis functions, we utilize a simple linear actor-critic model, and train it using data collected in an online fashion. At each time-step, the agent receives an observation from the o, and uses it to compute a state $s = f(o)$. With this state, a value function $V(s)$ and sampled action $a \sim \pi(a|s)$ are computed using linear transformations from a set of learned weight matrices. The sampled action is then used to act in the environment, producing a new observation o^* as well as a reward r. We train the model to maximize the discounted expected return $R = \sum_{t=0}^{T} \gamma^t r_t$ using the following temporal difference update rules.

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{III.1}$$

$$V(s_t)' = V(s_t) + \alpha \delta_t s_t \tag{III.2}$$

$$Q(s_t, a_t)' = Q(s_t, a_t) + \alpha \delta_t s_t \tag{III.3}$$

Where α is the learning rate and γ is the discount factor. We set these to 0.25 and 0.99 respectively. Actions are sampled by transforming the $Q(s,a)$ function into a categorical probability distribution using the softmax function, $\frac{\exp(\log(x_t/\tau))}{\sum \exp(\log(x_t/\tau))}$, and adjusting the weighting using a temperature parameter, τ, which we set to 0.01. We train each model for 200 episodes of either 150 time-steps, or the number of steps it takes to reach the goal, whichever comes first within the episode. We train all models with five different randomly selected initialization seeds.

III.1.2 Results

To assess performance, we examined the mean and median number of time-steps to reach the goal of the last 20 episodes in each of the five training runs per agent. The optimal policy in this task can reach the goal in 11 time-steps. We find that as expected, the one-hot basis function results in an agent which consistently learns an optimal policy for navigating to the goal ($Mean = 11.4, Median = 11$). In contrast, the basis function consisting of the raw observations from the environment results in an agent which is never able to arrive at the goal in any of the five random initializations ($Mean = 148.98, Median = 149$). These two results provide the extremes of a canonically good and bad basis function.

Unlike the observation basis function, the two basis functions based on the learned latent space from the generative temporal model are able to in general support learning optimal policies, though not with the same level of consistency or performance as the optimal one-hot basis function. While the resulting agent learns an optimal policy in all trials ($Mean = 11.06, Median = 11$), the "GS-Dist" model which utilizes the z softmax distribution took significantly longer to converge than the one-hot encoding. Additionally, the "GS-Sample" model, which utilizes a discrete sample from the z latent space of the model, is able to learn as quickly as the one-hot basis function, but failed to converge in one of the five runs ($Mean = 25, Median = 11$). See Figure 21 for the learning curves associated with these results.

We also recorded the estimated value in each state of the environment from each model during learning. These value estimates are presented in Figure 22. As expected from the performance results presented above, the observation basis function fails to learn a coherent value map. In contrast, the value maps for the three successful basis functions all assign value to both the goal location, as well as the path leading from the agent start location to the goal.

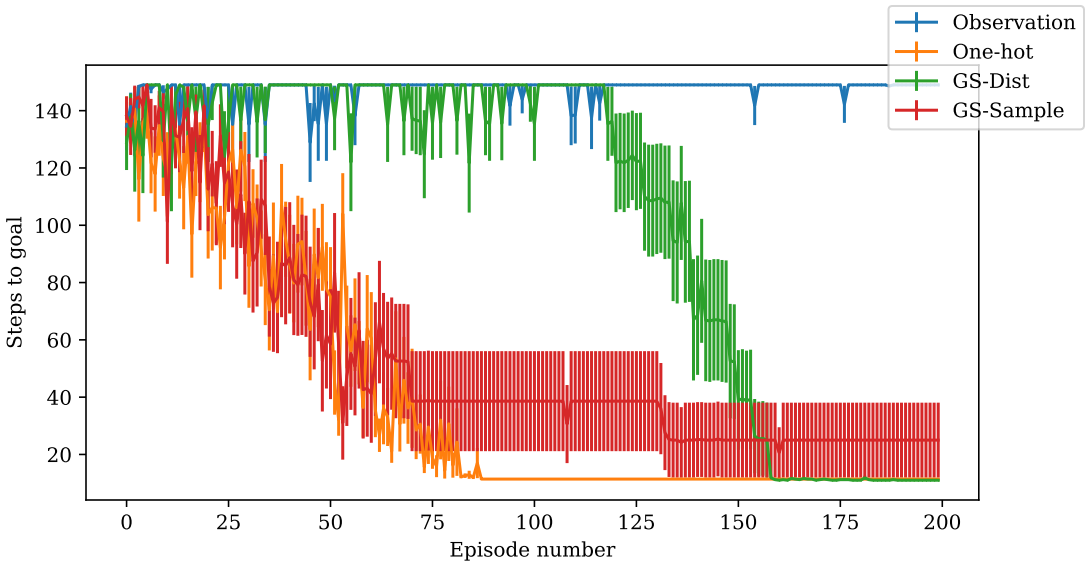

Figure 21: Actor-Critic agent mean time-steps per-episode for each basis function. Error bars represent standard error over five random initialization seeds.

III.2 State Cells for Successor Feature Learning

In the previous section we demonstrated that the learned latent space of a generative temporal model with a gumbel-softmax distribution can serve as a useful basis function for performing actor-critic learning. This was of interest due to the actor-critic model being a popular means of theoretically understanding the function of the ventral and dorsal striatum (O'Doherty et al., 2004), and the induced latent space in a GTM with a gumbel-softmax distribution bearing a strong similarity to hippocampal place cells.

Another model of interest for hippocampal-striatal learning is successor feature algorithm (Barreto et al., 2017). In this case, rather than dividing the learning problem into one with an actor and a critic, the representation of the reward $w(s)$ is dissociated from the representation of the environment dynamics $\psi(s)$. This dissociation is useful because it allows for a decoupling of the learning process between the two quantities, with the result being that a model can be trained to learn to quickly adapt to changes in either goal location (a change in $w(s)$), or to changes in environment structure or policy (a change in $\psi(s)$).

In terms of the biological realizability of this formulation, there is evidence that the

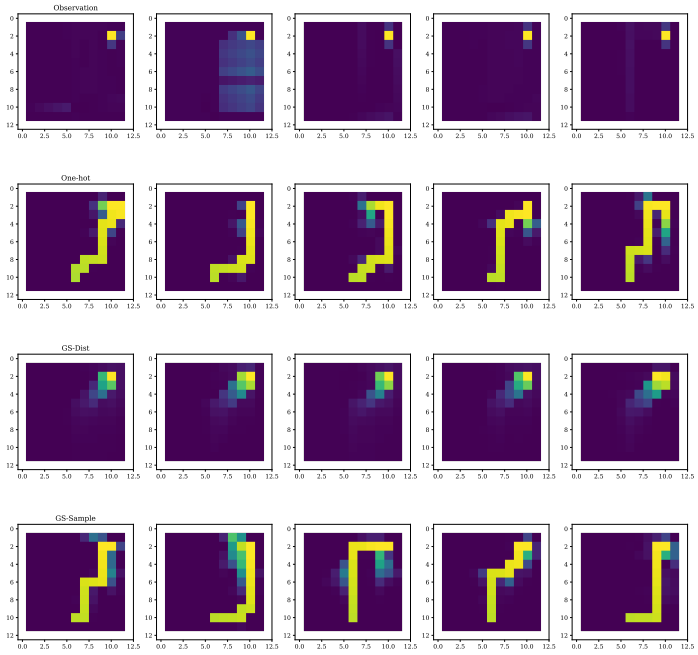

Figure 22: Example value maps for agents trained using different basis function. Shows five random initialization seeds.

"value" signal in the ventral striatum is relatively sparse, and as such could be better thought of as a reward representation $r(s)$ rather than a value estimate $V(s)$ in the traditional sense (van der Meer et al., 2010). In this case, the hippocampus would provide both the basis function s as well as the successor representation $\psi(s)$. This would correspond to CA1 output from the hippocampus (Stachenfeld et al., 2017). The ventral striatum would provide $w(s)$, and the dorsal striatum would take input from both and calculate the policy $\pi(a|s)$.

III.2.1 Evaluation Methods

We utilized a slightly modified environment compared to the previous section in order to assess the ability of agents using successor models to perform adaptation to goal position

changes during learning. Rather than an open-field square environment, we utilize a T-shape maze, in which the agent start location is at the far south end, and the goal is either in the north west or north east arm of the maze. At the beginning of a set of episodes, the goal is located at the north east arm. The agent interacts with the environment for 150 time-steps per episode, and a total of 200 episodes. At episode 100, the goal position is moved from the east arm to the west arm for the duration of the episodes. See Figure 23 for a schematic of this simple experimental design.

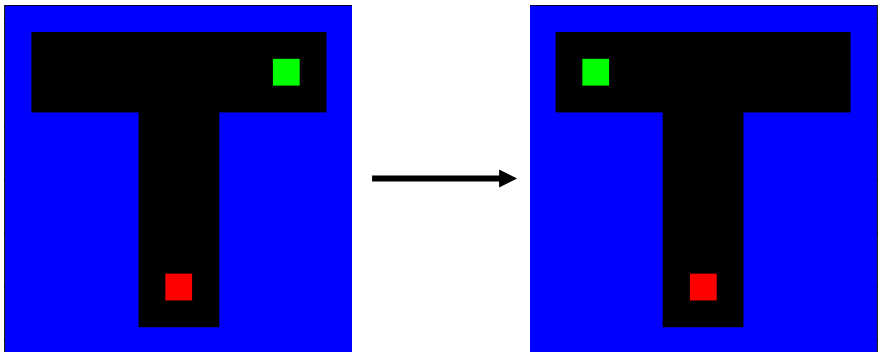

Figure 23: Diagram of experimental design for successor learning experiment. Position of goal changes halfway through training process. Blue corresponds to walls which the agent cannot pass through. Red corresponds to agent starting location. Green corresponds to goal location.

III.2.2 Modeling Methods

We expect the actor-critic algorithm to perform poorly in a context in which the goal location rapidly changes during the training process. As such, agents in this experiment are trained using both the actor-critic algorithm as a baseline, and an algorithm based on the successor representation (Dayan, 1993). In this case, the quantities being learned are $w(s')$

and $\psi(s,a)$, with the former corresponding to the learned reward function, and the latter corresponding to the learned successor representation. In both cases, the outputs are a linear function of the basis function s, which serves as input to the model. The reward function is updated using a simple learning rule as follows:

$$\delta_w = r_t - w(s) \tag{III.4}$$

$$w(s)' = w(s) + \alpha_w \delta_w \tag{III.5}$$

Where α_w corresponds to the reward learning rate. We set this to $\alpha_w = 1$ in our experiments to encourage fast adaptation to changing reward locations.

The update rule for the successor representation follows a familiar temporal-difference learning rule, with the state representation rather than the value being the propagated quantity:

$$\delta_\psi = s_t + \gamma\psi(s_{t+1}, a_{max}) - \psi(s_t, a_t) \tag{III.6}$$

$$\psi(s_t, a_t)' = \psi(s_t, a_t). + \alpha_\psi \delta_\psi \tag{III.7}$$

Where α_ψ corresponds to the successor learning rate. We set this to $\alpha_\psi = 0.2$. a_{max} corresponds to the action with the highest expected value, derived from the value function $Q(s,a) = \psi(s,a) * w(s)^T$. This equation is also used to arrive at the policy, where we convert the Q function into a categorical distribution using a softmax function, with a temperature of $\tau = 0.01$.

Due to the nature of the successor representation, only certain state representations are useful as the basis function for computing ψ with. In particular, state representations with continuous values (such as a gaussian latent space) cannot be accumulated using the above equations without changing their underlying meaning. As such, we only compare the one-

hot state representation to the "GS-Sample" representation, which is also discrete.

Unlike the open-field environment which we used in previous experiments, the T-Maze contains additional structure in the space of the environment. In order to allow our model to learn from this structure, we utilize a more complex observation space for the GTM-GS model. In addition to the $< x, y >$ components of the vector, we also include $< n, s, e, w >$ components, which each provides a normalized distance of the agent from the nearest wall in each of the four cardinal directions. These can be thought of as corresponding roughly to the activation properties of boundary cells (Lever et al., 2009). Together, the observation space is a vector of length 6, and we train a GTM-GS with a latent space of size 100.

III.2.3 Results

As expected, we find that in general the actor-critic models fail to learn an optimal policy after the goal change at episode 100, while the successor models are able to adapt to the change. In the case of the successor models, we find that both the learned latent state space and the one-hot state space are both able to serve as a basis function for an agent which learns an optimal policy for the task. As in the previous environment, an optimal policy requires 11 time-steps to reach the goal location. Both the one-hot ($Mean = 12.07, Median = 11$) and "GS-Sample" ($Mean = 11, Median = 11$) agents learn policies which reach this level by the final 20 episodes of the learning session for the agent. See Figure 24 for a visual presentation of these results.

Furthermore, we find that the learned state space results in agents which are able to even more quickly adapt to the change in goal location than the baseline agents. One potential reason for this is the distribution of states in the learned space. Whereas the one-hot encoding results in a completely uniform covering, the learned representation is biased by the states the agent encounters, where more representational resources are devoted to certain parts of the space than others. This results in a propagation of value information in a potentially more efficient manner.

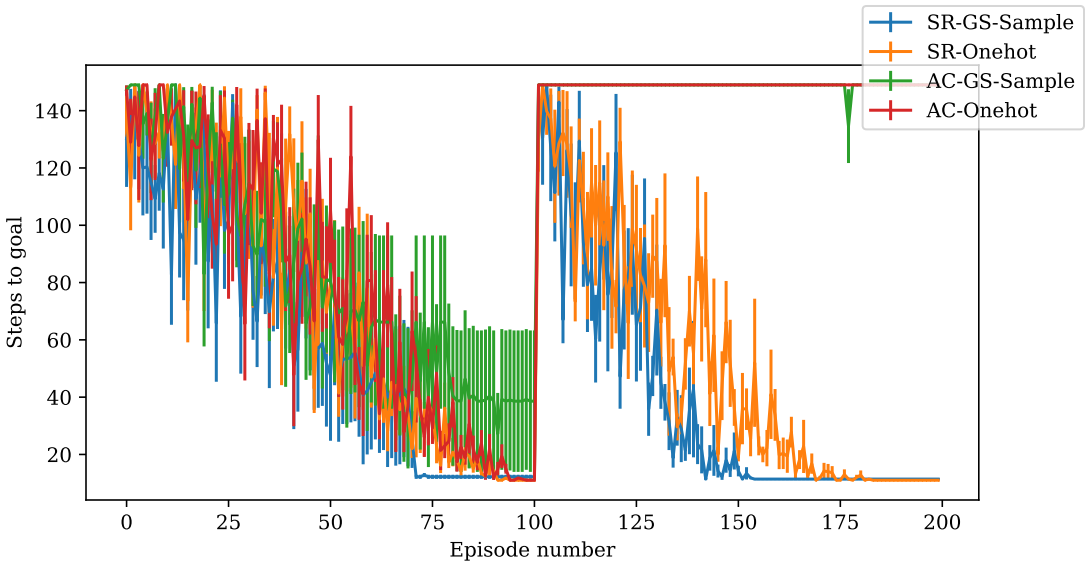

Figure 24: Mean time-steps per-episode for the two state space representations using either a successor representation or actor-critic learning algorithm. Goal location changes at episode 100. Error bars represent standard error over five random initialization seeds.

III.3 Fast Convergence with Successor Similarity Learning

In the previous section, we demonstrated that a successor-based agent using the latent space of a GTM-GS model can quickly adapt to changes in goal location during the learning process. A limitation of this model however is the need for a specific kind of state space in order for successor learning as described in (Dayan, 1993) and (Barreto et al., 2017) to perform well. This limitation comes from the fact that the reward and value functions must be linear functions in the state s and successor $\psi(s)$ spaces. This excludes the use of the gumbel-softmax distribution itself as a basis function, since it violates this requirement. As such, in the previous experiments we used a discretized sample from the latent space "GS-Sample," which effectively removes much of the useful information about the spread of a given state. This additional information can be interpreted as the model's probabilistic belief state about the agents true position in the world. We hypothesize that utilizing this extra information when computing and updating the value function would lead to faster convergence than learning exclusively from samples from the distribution.

Here we propose a modified version of the successor learning algorithm which allows for the use of successor features without the need for the strict linear function requirement, thus expanding the class of usable state space representations. We demonstrate that this algorithm enables much more rapid learning in a goal-directed navigation task by taking advantage of the full state information present in the latent space of the GTM-GS model. We do this by replacing the linear functions with cosine similarity computations, and as such refer to this new algorithm as Successor Similarity Learning (SSL).

III.3.1 Evaluation Methods

In order to examine the efficacy of the proposed SSL algorithm, we use the same environment T-Maze environment presented in the previous, but restrict the number of episodes from 200 to 100, and the number of time-steps per episode from 150 to 100. Both of these changes were done in order to provide a more challenging test of learning performance for the agents. We compare the traditional SR algorithm to our proposed SSL algorithm, using both the learned basis functions, and the pre-computed one-hot basis functions.

III.3.2 Modeling Methods

In order to arrive at the SSL algorithm, we make a few important changes to the traditional successor representation learning algorithm. First, in order to enable continuous-valued probabilistic basis functions, we replace the dot-product with a cosine similarity metric to compute the reward funciton: $r = \cos(w(s), s)$ and $V(s) = \cos(w(s), \psi(s))$. The cosine similarity between two vectors is defined as follows: $cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$. This has the property of ensuring that the reward and value functions are always bounded between 0 and 1, as long as the two vectors are positively valued. In addition, this allows us to bypass the requirement that these functions be linear combinations of the underlying quantities being compared. As such, we can take advantage of the additional information in the "GS-Dist" state space for learning. We also use a modified update rule for the reward function,

which sets $w = s$ if $r = 1$ and $w = 0$ if $w = s$ and $r = 0$. This effectively acts to cache the most recent rewarding state, and use it to compare incoming successor states $\psi(s)$ to determine value $V(s)$.

The result of these changes is that the reward and value functions now take on slightly different semantic meanings than in the case of classic successor learning. The reward function becomes a measure of how similar the current state is to the last known rewarding state. The value function becomes a measure of how likely the current state is to lead to a state like the last known rewarding state. We refer to this algorithm as Successor Similarity Learning (SSL).

III.3.3 Results

We train all model variants with five randomly initialized seeds in order to understand the performance and stability of each learning algorithm. We find that the proposed SSL algorithm with the more expressive "GS-Dist" state space ($Mean = 11.83, Median = 11.75$) outperforms the SR variants using the "GS-Sample" ($Mean = 22.14, Median = 13.3$) and "Onehot" ($Mean = 33.33, Median = 17.95$) representations. See Figure 25 for the respective learning curves.

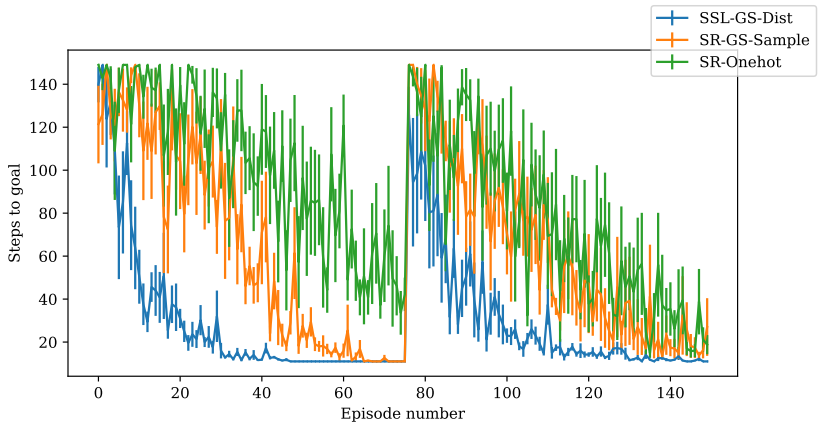

Figure 25: Mean time-steps per-episode for SSL and SR based learning algorithms with different basis functions. Error bars represent standard error over five random initialization seeds.

To ensure that the benefits gained from SSL are indeed related to greater representational capacity, and not solely from the cosine similarity metric, we also conducted an additional experiment comparing different state spaces all using agents trained with the proposed SSL algorithm. Here we find that it is indeed the combination of the more expressive state space representation "GS-Dist" with a fast-adaptation algorithm which can take advantage of it (SSL) that together confers the performance benefits we see in the first experiment ($Mean = 12.58, Median = 12.65$). In fact, we find that the performance curves for the SSL variants of "GS-Sample" ($Mean = 34.75, Median = 16.95$) and "Onehot" ($Mean = 16.19, Median = 14.75$)) state spaces are extremely similar to those of the agents trained using the SR algorithm, verifying our intuition that the benefit from SSL comes from supporting the utilization of a more expressive state space. See Figure 26 for the respective learning curves.

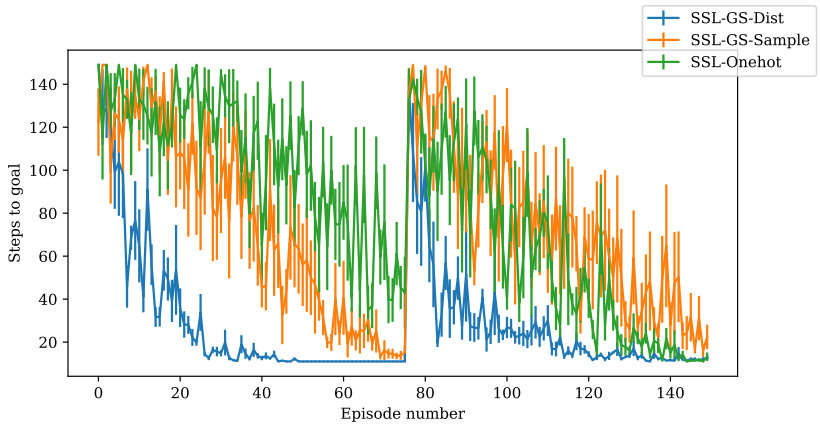

Figure 26: Mean time-steps per-episode for SSL based learning algorithms with different basis functions. Error bars represent standard error over five random initialization seeds.

III.4 Rollouts, Replay, and Dyna Learning

Thus far, we have demonstrated that the inferred latent space of a generative temporal model serves as a useful state-space for performing various kinds of reinforcement learning. By utilizing only the inferred latent space however, we are in effect throwing away half of

the trained GTM, since in doing so we are ignoring the forward model, and the trajectories through the latent space which it can generate.

A learned forward model has the potential to serve an additional purpose in the context of reinforcement learning, since it provides a model of the world which can be used to more rapidly train our value function and policy. The utilization of a learned model for this purpose in reinforcement learning is referred to as Dyna (Sutton, 1991). It has been shown to speed up learning in a number of contexts (Peng & Williams, 1993), including in biologically plausible learning using successor representations (Russek et al., 2017).

The natural analog to Dyna in the mammalian brain is the phenomena of hippocampal replay. In both cases sequences of experiences are "replayed" for the purpose of learning. In the case of hippocampal replay, this has traditionally been interpreted as serving largely a memory consolidation function (Foster, 2017). However, replay events are not random, and often involve trajectories to known goals (Pfeiffer & Foster, 2013). Additionally, the presence of replay events during rest is shown to correlate with better navigational task performance (Momennejad et al., 2018). These empirical results suggest that addition to supporting memory consolidation, there is also a significant behavior-learning component involved in replay, consistent with the role of Dyna in reinforcement learning algorithms.

In the following experiment, we build on the results in the earlier chapter showing that auto-regressively unrolling the forward model of a GTM results in the generation of a coherent trajectory of experiences. Here we show that periodically auto-regressively unrolling the model and using the pairs of latent states to update a successor representation can lead to more rapid goal-directed navigation learning than learning in a purely online manner for real experiences.

III.4.1 Evaluation Methods

In order to demonstrate the effectiveness of augmenting the online learning process with Dyna, we build on the previous navigation experiments in which successor-based agents

navigated a gridworld. Given the efficacy of the SSL algorithm introduced earlier in this chapter, we conduct these experiments using this algorithm. Here we compare multiple SSL agents, each utilizing the "GS-Dist" state space. One of the trained models updates in an online fashion, as described above, and the others update using both online experiences as well as different length trajectories (5, 10, and 20 steps) of "imagined" experiences that are the result of unrolling the GTM.

To better test for the usefulness of Dyna learning, and to make use of our more efficient successor learning algorithm, we also introduce a larger circular environment of size 21×21 which requires greater exploration on the part of the agent in order to arrive at the goal location than the previous environments. See Figure 27 for a diagram of this circular environment.

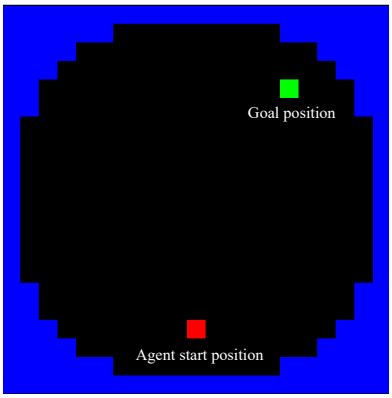

Figure 27: A large circular gridworld environment used to compare performance of purely online and Dyna-assisted learning.

III.4.2 Modeling Methods

In order to analyze the effectiveness of the Dyna procedure, we vary the length of the trajectories unrolled by the model. We hypothesize that maximum benefit from Dyna will take place with an intermediate trajectory length, since shorter trajectories may not provide much additional information, and longer trajectories may provide a corrupted learning sig-

nal, due to accumulation of errors in the unrolling process. For the agent which uses Dyna, at each time step there is some probability that an imagined trajectory will be initialized. Once initialized, the trajectory will unroll for a fixed number of time-steps, or for as long as it takes for the agent to imagine it has reached the goal location, whichever comes first. During updates within the unrolling, only the $\psi(s,a)$ is updated, and $w(s)$ is fixed, and used to determine the presence of an imagined goal, as well as to enable the computation of $Q(s,a)$, and guide the policy used during the imagined trajectory. We compare agents utilizing Dyna trajectories with a 20% probability of being activated each time-step, and unrolling the trajectory for either 5, 10, or 20 imagined time-steps.

III.4.3 Results

We find that in all cases the SSL algorithm using the GS-Dist state space are able to learn the navigation task within 100 episodes. Furthermore, we find that augmenting the online successor representation learning algorithm with an offline Dyna component enabled by unrolling the GTM is indeed able to lead to consistently faster learning on the task, leading to a near three times decrease in learning time.

Optimal performance in this task involves 17 time-steps from the agent start position to the goal. On all five runs, the agents using Dyna were able to learn to solve the navigation task optimally by the last episodes (Dyna-5: $Mean = 17.7, Median = 17.05$; Dyna-10: $Mean = 17.04, Median = 17$; Dyna-20: $Mean = 17.02, Median = 17$). In contrast, the agents without Dyna learned much slower, and less consistently ($Mean = 18.6, Median = 17.3$). See Figure 28 for a visual presentation of these results.

Comparing the number of episodes required to learn an optimal policy, we find that the Dyna-10 model, which used trajectories of length 10 when performing Dyna resulted in the fasted learning, with all five seeds converging to an optimal policy in less than 30 episodes each. In contrast, the Dyna-5 and Dyna-20 agents took over 40 episodes to converge, while the agents without any Dyna updates took over 70 episodes before all five agents converged

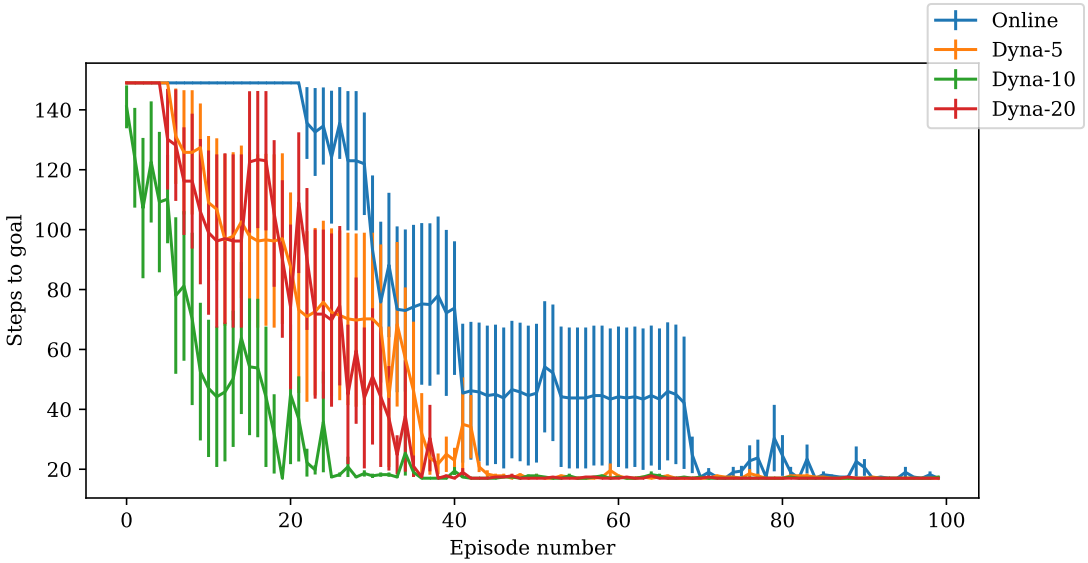

Figure 28: Mean time-steps per-episode for a fully online learning algorithm, and an online algorithm augmented with various rollout lengths of Dyna. Error bars represent standard error over five random initialization seeds.

to the optimal policy. These results confirm that Dyna can indeed greatly increase the learning process in a navigation task. It offers one strong possibility for explaining how it is that animals are able to learn navigation tasks in a few numbers of exposures to the environment or goal (few-shot learning).

III.5 Discussion

One important question about the learned representations of the hippocampus is their application to downstream tasks such as spatial navigation. In this chapter we have demonstrated that the latent space learned by a GTM-GS model can serve as a powerful state space basis function for performing different kinds of biologically plausible reinforcement learning.

We demonstrated the efficacy of these models using two canonical reinforcement learning algorithms thought to be biologically plausible, the actor-critic algorithm of striatal learning (O'Doherty et al., 2004) and successor representations algorithms (Dayan, 1993; Stachenfeld et al., 2017). In both cases, we demonstrated that the learned latent space is competitive with a pre-computed discretized latent space in terms of algorithm performance

when training an agent to perform goal-driven navigation tasks.

Beyond online model-free reinforcement learning, the forward model of the GTM provides a means of performing additional "imagined" learning using the Dyna algorithm, which we have demonstrated decreases convergence time. In addition to providing empirical benefits, Dyna is closely related to the process of internally generated sequences of place cell activations in the hippocampus found during animals at various times (Foster, 2017; Pezzulo et al., 2017). It has been hypothesized and theoretically demonstrated that this replay behavior serves to aid in learning (Russek et al., 2017), and here we provide additional theoretical evidence that this is indeed the case. The interpretation of replay and preplay within a Dyna framework is also just one of many possibilities. It has also been theoretically modeled as part of an explicit model-based planning scheme (Erdem & Hasselmo, 2012), rather than as an augmentation to model-free learning as is done in Dyna.

Our work in this chapter can be seen to complement that of (Russek et al., 2017). However, like the results presented in Chapter II related to (Schapiro et al., 2013), here we present results which build on previous work, but extend it to an end-to-end model. Whereas Russek et al. used exclusively a "one-hot" encoded state space, here we demonstrate that a state space that is learned from raw observations can be used for successor and actor-critic learning. In subsequent chapters, we will further extend this principle of demonstrating our findings in more ecologically valid settings, as we extend from simple observation spaces and Euclidean environments to high-dimensional visually realistic observations drawn from naturalistic fractal environments.

CHAPTER IV

CONTENT GENERALIZATION AND DUAL STREAM WORLD MODELS

In the previous chapters, we demonstrated that a simple generative temporal model can be used to learn a structured latent space which both displays a number of properties of hippocampal cells, while also serving as a useful basis function for performing downstream navigation tasks. Despite the demonstrated capabilities of this model, it is limited as a convincing model of the medial temporal lobe in a number of important ways. Firstly, all of the observations used were relatively low-dimensional, and in the case of many experiments, already contained relevant spatial information explicitly provided. Secondly, we trained a single model per environment, and demonstrated no capacity for generalization between environments. Thirdly, the perspective of the agent's observations and actions was allocentric, as opposed to egocentric, which is the reference frame which all embodied mammals actually utilize.

In this chapter, we seek to extend our generative temporal model in a number of important ways in order to achieve content-generalization, the ability to adapt to changes in the content of an environment, while the structure remains the same. In order to address this important capacity, we turn back to our original intention set out in the introduction, which was to provide a full model of the medial temporal lobe, taking inspiration from what we have referred to as the "language metaphor." If we were to interpret the previous model from the perspective of the metaphor, we would say that the simple generative

temporal model described in the previous chapters learns something akin to a highly pictographic language, where the signifier and the signified are intermingled together. From the perspective of memory and navigation, this corresponds to the what (content) and where (context) information being effectively fused into a single z representation. In the case where there is only low-dimensional spatial or temporal information in a signal, this is not an issue, since this fused representation reduces to a mostly where-based representation. Also, in cases where there is only a single environment with a fixed structure and set of objects of interest, then a "fused" model such as the one described above could be considered sufficient.

Of course, animals skillfully navigate not just one fixed environment, but any number of environments, which might vary in content and structure over time. They also sense the world through a series of sensory organs which provide a high-dimensional information signal. Issues for a simple generative temporal model arise when the underlying environment and observations which we are attempting to learn are higher-dimensional, contain non-spatial and spatial information, or vary over time. A canonical example of this situation is everyday egocentric narrative experience. In such cases, we take a series of actions, and experience a series of things in different places at different times. Modeling each moment of this stream using a single z, and then attempting to learn a forward model of these dynamics becomes an extremely daunting task. Especially when we would like to use the same model to make sense both of my experience making breakfast in my home, as well as the experience of making breakfast at a friend's house.

In this chapter we will introduce and validate a novel generative temporal model which we refer to as a Dual Stream World Model (DSWM). The main contribution of this model is that like other recent biologically-inspired GTMs, such as GTM-SM (Fraccaro et al., 2018), MBP (Wayne et al., 2018), and TEM (Whittington et al., 2019), it utilizes both a differentiable memory store, as well as a separation of what and where variables. Unlike each of these other models, it does so using general-purpose neural network building blocks,

which allow for it learn the dynamics of a variety of different environments with observational spaces ranging from simple vectors to high-dimensional visually realistic egocentric observations.

Concretely, this involves splitting the formerly "fused" latent state space into separate "definition" z and the "word" s representations. These two representations are then used together in a differentiable neural dictionary to enable storage and retrieval of experiences within an episode of learning. Instead of learning a dynamics model over both representations, we only learn the dynamics over the "words" s, which are inherently lower-dimensional and simpler to model. As we will show, this also enables generalization between environments with the same structure, but different objects or content within them. Taken together, this model can be seen as a complete implementation of the "language metaphor" and of an experience construction system described by Hassabis and Maguire (2009).

In this chapter, we will introduce the Dual Stream World Model (DSWM), and demonstrate how the separation of the latent space into a learned 'what' component z and a learned 'where' component s allows the model to learn the dynamics of complex environments with high-dimensional observations, and how this enables generalization between environments with similar structure. We will then demonstrate how the learned s latent space is a useful low-dimensional state space for performing goal-directed navigation. Next we demonstrate how this also allows for learning in egocentric observation spaces, and how the learned state space s in egocentric environments can also be used for performing goal-directed navigation in a visually complex 3D environment.

IV.1 Learning Content Agnostic Latent Representations

In order to extend the generative temporal model introduced previously, we make two main additions. The first is split the single encoding stream into two separate streams, each encoding the incoming observations. As such, instead of a single latent space z, DSWM

uses two latent spaces z and s, with the former representing 'what' information, and the latter representing 'where' information. This has a direct connection to the LEC and MEC regions of the MTL, which are hypothesized to convey content and context information downstream into the hippocampus (Deshmukh & Knierim, 2011; Hafting et al., 2005).

The second addition to the generative model is a mechanism by which this content and context information can be bound together and later separated in order to enable storage and retrieval of experiences within an episode for an agent. Here we use a simple differentiable neural dictionary (DND) module within the neural network (Pritzel et al., 2017). This DND is used to store and retrieve 'what' variables z using the 'where' s variables as the lookup keys. The DND consists of a list of these s, z pairs. The DSWM also consists of a forward model which is trained to learn the transition dynamics of only the 'where' variable $s_{t+1} \sim p(s_{t+1}|s_t, h_t, a_t)$. Doing so allows us to use different distributions for z and s, which can vary in both size as well as kind of distribution. We can also use different loss functions to train these two kinds latent spaces. See Figure 29 for a diagram of the complete DSWM and its three main components, a content and context encoder, a context forward model, and an associative look-up dictionary.

Key to the success of this model is that we can allocate representational capacity differently between the two latent variables. In the case of high-dimensional observations such as visual information, it is desirable to allocate a larger representational capacity to z. We can do so while maintaining a lower-dimensional latent space s which reflects the lower-dimensional transition dynamics of the environments. Consider for example a human walking around in a one-block park area, containing a few trees, sidewalks, and benches. While the sensory experience at any given time might be extremely rich, and require a complex latent space to represent, encoding one's location within the park is relatively straightforward, and requires significantly less representational capacity. Furthermore, the transition dynamics governing one's abstract location are much simpler than those governing exactly what one might see next after turning 90 degrees to the right, for example.

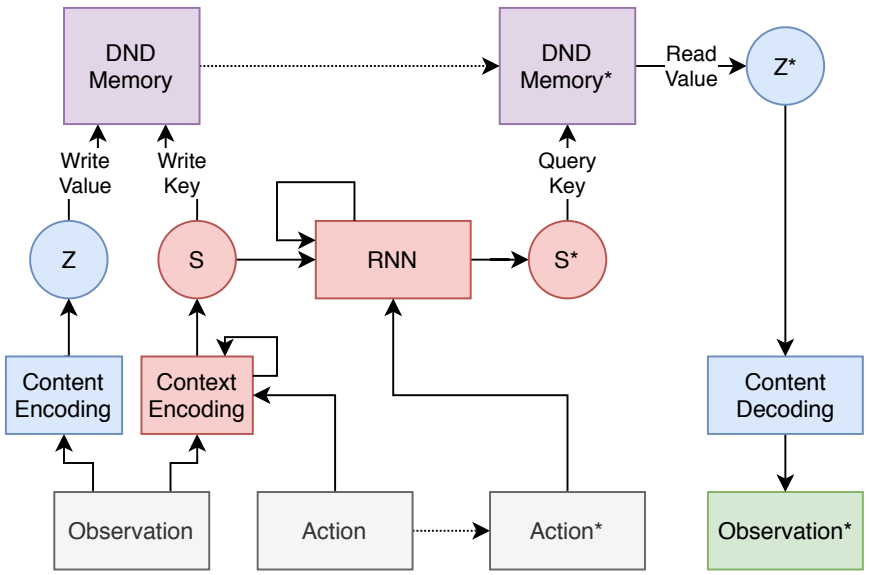

Figure 29: Diagram of the Dual Stream World Model. Blue represents content information. Red represents context information. Purple represents joint content and content information. White represents model inputs. Green represents model outputs. Nodes marked with a $*$ indicate information at the next time step of the simulation.

Key to this dissociation is the loss functions used to train each of the latent spaces of the DSWM. Here we will use the same loss function for the z space as before, a simple reconstruction loss paired with a regularization term to promote the disentanglement of representations. There are multiple candidates for losses which can be used to train s. Here we choose to use the ability to decode the position and orientation of the agent within the environment to derive the loss function used to train the s representation.

This model can be seen as an instantiation of the memory indexing theory of (Teyler & DiScenna, 1986). In this case, z represents the state of the cortex, and s serves as an index for that state. Rather than hand-designing an index to be used, we learn the index using a latent space which contains sufficient statistics which can be used to derive spatial information about the agent's location within an environment. Importantly, while we use spatial information as a training signal to the model, this information is not available during test time.

In this section we will demonstrate that the learned index s shows similarity to place cells when trained in a series of maze environments with higher-dimensional visual obser-

vations. We will demonstrate that a DSWM outperforms a single stream world model in a trajectory prediction task when the agent is exposed to environments with novel visual properties (content information). In addition, we will demonstrate that all of the relevant properties of the generative temporal models discussed in Chapter II have been retained in the DSWM.

IV.1.1 Evaluation Methods

In order to better test the capabilities of the DSWM, we use a new set of environments with more complex topographic structure, higher-dimensional observations, and greater variability in appearance. Like in previous chapters, each environment is instantiated as a 2D gridworld, from which the agent can move in the four cardinal directions, but cannot move through walls. Each environment is composed of 11×11 units. Instead of a simple observation space of spatial coordinates or distances from walls, here we use images drawn from a sliding window over a larger visual pattern map juxtaposed on the environment.

These "pattern maps" are generated by randomly selecting either a green or red pixel to be placed in each unit of the environment that does not contain a wall. This can be thought of as akin to changing the wallpaper or carpets within the same floor of a building, the content changes, but the structure remains the same. In order to derive an observation, the agent is provided with a 5×5 unit window around its current location, which displays the content of the pattern map as well as the location of any walls within the environment, which are represented as black squares. Each observation is presented to the agent as a $5 \times 5 \times 3$ image.

We use environments with four different topographies. These consist of an open area *OpenMaze*, an environment with four connected rooms *RoomsMaze*, an environment with a symmetrical obstacle in the middle *RingMaze*, and an environment with four symmetrical obstacles *HallwayMaze*. For each of these topographies, we generate 1000 different fractal maps to provide a variety of different objects for the agent to observe. See Fig-

ure 30 for examples of these environment topographies, the pattern maps, and the derived observations.

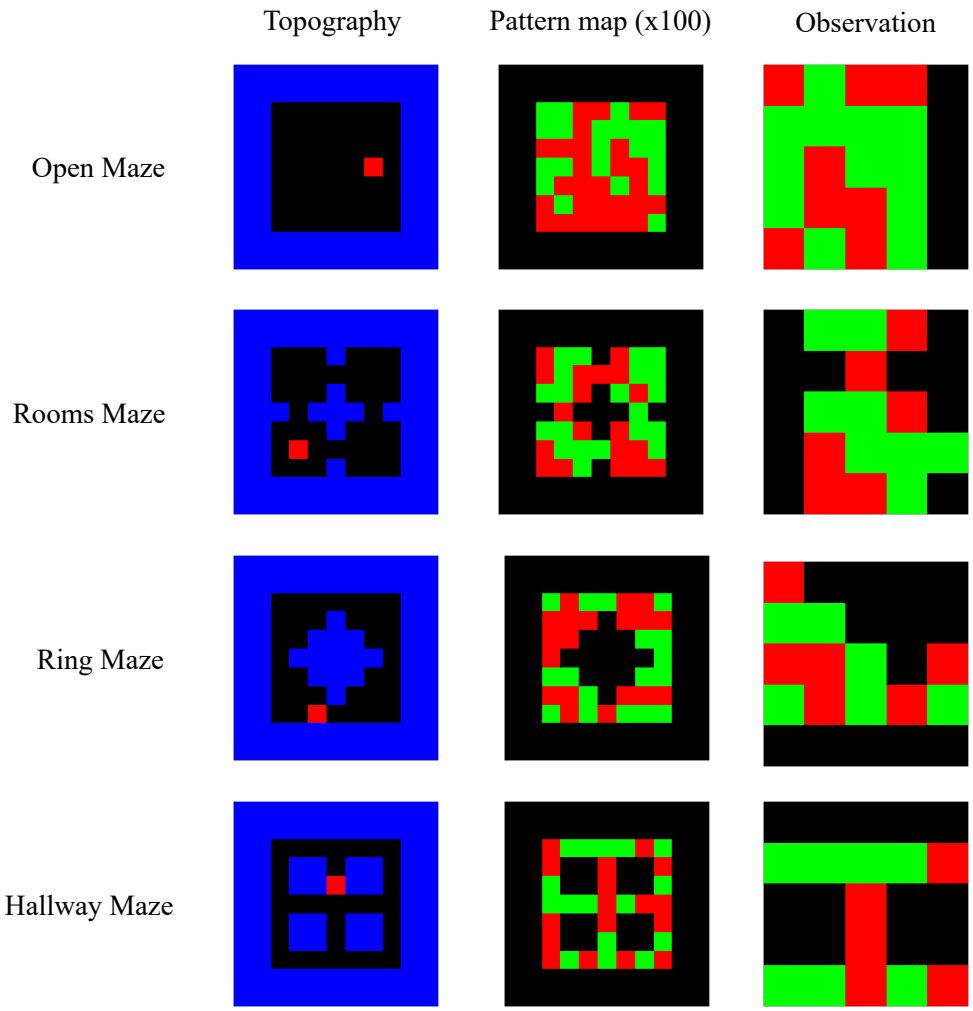

Figure 30: Four variable content environments with different topographies. Left: environment topography. Blue corresponds to walls. Red corresponds to agent position. Middle: Randomly generated pattern image used to derive observations based on agent location. Right: Agent observations provide a 5×5 window around the agent position.

The datasets used to train each model was collected by running a semi-random behavioral policy for 1000 episodes of 50 steps each. In this case, we create four different

datasets, one for each unique topography, and randomly select one of 1000 pattern maps to use for each episode.

IV.1.2 Modeling Methods

The DSWM consists of four main components. A content auto-encoder, a context encoder, a forward model, and a differentiable neural dictionary. Concretely, we utilize a variational encoder with a gumbel-softmax distribution for both the context and content components (Jang et al., 2016). For the forward model, we utilize the same gated recurrent unit (GRU) from the simpler GTM, and use as input both the latent 'where' state s as well as the current action a. The differentiable neural dictionary (DND) is similar to that used by Pritzel et al. and uses the latent context variables as keys, and the latent content variables as values. The lookup process uses cosine similarity between a query key and the stored keys to determine a similarity score. The top five stored values are then weighted by their similarity scores using a softmax function to derive the retrieved z.

For any given time-step of simulation, the following series of steps take place. First a new observation is observed from the environment. Next, that observation o_t is used to infer the latent 'where' s_t and 'what' z_t variables. The inferred 'where' variable s_t and 'what' variable z_t are then stored together as a key-value pair in the DND M_t. The forward model is then unrolled using both the next action a_t the agent takes, and the current inferred 'where' variable s_t to produce a new 'where' variable s_{t+1} that is used to query the memory to read a new 'what' variable z_{t+1}, which is decoded into a predicted observation o_{t+1}. This process is described in Figure 29. Concretely this corresponds to an inference and a generation phase, which are described below.

Inference phase:

$$z_t \sim p_{enc}(z_t|o_t) \tag{IV.1}$$

$$s_t \sim p_{enc}(s_t|o_t) \tag{IV.2}$$

$$M_t = f_{write}(M_{t-1}, s_t, z_t) \tag{IV.3}$$

$$h_t = f_{forward}(s_t, a_t, h_{t-1}) \tag{IV.4}$$

Generation phase:

$$s_{t+1} \sim q_{forward}(s_{t+1}|s_t, a_t, h_t) \tag{IV.5}$$

$$z_{t+1} \sim q_{read}(z_{t+1}|M_t, s_{t+1}) \tag{IV.6}$$

$$o_{t+1} = f_{decode}(z_{t+1}) \tag{IV.7}$$

The model is then trained to minimize four objectives. Content reconstruction error: mean squared error between original and predicted observations. Spatial information decoding: mean squared error between true and predicted position along with KL divergence between predicted and true orientation, where applicable. Sequence coherence: KL divergence between inferred and generated 'where' variables. Latent variable regularization: the negative entropy of the 'what' and 'where' variable distributions, which acts as a regularization term.

$$L_{Obs} = \frac{1}{n}\sum_{n=1}^{N} |o_t^q - o_t^p|^2 \tag{IV.8}$$

$$L_{Pos} = \frac{1}{n}\sum_{n=1}^{N} |pos_t^q - pos_t^p|^2 \tag{IV.9}$$

$$L_{Ori} = D_{KL}(p(ori_t|o_t)||q(ori_t|s_t)) \tag{IV.10}$$

$$L_S = D_{KL}(p(s_t|o, s_{t-1})||q(s_{t+1}|s_t, a_t)) \tag{IV.11}$$

$$L_{Total} = L_{Obs} + L_{Pos} + L_{Ori} + L_S - \beta_s H(s) - \beta_z H(z) \tag{IV.12}$$

In the DSWM, we compose the z latent space using eight gumbel-softmax distributions of size 16 each for a total of 128 units. We compose the s latent space with a single gumbel-softmax distribution of size 49. In the WORLD baseline models (referred to as GTM-SM in previous chapters), we use the same size latent space for z. In both model types we use 256 units for the GRU hidden layer. We train each model using mini-batches of three trajectories, each of length 50 for 10000 training iterations using a learning rate of $\alpha = 5e-4$ and regularization terms $\beta_s = 0.01$ and $\beta_z = 0.0001$.

IV.1.3 Results

The most immediate quantity to compare between the WORLD model and the DSWM is the reconstruction accuracy of the model's auto-regressive rollouts in a novel environment. It is here that we expect that the additional complexity of the DSWM over the WORLD will allow for better predictions. We use a separate set of five held-out pattern maps to create five novel environments for each of the four different topographies to use as a test set. We collect predictions based on first allowing the agent to run for 30 time-steps within an environment, and then auto-regressively predicting the next 20 observations.

We find that for all tested environments the DSWM is able to more accurately predict sequences of observations in these novel environments which were not part of the dataset

used for training (DSWM $Mean = 6.025, Std = 6.573$, WORLD $Mean = 8.752, Std = 4.594$, $p < 0.001$). See Figure 31 for the individual losses within each environment. These results suggest that DSWM does indeed have additional generalization capacity compared to the WORLD model.

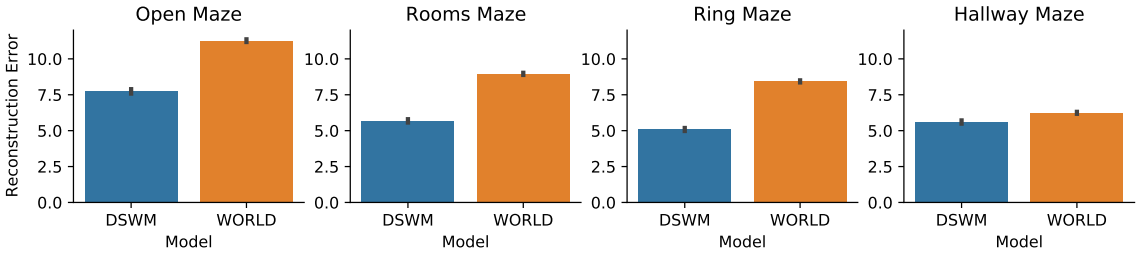

Figure 31: Reconstruction errors from rollouts of both World and DSWM models in four different topographical environments. Error bars represent standard error. In all environments, DSWM is able to significantly better predict trajectories of future observations than the WORLD model.

We can also inspect qualitatively the predictions produced by each model. Example auto-regressive rollouts from the two models are presented in Figure 32. We can see that while both models are reasonably accurate at predicting the structure of the environment, the WORLD model fails to predict the correct content in novel environments, whereas the DSWM is able to predict both the content and structure. As such, this provides evidence that the DSWM is able to adapt to an environment's novel visual content as long as it retains a familiar topographical structure.

We next examined the learned latent representations within the DSWM, asking whether the learned representation of the s latent space reflects place-like firing properties. Given the loss function which induces a representation from which the agent position can be decoded, we would expect that such a representation would arise. This is not guaranteed however, since the observations being encoded into s contain both spatial and non-spatial information, and in some cases the non-spatial information dominates the observation.

To answer this question, we can qualitatively examine the learned representations of s mapped onto the environment topography. The firing affinity of cells within the learned

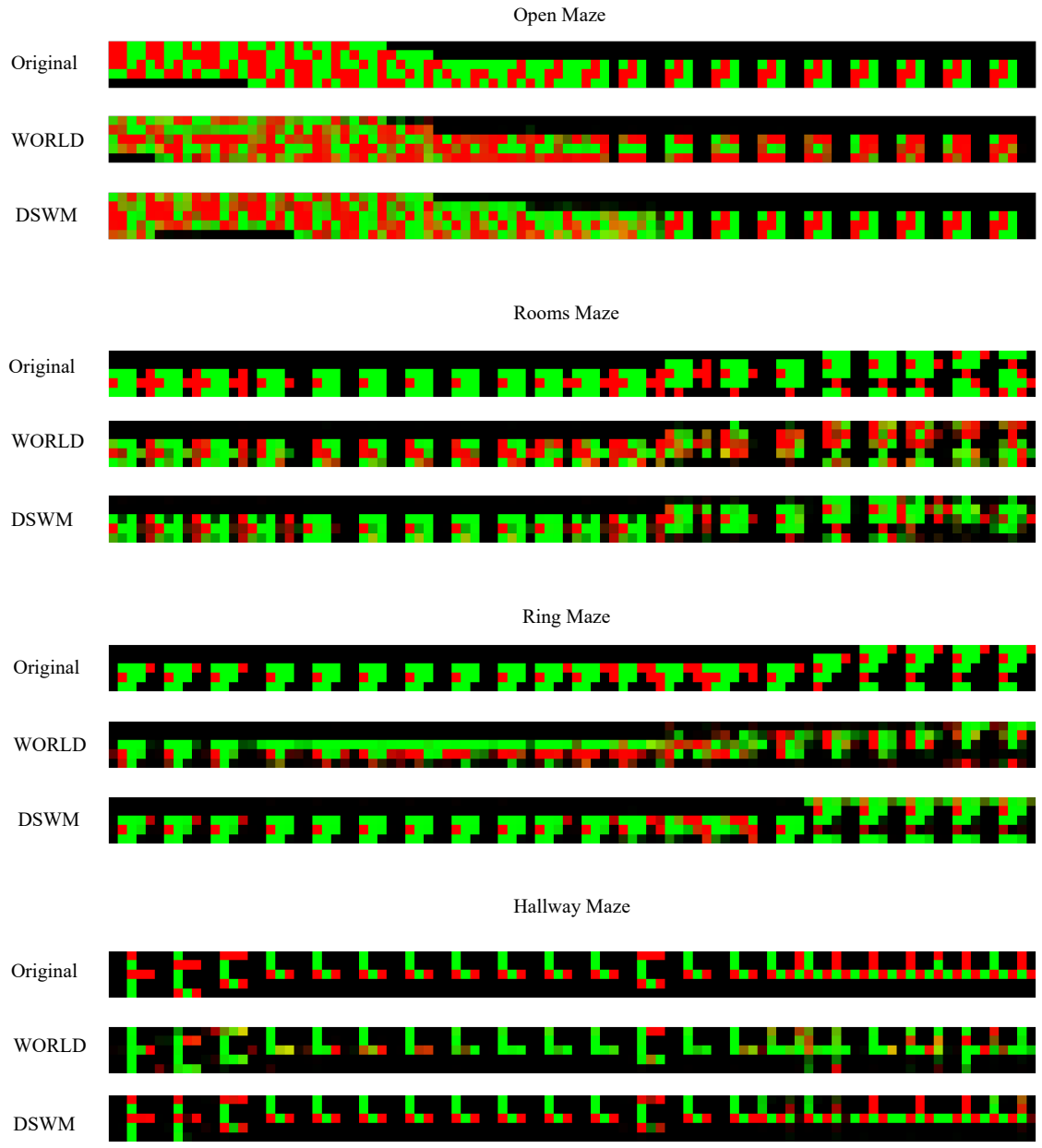

Figure 32: Examples of reconstructed observations from rollouts of both World and DSWM models in four different topographical environments. Environments use pattern map reserved for testing, and not seen during training. In all environments, DSWM is able to better predict the true trajectory of future observations within the novel environment.

representation is presented in Figure 33. We find that the representations can be best described as indeed being place-like in their firing affinities. In particular, we find that the inferred s_t units are highly spatially local, whereas the s_{t+1} units generated by the forward model have wider spatial selectivity. This can be seen as connected to the dentate gyrus /

CA3 and CA1 regions of the hippocampus, with the two regions being involved in either latent state inference (pattern separation) or generation (pattern completion).

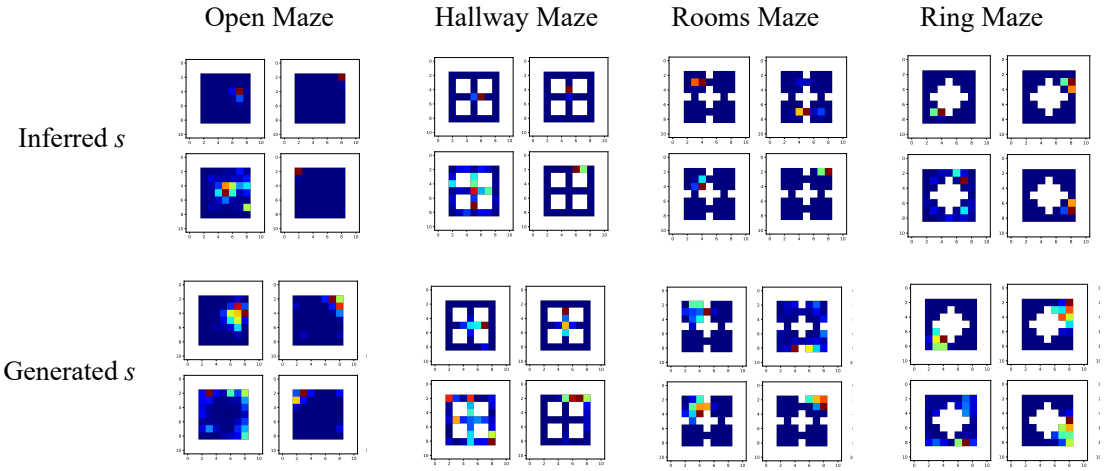

Figure 33: Examples of activations of first four units of inferred and generated s from DSWM model in each of the four different environment topographies.

IV.2 Goal-directed Navigation in Environments with Novel Content

Given the evidence that the DSWM is able to adapt to novel environment content when being used to generate imagined trajectories, the next question we can ask is whether it can do the same when serving as a state-space for performing goal-directed navigation. In this section, we use the learned latent spaces from the trained models in the previous section as basis functions for performing reinforcement learning as done in Chapter III. Instead of performing navigation within the same environment used for training, we use a set of environments with the same structural topographies, but different pattern maps, providing different 'content' information within each observation of the environment.

Here we compare the DSWM context latent space s to that of the WORLD model latent space z, as well as to a onehot-encoding baseline. We find that the DSWM latent space provides a basis function for learning which results in both faster learning and overall better

performance than either the latent space form the WORLD model or the pre-computed onehot encoding. Furthermore, we find that the DSWM can be used to perform additional offline learning using the DYNA algorithm to further improve learning performance.

IV.2.1 Evaluation Methods

In order to examine the goal-directed navigational abilities of agents using the learned state spaces, we use the same test environments from the previous section. We employ a goal-directed navigation task which involves the agent finding a hidden goal in one of the states of the environment. Halfway through a given training session, in this case, 50 episodes into training, the location of the goal changes to a new location. We use the same set of goal locations for all topographies in order to allow for the consistent comparison between results. As such, in all environments except for the "Rooms Maze," there exists the same optimal policy for each goal. Due to the nature of the topography of the "Rooms Maze" environments, this optimal policy is slightly different, and involves dealing with the bottleneck between rooms. See Figure 34 for a visual representation of the goal locations before and after the change for each environment topography.

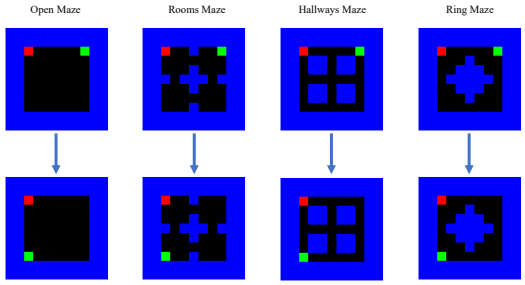

Figure 34: Four different environment topographies, each showing the initial goal location for the first 50 episodes (top) and the second goal location for the following 50 episodes (bottom). Red corresponds to agent start location. Blue corresponds to wall/obstacle location. Green corresponds to goal location.

All agents are trained using the Successor Similarity Learning (SSL) algorithm, introduced in Chapter III. All agents are trained using a learning rate of $\alpha = 0.1$. Agents are

trained for 100 episodes each, with a maximum of 100 steps per episode using an environment from the test set of pattern maps. Each training session is repeated with five separate agent initialization seeds in order to better understand learning dynamics.

IV.2.2 Results

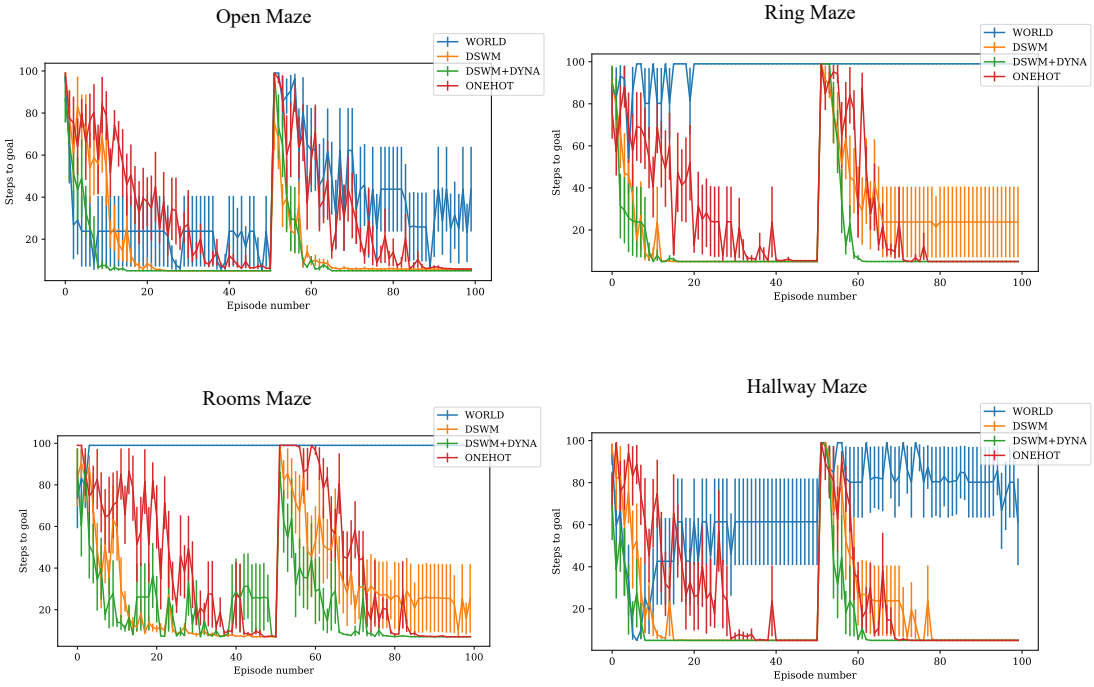

Figure 35: Learning curves in goal-directed navigation task for each of the four unique environmental topographies. Each curve represents the average of five separate initialization seeds for the agent. Error bars represent standard error.

We find that for all four environments, the state space derived from the DSWM model latent space s is able to match or outperform both the state space derived from the WORLD model latent space s as well as the one-hot state space encoding. See Figure 35 for the relevant learning curves for each agent. See also Table 1 for the reported mean and median time-to-goal of the final 20 episodes of training for each agent.

We furthermore find that in all environment topographies, the addition of the Dyna algorithm improves the performance of the DSWM state space-based agents, and results

Topography	Optimal	Statistic	WORLD	DSWM	DSWM+DYNA	ONEHOT
Open	5	Mean	32.1	5.81	5.0	7.76
		Median	7.45	5.0	5.0	7.1
Rooms	7	Mean	99.0	23.93	7.04	8.64
		Median	99.0	7.6	7.0	7.55
Ring	5	Mean	99.0	23.8	5.0	5.0
		Median	99.0	5.0	5.0	5.0
Hallway	5	Mean	79.22	5.0	5.0	5.0
		Median	99.0	5.0	5.0	5.0

Table 1: Statistics from final 20 episodes of each training session for goal-directed agents. DSWM+DYNA results in most consistent learning, with near optimal performance in all four topographies.

in optimal performance for three out of the four environments, with the "Rooms Maze" performance being slightly below optimal. We can interpret these results as a clear sign that the learned latent space in the DSWM model is both useful for predicting trajectories of experience in novel environments, but also in subserving goal-directed navigation in novel environments. Additionally, the DSWM+DYNA model performing best suggests that the DSWM has learned a coherent model of the dynamics of the environment which are able to abstract away the specific content of the environment.

IV.3 Learning from Egocentric Observations

Thus far we have demonstrated the properties of generative temporal models using exclusively environments with observation and action spaces which are defined with respect to allocentric coordinate systems. As such, we have missed out on a critical aspect of animal learning and acting, the fact that they do so from a limited egocentric perspective. In this section we introduce a new three-dimensional environment from which high-dimensional egocentric visual observations can be derived. We then show that an agent using a DSWM model can learn to predict trajectories though this more complex environment. Crucially, in animals this ability involves the transformation of the purely egocentric sensory obser-

vations and actions into an allocentric reference frame, and then a reverse transformation back into a egocentric coordinate space for prediction and goal-directed action (Zaehle et al., 2007). We furthermore demonstrate that the DSWM is able to accomplish this in a largely unsupervised manner.

IV.3.1 Evaluation Methods

In order to examine the properties of various generative temporal models within environments with an egocentric reference frame, we use a novel three-dimensional environment built using Unity, a 3D rendering and physics engine, taking advantage of the ML-Agents toolkit in order to enable the agents to interface with this environment (Juliani et al., 2018). The environment can be thought of as a three-dimensional version of the gridworld environment presented above. The environment consists of a set of nodes which the agent or a wall can take up.

At a given time-step, the agent is presented with an observation derived from the agent's current position and orientation within the environment. This observation consists of a $64 \times 64 \times 3$ color image presenting a 120-degree field of view. See Figure 36 for renderings of the environment from multiple different angles, including the agent's perspective. Within the environment, the agent can take one of five actions: either move forward, move left, move right, rotate 90 degrees to the left, or rotate 90 degrees to the right. There are four possible orientations which the agent can take, consisting of facing each of the four cardinal directions. As such, in a 7×7 environment, there are 196 possible states the agent can be in, assuming there are no wall obstacles within the environment.

In order to test the generalization ability of the agent, we use a similar set of topographies and pattern maps as was done in the previous section. Instead of the pattern map being overlaid on the open spaces of the environment, as was done in the 2D case, here we overlay them on the wall obstacles instead. Additionally, we use three possible colors, red, green, and blue, when defining the map in order to introduce greater visual variety to the

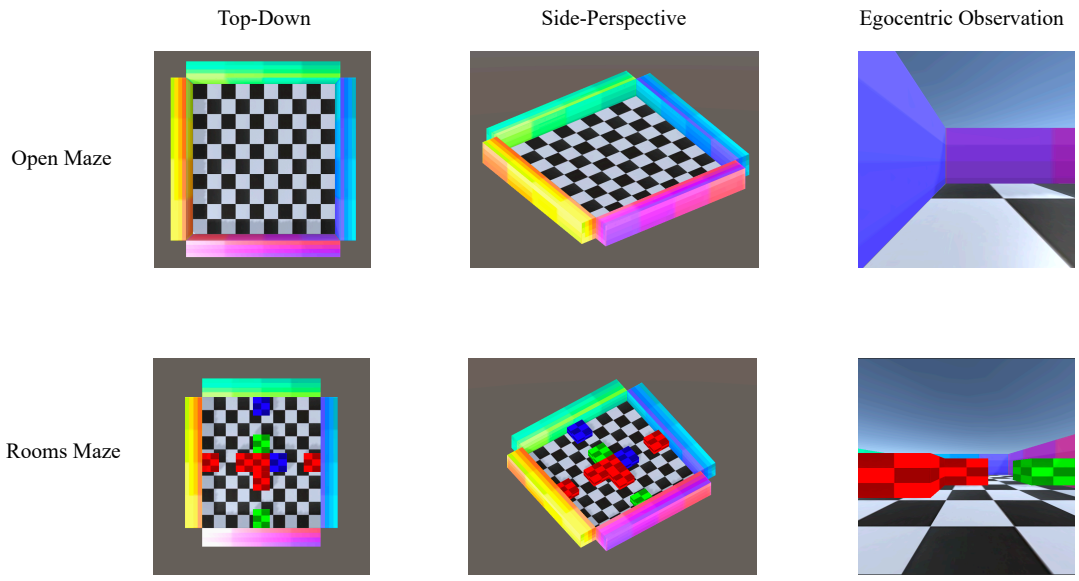

Figure 36: Three dimensional gridworld environment rendered using Unity. Two example topographies shown. Top: open maze. Bottom: rooms maze. Left: top-down perspective of environment. Middle: side-perspective of environment. Right: egocentric observations provided to agent within environment.

"content" of a given environment. Figure 36 contains an example of this pattern map in the 3D version of the "Rooms Maze" environment.

As was the case in the 2D environments, in order to generate a training set, we generate 100 pattern maps for each topography, and use four different topographies: "Open Maze," "Ring Maze," "Rooms Maze," and "Hallway Maze." We then collect 100 episodes of 50 time-steps each for each topography, and train a separate WORLD and DSWM model on each dataset. Due to the larger state space, we use a vector of length 128 for the latent context space s in the DSWM. We also modify both the WORLD and DSWM models to use a three-layer convolutional neural network (CNN) (LeCun, Bengio, & Hinton, 2015) to encode the image-based observations, and likewise use a de-convolutional network to decode the predicted observations from these models. We use the same convolutional architecture described by Ha and Schmidhuber (2018). Otherwise use the same hyper-parameters defined in the 2D experiments, including training for 5000 iterations.

IV.3.2 Results

Examining the reconstruction error on a test-set of pattern maps for each environment, we find that in all cases the DSWM ($Mean = 223.35, Std = 213.39$) is able to significantly better predict future trajectories of observations than the WORLD ($Mean = 304.17, Std = 215.23$) model ($p < 0.001$). Figure 37 presents the reconstruction error for each model and topography visually.

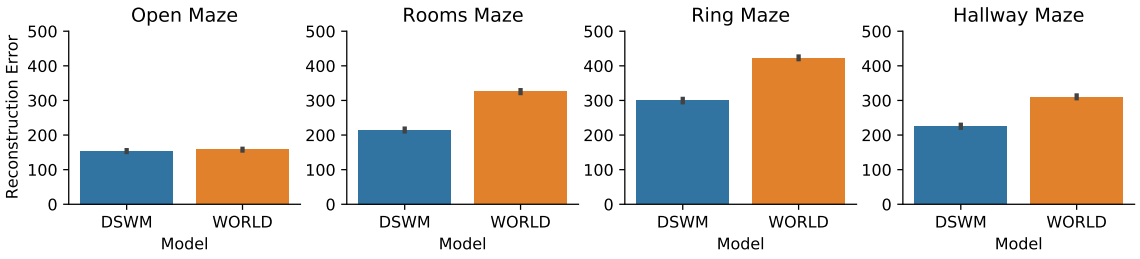

Figure 37: Reconstruction errors from rollouts of both World and DSWM models in four different topographical environments. Error bars represent standard error. In all environments, DSWM is able to significantly better predict trajectories of future observations than the WORLD model.

We can also examine the quality and coherence of the predictions of the two models. In Figure 38 we present example rollouts in each of the four topographies on test-set pattern maps. In all cases, the WORLD model produces trajectories with differ more severely than the DSWM. In particular, the DSWM is better able to track the correct colors of the wall obstacles, whereas the WORLD model often predicts incorrect colors.

Lastly, we can also examine the learned representations of the DSWM latent space s. Critically, here we are learning the latent representation from egocentric observations, which by definition do not a priori contain necessary allocentric information from which a place code could be derived. Given our loss function and the existence of a recurrent neural network processing these observations, there is reason to believe that the model could learn to integrate the observation stream into an allocentric place code. In Figure 39 we present example activation patterns of the latent code s from a trained DSWM in each of the four topographies. We find that a place-like code does indeed develop within the model,

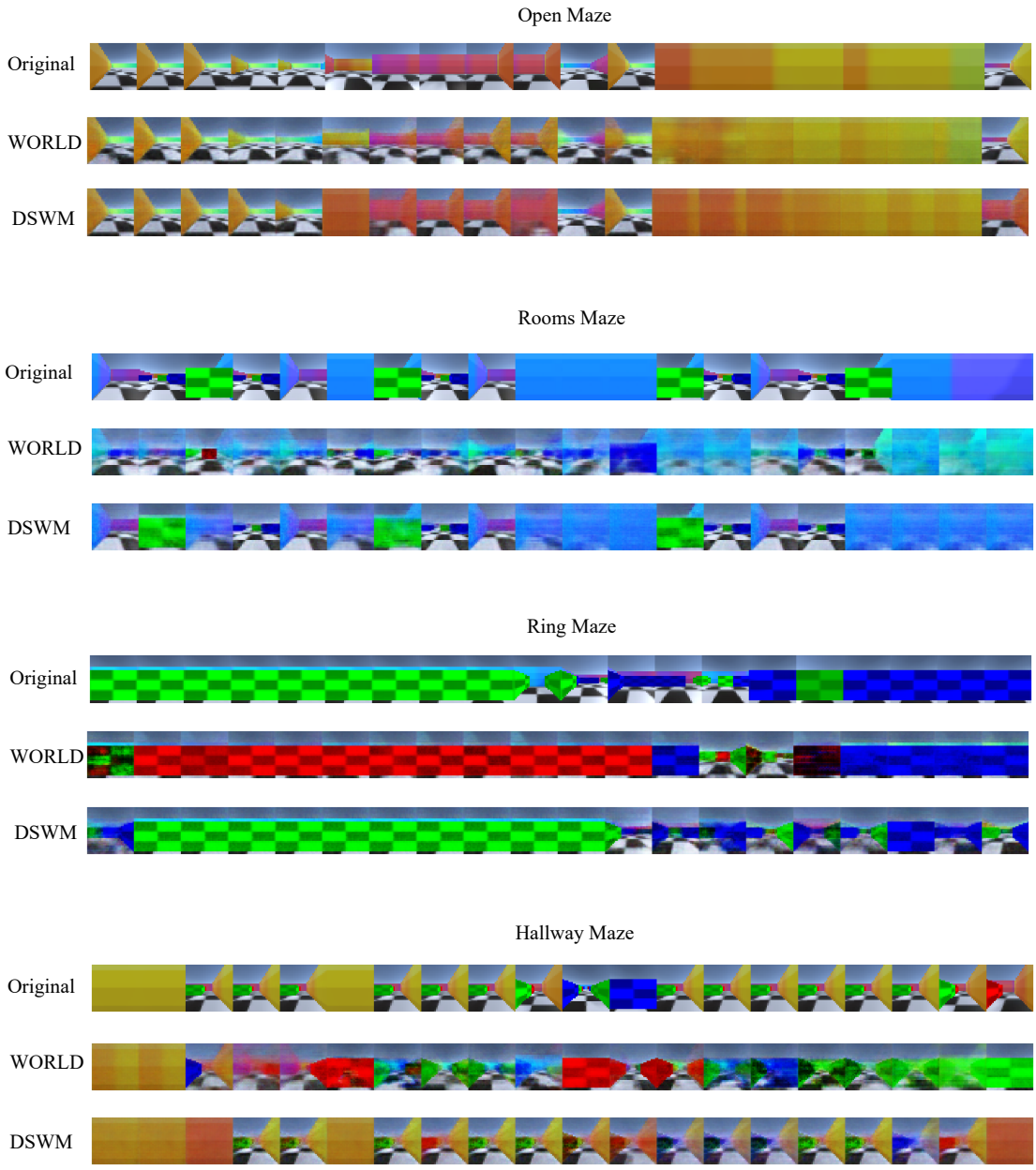

Figure 38: Examples of reconstructed observations from rollouts of both World and DSWM models in four different topographical environments. Environments use pattern map reserved for testing, and not seen during training. In all environments, DSWM is able to better predict the true trajectory of future observations within the novel environment.

providing evidence for a learned translation from egocentric to allocentric representation.

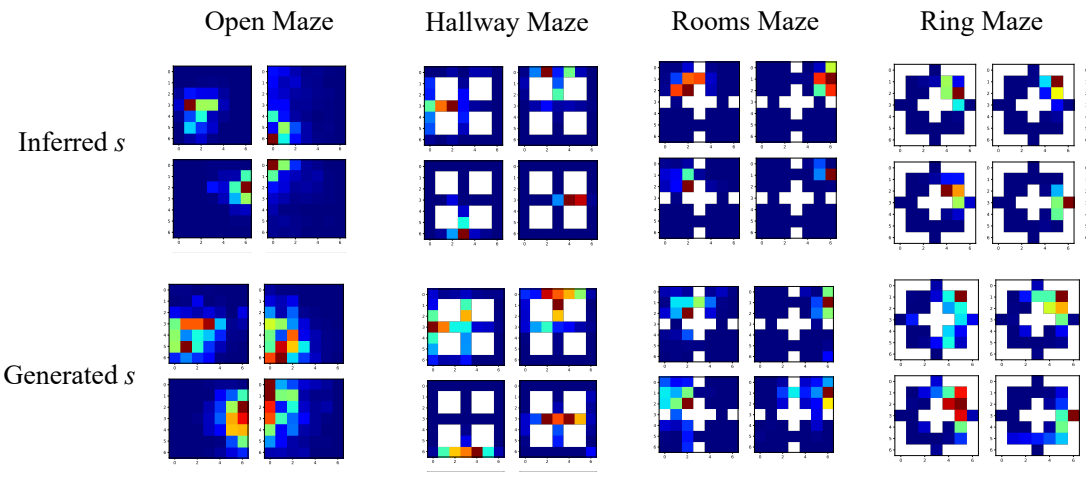

Figure 39: Examples of activations of selected four units of inferred and generated s from DSWM model in each of the four different environment topographies.

IV.4 Goal-directed Navigation from Egocentric Observations

Generating coherent trajectories and possessing a structured latent space can be useful to the extent to which it supports useful goal-directed behavior for the animal. While we have previously demonstrated that the DSWM latent space supports goal-directed navigation in the allocentric case, here we demonstrate that it also does so in environments with high-dimensional egocentric observations.

IV.4.1 Evaluation Methods

We use the same environments described in the previous section to test for goal-directed navigational abilities. We compare a DSWM state space, a DSWM model augmented with Dyna, and a one-hot encoded state space. The WORLD model state space is excluded here due to the poor navigational performance in the 2D environments, suggesting that learning would not be possible in 3D environments either. We use the SSL algorithm to train each agent.

Due to the expanded state and action spaces of the environment, we use a fixed-goal navigation task where the goal location remains constant throughout training. See Figure 40 for the starting agent and goal positions for each of the four topographies. For each training session, we train the agent for 100 episodes of a maximum of 200 time-steps. Training sessions are repeated five times, each with a different initialization seed for the agent.

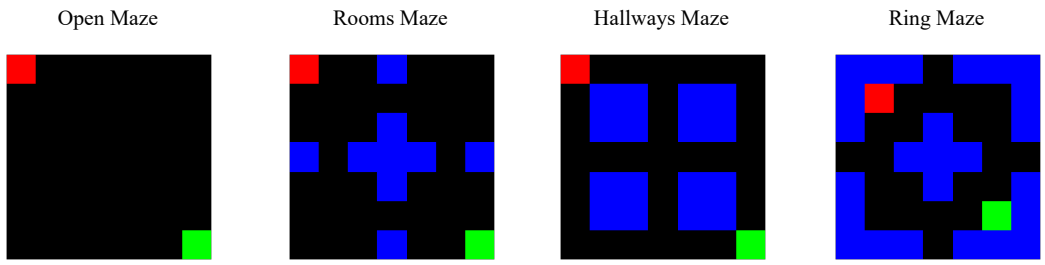

Figure 40: Starting agent and goal positions for each of the four topographies in the 3D environment. Red: agent position. Green: goal position. Blue: wall positions.

IV.4.2 Results

We find that for all agents learning is more difficult in the 3D environments than in the 2D variants. This is true both for the time to convergence, as well as for the ability for a given algorithm to converge. This can be seen in the learning curves presented in Figure 41 and the full table of results presented in Table 2. Despite the general increased difficulty in learning, we find that the DSWM state space augmented with DYNA is either competitive with or outperforms the pre-computed onehot state space. In both the open maze and Ring maze, we find that agents with the DSWM state space by itself fail to converge to an optimal policy in some of the five training sessions, while they converge in others.

Taken together, these results suggest that the DSWM is able to learn a useful state space which is competitive with a pre-computed onehot encoding of the environment state. The quality of this state space varies however based on the topography. The difficulty of the

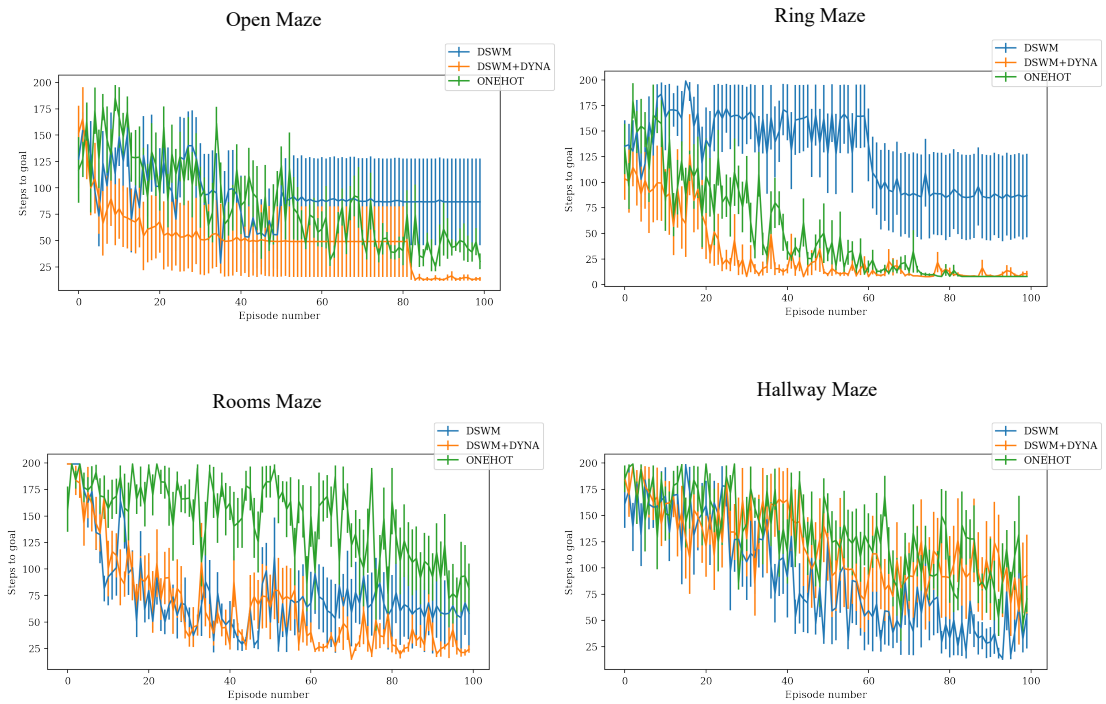

Figure 41: Learning curves in goal-directed navigation task for each of the four unique environmental topographies. Each curve represents the average of five separate initialization seeds for the agent. Error bars represent standard error.

learning problem also varies by topography. We find that learning in the Open and Ring Maze environments is easier, whereas the Rooms and Hallway Mazes are more difficult.

IV.5 Discussion

In this chapter, we introduced the Dual-Stream World Model, and analyzed its properties with respect to both the coherent generation of trajectories of experience in environments with novel content, as well as the ability to provide a support for goal-directed navigation. This proposed model takes inspiration from recent generative temporal models which include differentiable memory stores, such as the Model-Based Predictor (MBP) (Wayne et al., 2018), the Generative Temporal Model with Spatial Memory (GTM-SM) (Fraccaro et al., 2018), and the Tolman-Eichenbaum Machine (TEM) (Whittington et al., 2019). While related to each model, there are important differences which set the DSWM apart.

Topography	Optimal	Statistic	DSWM	DSWM+DYNA	ONEHOT
Open	12	Mean	86.9	18.0	45.31
		Median	12.5	12.0	35.85
Rooms	12	Mean	65.51	30.06	103.58
		Median	25.1	26.9	118.7
Ring	8	Mean	87.18	9.96	8.17
		Median	16.3	8.0	8.0
Hallway	12	Mean	38.28	96.5	94.3
		Median	12.0	113.8	91.05

Table 2: Statistics from final 20 episodes of each training session for goal-directed agents in 3D environment. DSWM+DYNA results in most consistent learning, with near optimal performance in all four topographies.

While both the GTM-SM and DSWM use a similar DND as a storage and look-up mechanism (Pritzel et al., 2017), DSWM uses a more general-purpose representation for the 'where' variable s. Furthermore, we demonstrate the usefulness of this representation in both a trajectory generation task as well as a navigation task, whereas the GTM-SM is used only for a trajectory generation task. We believe that the learned representation described in (Fraccaro et al., 2018) is not suitable to reinforcement learning, as it corresponds to continuous-values x and y coordinates, which we demonstrated in Chapter III do not allow for convergence during learning.

Likewise, whereas the TEM takes more specific inspiration from hippocampal anatomy, and thus could be seen as more biologically plausible, the authors do not demonstrate the usefulness of the learned representations for any navigation tasks. The TEM is also only demonstrated with hand-crafted low-dimensional observations, whereas we have demonstrated the efficacy of the DSWM on high-dimensional egocentric observations similar to those an animal would encounter during navigation.

Lastly, we can compare the DSWM to the MBP. Both of these models are validated using high-dimensional observations on tasks of both trajectory generation and goal-directed navigation. The MBP however utilizes a single latent state, and was not tested in the domain in which the environment content significantly changes, and the capability of the model to

adapt to these changes is not clear.

As such, the DSWM can be seen as a meaningful addition to this growing ensemble of dictionary-based models of hippocampal learning, with clearly demonstrated properties of adaptability to changes in environmental content, while maintaining the ability to generate coherent trajectories of experience, and support goal-directed navigation.

One potential weakness of the DSWM compared to the other models described is the need for an auxiliary loss function based on spatial information in order to train the latent representation s. In the other models described, this learning signal is not used, or is built more explicitly into the models, as is the case of the GTM-SM. Given the evidence for both representations of spatial position and head direction in regions adjacent to the hippocampus, namely the MEC (Hafting et al., 2005) and subiculum (Taube et al., 1990), we believe that it is not implausible for this signal to help guide the place cell representations within the hippocampus proper. Still, we acknowledge that the ability to induce a similar representation in an unsupervised manner would amount to a significant improvement of the model presented here.

The DSWM can be seen as providing a means of largely actualizing the cognitive mapping system described in the introduction of this work. We have presented a model which can adapt to changes in both goal location and the content of the environment in a rapid manner consistent with experimental evidence in mammals. One aspect missing from the proposed model so far however is the ability to adapt to changes in the structure of the environment itself. It is this capability which we turn to in the next chapter.

CHAPTER V

STRUCTURAL GENERALIZATION AND CONTEXT MODELS

In the previous chapter, we introduced a Dual Stream World Model, and demonstrated its capacity to model a number of known properties of the medial temporal lobe. In particular, we demonstrated the ability for the model to learn to adapt to changes in environment content, and to learn an allocentric representation useful for navigation from an egocentric observation signal. So far however, we have focused on environments with fixed topographic structure. In doing so, we have ignored a key property of the cognitive map in animals, the ability to adapt to structural changes in the environment (Tolman, 1948).

For a living animal, such structural changes can either take place in the form of arriving in a novel environment, or in the introduction of shortcuts or roadblocks into a familiar environmental structure. Both of these capabilities are based on the ability of the animal to adapt to changes in environment topography, and to take advantage of non-reactive, and generalized representation of space. In this chapter we explore a series of extensions to generative temporal models which provide them with some capacity for structural generalization.

In addition to the ability to adapt to changes in the structure of an environment, cognitive maps in animals also support the ability for animals to make sense of their surroundings in many different environments. This making sense involves the generation of coherent imagined trajectories of experience, as well as the ability to perform goal-directed naviga-

tion. In the case of imagining trajectories of experience, these can be from environments which the animal does not currently inhabit (Karlsson & Frank, 2009). This can be seen as the ability to store and retrieve not only the content of a specific map, but the ability to store multiple such maps simultaneously.

The maintenance of multiple cognitive maps, and the ability to adapt to changes in environment structure within a single map are both instances of a more general property of the medial temporal lobe, and the cognitive maps they support. In both cases what is additionally being represented alongside 'what' and 'where' information is an additional 'how' variable. In cases where environmental changes are large and discrete, this 'how' represents simply the different maps. In cases where changes are more subtle within the environment, this 'how' represents specifics about the nature of the task. Returning to the language metaphor presented earlier, this 'how' information can be thought of as the specific grammar rules which apply at a given time and context.

Building on the previously demonstrated generative temporal models, and our understanding of the medial temporal lobe, we propose two new models, a context augmented generative temporal model, or Contextual World Model (CWORLD), and a Tri-Stream World Model (TSWM), which learns separate representations for 'what,' 'where,' and 'how' (or context).

In this chapter we will formally define the CWORLD and TSWM models, and define a few possible loss functions which can be used to train the contextual representation, c. We will demonstrate the efficacy of each with respect to both the quality of the model's trajectory predictions in novel environments, and explore the nature of the learned representation c, which we connect with the contextual scene representation found within the parahippocampal area in humans (R. A. Epstein, 2008).

V.1 Learning an Index-based Context Representation

Thus far we have examined the capabilities of generative temporal models in environments with a single, fixed, structural topography. Animals in the wild however are able to adapt their behavior to multiple different environments. In this chapter, we explore the ability for a context-augmented generative temporal model (CWORLD) to learn to model the dynamics of more than one environment structure. In doing so, the question which arises as to what the ideal loss function is to train such a representation.

In this section, we explore one of the simplest possible objective functions to train a contextual representation c. When there is a known, fixed set of environment topographies, we can train the context representation to simply be useful for predicting the identity of the environment topography. We find that the CWORLD model is indeed able to be trained to perform this identity prediction task for the current topography the agent is exploring. Furthermore, we show that the latent representation which supports this identification allows the generative model to make more accurate predictions of trajectories of future observations than a WORLD model without any explicit contextual representation.

V.1.1 Modeling Methods

We compare a WORLD model, as described in Chapter II to a Contextual World Model (CWORLD) described below. Similar to the DSWM, the CWORLD model uses two separate encoding streams, and two latent variables, z and c. Once encoded, both variables, along with the selected action from the agent are used as input in the RNN layer of the network to produce a generated z from which a predicted observation is decoded. Here we train the inferred z representation using the traditional variational autoencoder loss function as done in all previous models. We train the contextual representation c to predict the map identity using a cross entropy loss function after a series of decoding layers, in addition to the variation regularization loss for the gumbel-softmax distribution. The equa-

tions governing the inference and generation process are provided below. For a graphical representation, see Figure 42.

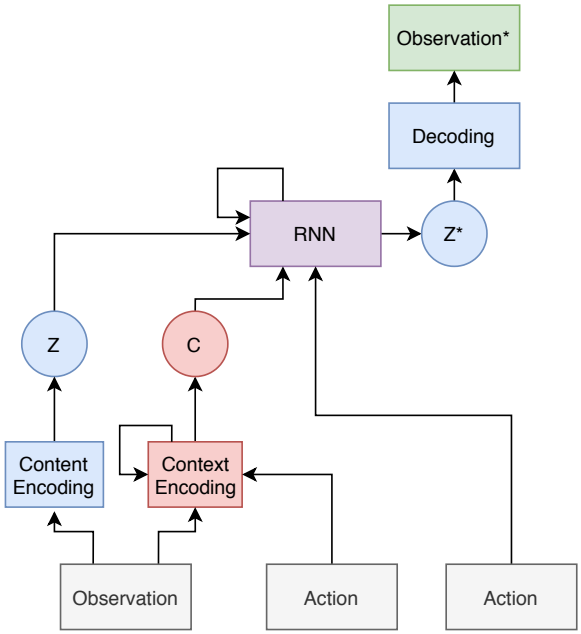

Figure 42: Diagram of a Contextual World Model. Blue represents content information. Red represents context information. Purple represents joint content and content information. White represents model inputs. Green represents model outputs. Nodes marked with a $*$ indicate information at the next time step of the simulation.

Inference phase:

$$z_t \sim p_{enc}(z_t|o_t) \tag{V.1}$$

$$c_t \sim p_{enc}(c_t|o_t, h_t^c) \tag{V.2}$$

$$h_t^c = f_{forward}(o_t, a_{t-1}, h_{t-1}^c) \tag{V.3}$$

$$h_t^z = f_{forward}(z_t, a_t, h_{t-1}^z) \tag{V.4}$$

Generation phase:

$$z_{t+1} \sim q_{forward}(z_{t+1}|z_t, c_t, a_t, h_t^z) \tag{V.5}$$

$$o_{t+1} = f_{decode}(z_{t+1}) \tag{V.6}$$

V.1.2 Evaluation Methods

In order to examine the ability for structural adaptation, we use a set of two-dimensional environments with novel topographies. We generate these topographies using the inverse Fourier fractal generation method (Bies, Boydston, et al., 2016). We use a value of $\beta = 2.0$ to define the fractal complexity, and threshold each generated fractal map at 0.65, setting all values less than the threshold level to 0 and all values above it to 1. We then use the values set to 1 to represent wall and obstacles within the environment, and all values set to 0 to represent the navigable ground. We generate each topography by providing a unique seed to the random number generated used in the fractal generation process. We use this process to generate sixteen unique topographies within environments of 13×13 units, including two units on each edge for observation padding (see below). Figure 43 provides a visual representation of these environments.

We collect 1000 trajectories of 50 time-steps each. For each trajectory, we randomly select one of the sixteen environments, and a random starting position for the agent within the environment. The agent follows a semi-random allocentric movement policy as described above. The observation provided to the agent at each time-step is a 5×5 window around the agent's current position providing information about the presence or absence of a wall in each location. As such, the total size of the observation vector is 25.

We train both models using a latent z representation composed of four gumbel-softmax distributions with size 16 each, for a total representation size of 64. We likewise use the same number and size of distributions for the c representation. We train the entire model end-to-end using a learning rate of $\alpha = 5e^{-4}$. We train both models for 5000 iterations on batches of entire trajectories, using a batch size of 3.

V.1.3 Results

We can first evaluate the ability of the CWORLD model's contextual representation c to enable prediction of the identity of the environment topography. Presented in Figure 44

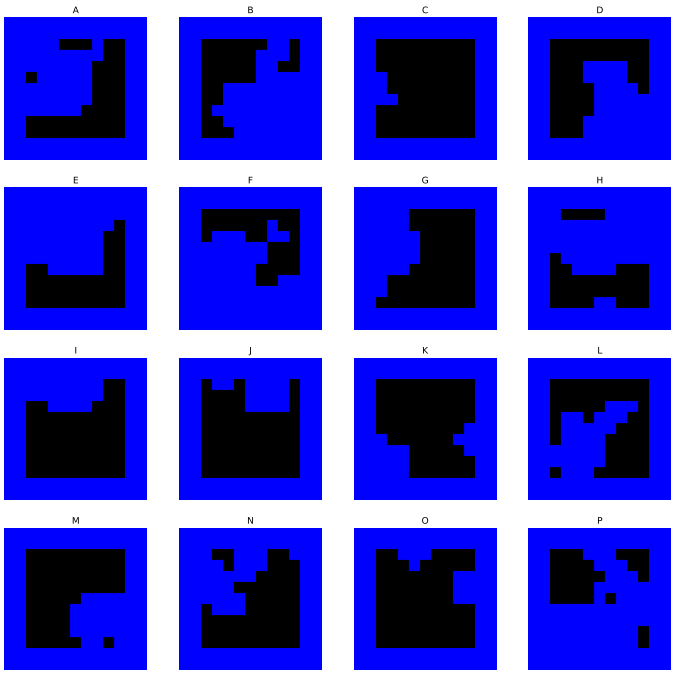

Figure 43: Examples of sixteen environments with fractal topographies. Blue represents walls. Black represents navigable space for the agent.

is the correlation matrix for the predicted and actual environment topography identities. Predictions are taken from the model after a "burn-in period" of an initial 30 time-steps within the environment to provide the agent an opportunity to develop the contextual representation. Predictions are averaged over 100 episodes, and the final 20 time-steps of each episode.

We find that in 11 of the 16 environments the model assigned a 50% or greater probability to the correct environment identity. In an additional two environments, the model assigned a 40% probability (a plurality) to the correct environment identity. This leaves only three environments, 'I,' 'J' and 'O,' which the agent had difficulty identifying. If we examine these environments, we find that they all share largely the same topographic features, with a prominent protrusion in the northwest corner, and largely open space other-

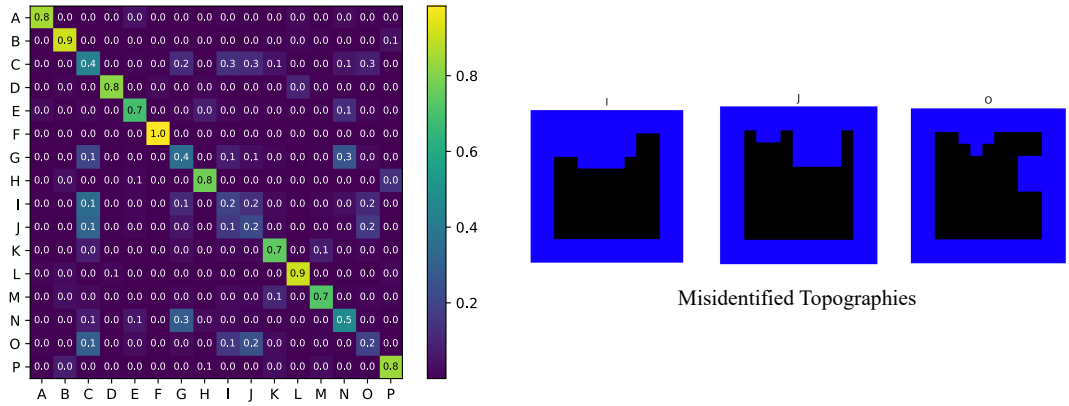

Figure 44: Classification accuracy of index-based contextual world model. Left: Correlation matrix of predicted and actual environment topography identities. Right: Three environment topographies which were misclassified by the CWORLD model.

wise. We can determine that overall this suggests the model is able to successfully classify environment identify based on partial information concerning their topography which is obtained via the observations available to the agent.

When examining the reconstruction errors of the WORLD and CWORLD models, we find that there is a significant difference in quality between the two models. The WORLD model produces poorer reconstructions ($Mean = 2.98, Std = 3.04$) than the CWORLD model ($Mean = 2.85, Std = 3.03$), $t(67198) = 5.37, p < 0.001$. These results are presented visually in Figure 45.

We can interpret these results as providing evidence that the addition of the contextual variable c does indeed allow for greater reconstruction accuracy when predicting trajectories of observations in different environments. Put more simply, the model having a sense of which environment it is in allows it to better make predictions about what it will observe in that environment.

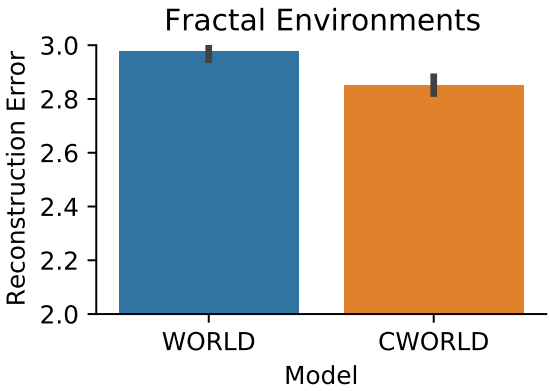

Figure 45: Reconstruction error for predicted trajectories of future observations for both WORLD and CWORLD models. CWORLD model is able to predict observations with significantly less error than WORLD model.

V.2 Learning a Map-based Context Representation

In the previous section we demonstrated that a generative temporal model could learn a useful context representation c by training the model to predict the identity of the environment topography the agent was in. While the efficacy of this approach has been demonstrated, it has a number of drawbacks. Firstly, it is not very biologically plausible, due to the lack of an explicit numbered representation of each environment an animal experiences. Furthermore, using a fixed index results in a model only capable of learning a context for a fixed number of environments. We know that animals learn to make sense of a large number of environments. More importantly, using fixed indices for each environment imposes a strict representational boundary between each environment, and prevents any use a generalized knowledge gained in one environment to be applied to another. Finally, while this context representation resulted in a statistically significant decrease in reconstruction error when predicting trajectories of observations, this decrease was relatively modest. In this section, we propose a second loss function to train the context representation c which addresses these issues.

We propose a new objective, which consists of training the context representation c to be useful for predicting the structure of the topography of the environment the agent is cur-

rently within. This training objective has the benefit of not being bounded by the number of training or testing environments, since each environment can be defined uniquely by their topography. Secondly, since we are not training the model to predict a fixed set of information, as was the case when predicting the index of the environment, this new objective allows for generalization to novel environmental structures. In this section we demonstrate the efficacy of this approach compared to both the context-less WORLD model and the contextual world model using index learning (CWORLD-I). We refer to the approach we propose and compare here as a Contextual World Model with Map-Prediction (CWORLD-M).

V.2.1 Evaluation Methods

We use the same process to generate training environments from the previous section, using inverse Fourier fractal generation. In addition to collecting the 5×5 observations for each trajectory, we additionally collect a vector representing the environment topography. In the case of a 13×13 environment, this corresponds to a binary vector of size 169.

V.2.2 Modeling Methods

We compare a WORLD model to CWORLD-I, and the proposed CWORLD-M. The latent representation c of the CWORLD-M model is trained using a binary cross entropy loss function. This loss function compares the true map topography vector to the predicted vector, deriving a gradient used to improve the representation of c during training.

$$L(c) = -\Sigma(p\log(q) + (1-p)\log(1-q)) \tag{V.7}$$

We use the same set of hyperparameters from the previous experiment with the CWORLD-I model.

V.2.3 Results

We can first examine whether the prediction loss used by the CWORLD-M model was able to produce a representation c which is indeed able to predict the environment topography. While there is no baseline to measure this model's prediction error against, we can qualitatively evaluate the predicted topographies.

Figure 46 presents example topography predictions from the CWORLD-M model alongside the true map topography. In most cases, the model is initially unsure of the correct topography, and assigns medium probability of wall location to most of the units in the environment. As the agent moves around and collects more evidence via observations, the prediction becomes more certain, and in most cases eventually reflects the true underlying environmental topography.

Figure 46: True environment topography alongside predictions from the CWORLD-M model at test-time for environment topographies A-E. White corresponds to regions of high certainty there is a wall, which black corresponds to regions of low certainty. Eighteen predicted topographies from the model are shown, consisting of the first and last nine of the "burn-in" period.

We can next ask whether this map prediction task, while able to produce qualitatively convincing predictions of the environment is actually useful for predicting future observation trajectories. We do so by comparing the reconstruction error produced by each model

when unrolling a trajectory of predicted future observations. In each case, we allow the model a 30 time-step "burn-in" period, followed by a 20 time-step auto-regressive unroll of the model. We find that there is indeed a significant difference in performance between the three models in this prediction task ($F(2, 100797) = 44.32, p < 0.001$), with the CWORLD-M model ($Mean = 2.75, Std = 2.99$) producing predicted observations with significantly less deviation from the true observations than other other two models ($p < 0.001$). Figure 47 presents these results visually.

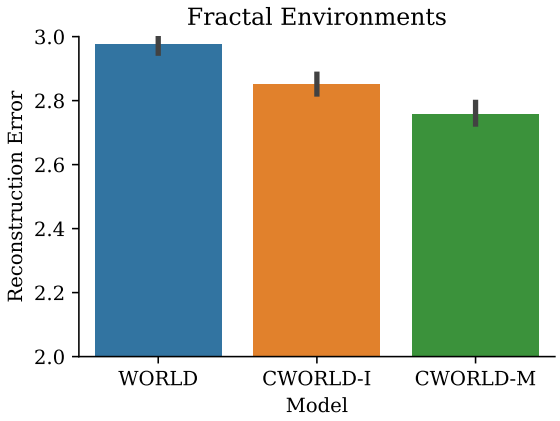

Figure 47: Reconstruction error for predicted trajectories of future observations for both WORLD and CWORLD models. CWORLD model is able to predict observations with significantly less error than WORLD model.

V.3 Learning Implicit Context Representations

In the previous sections of this chapter we have presented two different objective functions which can be used to train a useful context representation c within a generative temporal model. In particular, the second function presented, the topography prediction task, can be used in novel environments, since it does not rely in predicting a quantity limited by the training dataset. One remaining issue with this objective function is that there is not a biologically plausible complete topographical representation of the environment which an animal might use to train such a context representation. Indeed, it seems unlikely that animals would maintain literal topographic representations of space, unless explicitly trained

to do so. In this section, we propose an unsupervised loss function which shapes the context space c to be useful for prediction tasks.

The insight we build on is that what we'd like our model to optimize is the predictive quality of the dynamics model over z^*. Instead of developing a surrogate loss function for this optimization problem, we can directly learn a c which helps to optimize this quantity. The unsupervised loss function which we propose to train c simply consists of allowing the gradient from the forward model to pass through the c and observation encoder. We augment this with an additional forward model over the context c, so that both z and c can evolve independently during auto-regressive trajectory predictions. We refer to this model as CWORLD-U, due to the unsupervised nature of the learned context. We demonstrate that this new model variant results in significantly greater predictive accuracy than either of the previous proposed contextual world models.

This separate contextual representation which evolves on its own and guides the learning process of the z forward model can be seen as a kind of hierarchical system, where more abstract information about the dynamics of the environment are encoded into c, while only relevant local information is encoded into z. This has a connection to the interplay between the hippocampus and the parahippocampal area, which is known to respond preferentially to stimuli containing structural information (R. Epstein & Kanwisher, 1998), and contains a more general contextual representation of the current scene (R. A. Epstein, 2008). This region is known to tightly interface with the hippocampus to pass this information onward (Van Hoesen, 1982). Here we propose that a potential purpose for this interplay is to aid the trajectory generation which takes place within the hippocampus by providing it the correct context with which to generate coherent sequences of activation.

V.3.1 Modeling Methods

The CWORLD-U model consists of a modified version of the previous CWORLD models. In this case, we augment the model with an additional forward model over the c latent

state. This is needed in order to ensure that when unrolling the model to predict trajectories of observations both the z and c states are kept up-to-date. This was not necessary in CWORLD-I and CWORLD-M, where the c could be interpreted as a fixed quantity (either an environment index or map representation). In the case where c is learned in an unsupervised fashion, we expect that the representation will evolve over time, and therefore a forward dynamics model is needed. Critically to training this model, the c and z dynamics models take as input both c and z from the current time-step, but the gradient is only allowed to flow backwards from the z dynamics model into c. This ensures that the latent state c is formed to aid the development of z, and not the other way around. See Figure 48 for a visual representation of the network flow.

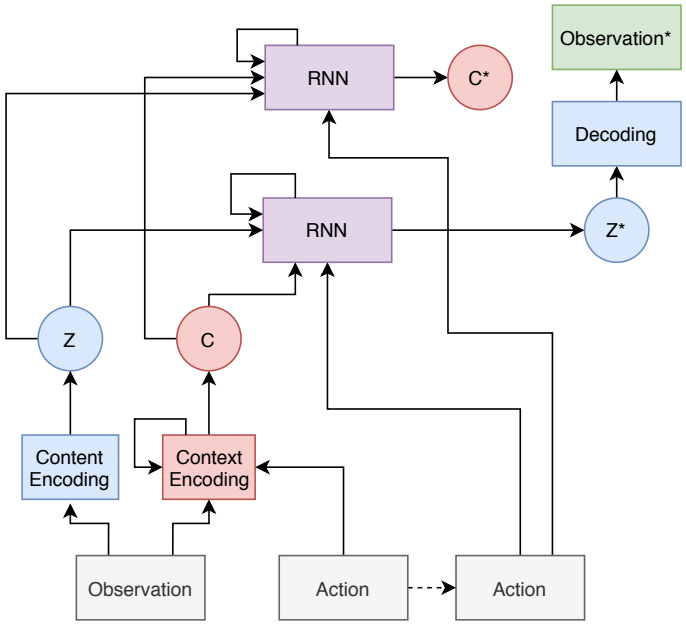

Figure 48: Diagram of CWORLD-U model. Red represents context information. Purple represents joint content and content information. White represents model inputs. Green represents model outputs. Nodes marked with a $*$ indicate information at the next time step of the simulation.

$$c_{t+1} \sim q(c_{t+1}|c_t, zt, at, h_t^c) \tag{V.8}$$

$$z_{t+1} \sim q(z_{t+1}|c_t, zt, at, h_t^z) \tag{V.9}$$

V.3.2 Evaluation Methods

In order to evaluate this novel model variant, we utilize both the same dataset consisting of 16 fractal topographies, along with an additional larger dataset containing environments with 100 different fractal topographies. As done previously, this new dataset consists of 1000 episodes of 50 time-steps each. Each trajectory is sampled from a random fractal topography, and the agent starting position is randomized within the open space in the environment.

In the smaller dataset, we compare the WORLD model to the three contextual variants, CWORLD-I, CWORLD-M, and CWORLD-U in their ability to generate coherent trajectories of imagined observations in each of the sixteen environments. In the second experiment using the larger dataset of 100 topographies, we then test on a set of hand-crafted environment topographies. See Figure 49 for these test environment topographies. We use this separate set of environments in order to examine the generalization ability of these model with respect to their ability to predict trajectories of observations in these unseen environments. The CWORLD-I model is excluded from this analysis, as we do not expect the loss function used to induce a representation which improves generalization.

We train all models for 5000 iterations, using a learning rate of $\alpha = 5e^{-3}$ and a batch size of 3.

V.3.3 Results

We first evaluate the reconstruction accuracy of imagined trajectories from all four model variants when trained and tested using the smaller dataset of 16 environment topographies.

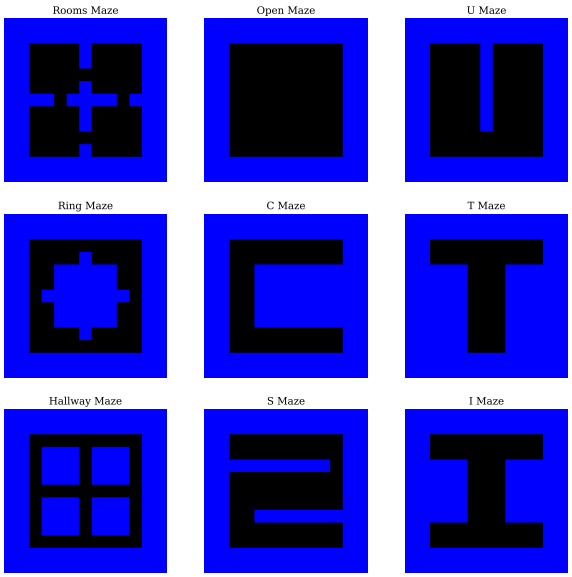

Figure 49: The nine test environments with hand-crafted Euclidean geometries. Blue represents walls. Black represents navigable space for the agent.

Figure 50 presents these results graphically. We find that there is a significant difference between the model's predictive accuracy, with CWORLD-U ($Mean = 2.572, Std = 2.912$) being able to predict future observation trajectories with significantly less error than all other models ($p < 0.0001$).

We can interpret this result as providing clear evidence that a learned contextual representation optimized to improve the dynamics model does indeed provide a better context than either an index-based or map-based representation. This is likely because of the adaptive nature of the learned c in CWORLD-U, which can change based on the current needs of the prediction problem, whereas the c in CWORLD-I and CWORLD-M is fixed.

We next turn to examining the predictive ability of the models trained using the larger dataset of 100 fractal environments. As mentioned above, we evaluate these models on a held-out set of nine hand-crafted topographies. See Figure 51 for a graphical presentation of reconstruction errors. We find that the CWORLD-U model ($Mean = 3.171, Std = 3.256$) is able to predict trajectories of future observations significantly better than the WORLD ($Mean = 3.582, Std = 3.296$) or CWORLD-M ($Mean = 3.443, Std = 3.127$) models ($p < 0.001$). These results validate our intuition that the CWORLD-U model is able to learn an

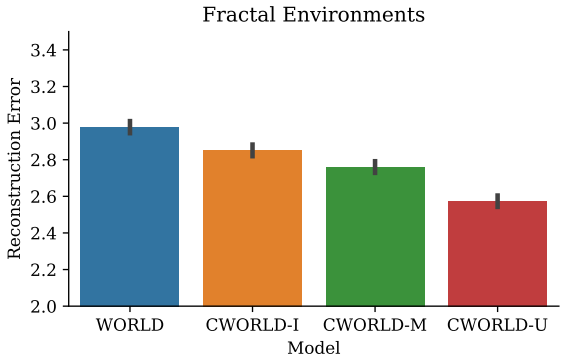

Figure 50: Reconstruction error for predicted trajectories of future observations for both WORLD and contextual variants. CWORLD-U model is able to predict observations with significantly less error than WORLD or other CWORLD models when evaluated on the same sixteen environment topographies the models were trained on.

evolving context representation c which allows for better generalization to unseen environments than a model without a context, or one with a fixed context.

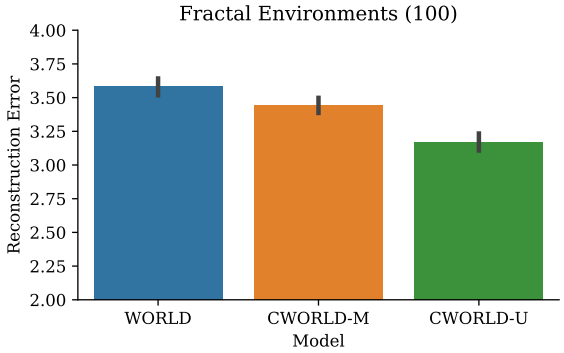

Figure 51: Reconstruction error for predicted trajectories of future observations for both WORLD and contextual variants. CWORLD-U model is able to predict observations with significantly less error than WORLD or other CWORLD models when evaluated in a set of nine hand-crafted environment topographies.

V.4 Adapting to Changes in Context and Content

Having demonstrated that context-enhanced generative temporal models can adapt their understanding of the transition dynamics to novel environment structures, we turn our attention to combining this ability with the content generalization explored in the previous chapter, and made possible by the Dual Stream World Model.

V.4.1 Modeling Methods

In the previous chapter we introduced the Dual-Stream World Model, which encoded incoming observations into separate z and s streams. Just as we augmented the WORLD model to produce the CWORLD model, we can likewise augment the DSWM with an additional context streams c to produce a model which we refer to as a Tri-Stream World Model (TSWM). These three streams can be thought of as roughly corresponding to transforming the incoming series of observations into 'what' z, 'where' s, and 'how' c representations. In this model, z and s function largely how they did in the DSWM, but the context variable c is used as an additional input to the forward model over the s variables. Likewise, s and c are additionally provided as input to the c forward model to generate c^*. See Figure 52 for a visual representation of this network flow. By augmenting the DSWM in this way, we gain a generative model which is both capable of content generalization (DSWM) but also structural generalization (context variable).

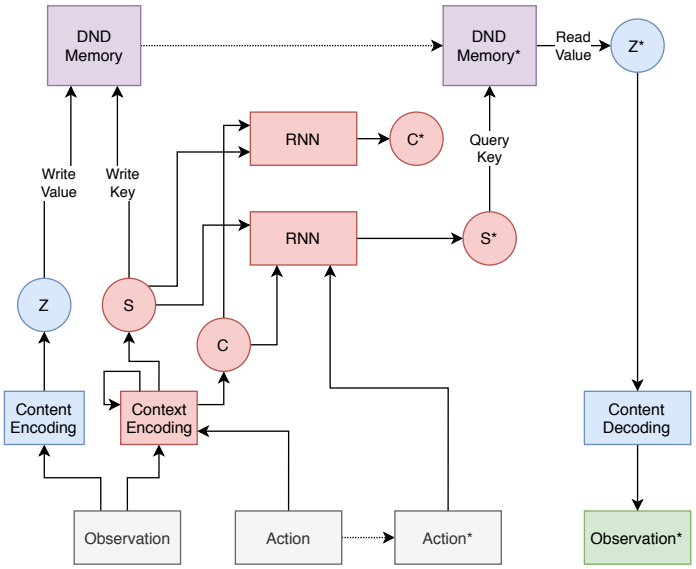

Figure 52: Diagram of Tri-Stream World Model. Red represents context information. Blue represents content information. Purple represents joint content and content information. White represents model inputs. Green represents model outputs. Nodes marked with a $*$ indicate information at the next time step of the simulation.

V.4.2 Evaluation Methods

In order to evaluate the adaptation ability of the TSWM, we again utilize a set of 100 fractal environments of size 13×13, which are generated using the inverse-Fourier method (Bies, Boydston, et al., 2016). Because we will be evaluating models capable of content and context generalization, we use the same 2D environment content found in Chapter IV, where each open space in the environment is filled with either a green or red pixel. We use this dataset to evaluate the generative modeling capabilities the TSWM compared to other baseline models.

V.4.3 Results

We first evaluate the generative modeling performance of the TSWM compared to other baseline models introduced previously. We find a significantly difference between the performance of all models ($F(6336, 32) = 192.26, p < 0.001$). We find that the CWORLD ($Mean = 10.832, Std = 4.887$), DSWM ($Mean = 10.741, Std = 6.507$), and TSWM ($Mean = 10.335, Std = 6.260$) models all outperform a baseline WORLD model ($Mean = 11.711, Std = 5.138$) ($p < 0.001$). Between these more complex models, there is no significant difference between the CWORLD and DSWM models ($p = 0.123$). We furthermore find that the TSWM is able to predict observations trajectories with significantly less reconstruction error than either the WORLD, CWORLD, or DSWM models ($p < 0.001$), suggesting that the contributions of the CWORLD and DSWM models are independent and complementary. We present these results in Figure 53.

We next turn our attention to the kinds of representations being learned within the c latent space. An examination of the activation patterns of units within the latent space suggests that there is a general affinity of structural motifs within an environment. See Figure 54 for example of cells and their firing properties. We see that certain cells respond to corners of the environment, while others consistently respond to open spaces. Likewise, some respond to long walls, while others respond to dead-ends. Collectively, this set of con-

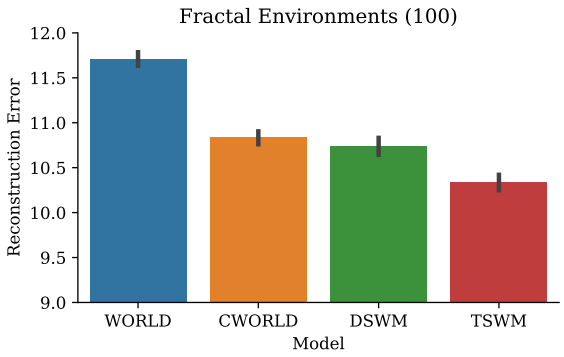

Figure 53: Reconstruction error for predicted trajectories of future observations for both WORLD and contextual variants. TSWM model is able to predict observations with significantly less error than WORLD, CWORLD, or DSWM models when evaluated in a set of nine hand-crafted environment topographies.

textual units provides a full picture of the nature of the environment topography, and thus provides necessary information to the generative model to allow for predicting trajectories of latent state representations s, and ultimately decoding imagined observations.

V.5 Discussion

The ability to skillfully imagine and navigate novel spaces involves the capacity to adapt not only to changes in the content within an environment, but also to changes in the structure of the environment itself. In this chapter, we introduced a class of generative temporal models augmented with a contextual representation meant to enable this second class of generalization. We demonstrated that there are a number of viable objective functions which can be used to learn such a contextual representation, with both supervised and unsupervised learning methods resulting in a working context representation.

Among supervised learning objective functions, we demonstrated that environment classification and map prediction were both viable to induce a context representation useful for observation trajectory prediction. Given their limitations and lack of biological plausibility, we then demonstrated that an unsupervised learning signal was a more powerful and biologically plausible option, outperforming the supervised learning alternatives. Finally,

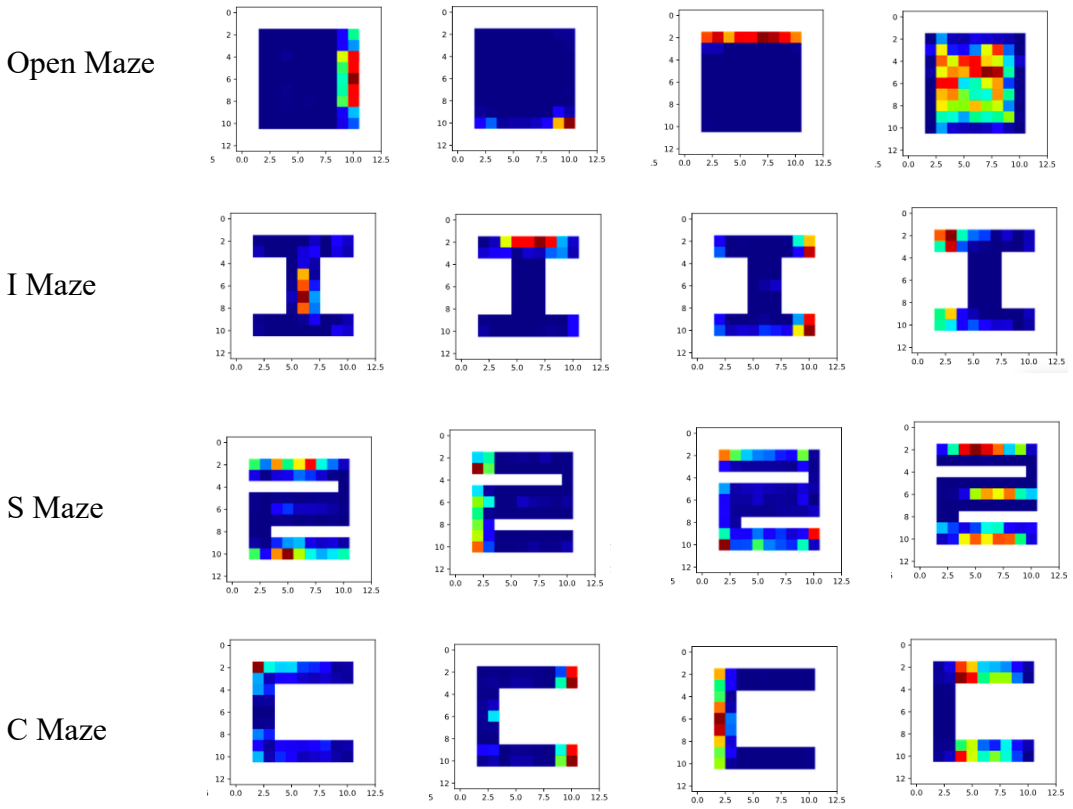

Figure 54: Examples of units from the c latent space of a TSWM model. Each row consists of hand-selected units chosen to demonstrate the structural selectivity of the cells. Each cell responds to a specific structural motif in the environment.

we combined insights from these contextual models with the advances introduced in the previous chapter regarding content generalization to propose a Tri-Stream World Model, capable of both content and context generalization.

While the Tri-Stream World Model learns to generalize to unseen environments with both novel content and structure better than those we compared it to, it is far from perfect in its predictions. We believe that there are a number of promising future approaches which can be taken to enable stronger conditioning of the dynamics model on the contextual representation, such as the use of hypernetworks (Ha, Dai, & Le, 2016).

The ability to navigate and imagine sequences of observations in novel environments depends on the brain's ability to form both a representation of what is being observed,

where it is being observed, and how a given observation related spatially to others. Within humans, this final contextual representation can potentially be localized to a number of brain regions, depending on the nature of the task, and level at which 'context' is defined. One meaningful candidate which to draw a comparison with however is the parahippocampal gyrus, specifically the parahippocampal place area (PPA). While early research connected the PPA to the identification of places and scenes in the brain (R. Epstein & Kanwisher, 1998), more contemporary work has suggested that the PPA forms a contextual representation of the local scene, useful for navigation (R. A. Epstein, 2008).

CHAPTER VI

HUMAN AND AGENT BEHAVIOR IN COMPLEX ENVIRONMENTS

Throughout this work, we have taken continuous inspiration from biological findings, both behavioral and neural. These findings have guided the classes of models considered, the objective functions used to train them, and the environments and tasks within which they are evaluated. This has led us to a class of generative temporal models which can demonstrate a number of known properties of the medial temporal lobe. What has been absent from this analysis is a contribution of novel biological results to help validate the models proposed and evaluated in purely theoretical contexts. In this chapter, we turn directly to this issue, and seek to understand the relationship between human goal-directed navigational behavior and that of the class of models we have discussed thus-far.

In particular, there is a wealth of research exploring the specific kinds of navigational strategies which humans and other mammals employ. We reviewed much of this in Chapter I, pointing out that these behaviors have largely been grouped into categories of model-free strategies, model-based strategies, and hybrid strategies (Daw et al., 2005; Momennejad & Haynes, 2012). The hybrid strategies being of particular interest for their typical instantiation in the successor representation, and successor-based learning (Momennejad et al., 2017). We have utilized both a model-free strategy (Actor-critic) and a hybrid strategy (Successor learning) when modeling the policy learning behaviors demonstrated in previous chapters. Here we seek to compare these two strategies to empirical data from humans

performing a novel navigational task.

In this chapter, we will demonstrate that humans are able to adapt to changes in environment content and goal location in a rapid fashion, but adapt slower to changes in environment structure, suggestive of a hybrid decision making strategy. Similar results have been shown in relatively artificial contexts such as the two-step task (Momennejad et al., 2017), and in simplified euclidean virtual environments (de Cothi, 2020). Here we present what we believe to be the first work demonstrating a hybrid decision making strategy in complex 3D environments involving surface, goal, and structural changes in the environment during learning.

We utilize a visually realistic set of virtual fractal island environments, building on earlier work exploring human navigational performance with respect to varying levels of environment complexity (Juliani, Bies, Boydston, Taylor, & Sereno, 2016). Such virtual environments allow for programmatically varying the environmental appearance, structure, and goal location. Furthermore, fractal topographies are found in various aspects of natural environments (Mandelbrot, 1983), such as coastlines and mountain ranges, and thus are suited to ecologically valid simulation of navigation in the natural world.

As a first step, we provide a replication of the main findings of an earlier work utilizing fractal environments, demonstrating that humans are better able to navigate fractal environments with a low-to-mid range value for the dimension, or complexity of the environment. This finding provides further evidence for the fractal fluency theory, which proposes that various aspects of the human visual and cognitive systems are most adapted to this range of complexities (Juliani et al., 2016; Bies, Blanc-Goldhammer, Boydston, Taylor, & Sereno, 2016).

Next, we examine the effect various changes on the environment have on the learning process. We do this in order to find evidence for either a model-based, model-free, or hybrid decision making strategy. Much previous work has been dedicated to determining when and how humans utilize different kinds of decision making strategies. Some recent

work has suggested that that humans navigate using a hybrid strategy (Daw et al., 2005; Momennejad & Haynes, 2012), which manifests as selective disruption to various kinds of environmental changes. This hybrid strategy has been modeled in the past using a successor representation learning algorithm (Momennejad et al., 2017; de Cothi, 2020). Here we present partial evidence for a hybrid strategy, with humans showing no disruption for visual changes, apparent disruptions for changes to the terrain and goal location, but importantly a consistent recovery from changes in goal location, but a less consistent recovery from changes in terrain.

We also train a set of artificial agents to perform a modified version of this task using Deep Reinforcement Learning. We proposed three different states spaces to use as input into these agents, one based on the pre-computed location and orientation of the agent, one based on the inferred state space s from a TSWM model, and one based on the inferred state space z from the same model. We find that all agents were able to perform the task well, but that only the agents trained using the inferred s latent state showed adaption to changes in goal consistent with human behavior. Because of the nature of this representation, and it sharing the property of geodesic representation with the successor representation, which has been previously used to demonstrate hybrid behavioral strategies, we can interpret these results as providing an additional approach to the question of human behavioral strategy. Rather than focusing exclusively on the learning algorithm, we demonstrate the value of examining the impact of the underlying representations utilized in learning on the induced behavior.

VI.1 Human Experimental Methods

We recruited subjects for this study from the University of Oregon Human Subject Pool. Participants were granted class credit for participating. Due to restrictions in place as a result of the COVID-19 pandemic, all experiments were conducted online, at the participants' convenience. Each participant was given an hour to complete the study, and in most

cases reported completing in in less time.

From the perspective of the participant, the study consisted of a website in which a 3D virtual environment was rendered from a first-person perspective. This environment consisted of an island surrounded by water. The participants are instructed that they can control their virtual avatar by moving it in the forward or backwards directions, or rotating the perspective of the avatar to the left or right. These controls were provided via keyboard buttons. Participants were instructed that their task was to follow prompts presented on the screen, and to find a goal location hidden on the island. Figure 55 provides a series of example screenshots of the perspective of the participant while exploring the island.

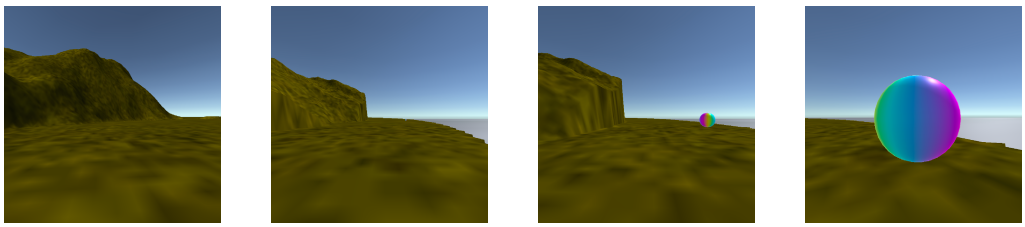

Figure 55: Example first-person perspective of participant performing navigation task. A: Participant begins trail in random location on island. B: Participant navigates around island looking for goal location. C: Participant finds goal location indicator, which grows in size as participant approaches. D: Participant touches goal indicator, ending trial.

When a participant navigated their avatar within a 10-meter radius of the hidden goal, a sphere begins to be rendered, and its size increases the closer the avatar is to the goal location. The trial ends successfully when the avatar makes contact with this sphere. Alternatively, the trials ends unsuccessfully if the participant goes 30 seconds without contacting the sphere. In either case, at the start of a new trial the location and orientation of the avatar is randomized, and the participant is instructed to find the sphere again.

The experiment consists of six blocks of 20 trials each. In each block, the first ten trials keep all environment properties fixed. After the 10th trial, depending on the condition of the block, one of five changes can take place. Note that at this point participants are notified by a message on-screen that a change may have taken place.

The five possible changes consist of the following: either the superficial appearance of the island and water changes, the goal location changes, the terrain changes by adjusting the fractal ground threshold up, or the terrain changes by adjusting the fractal ground threshold down, or no change takes place. In order to acquaint the participants with the fact that the environment changes, the first block is always a color change condition, and the next five consist of a random permutation of all five conditions, such that each participant experiences all conditions at least once during the experiment, and the color-change condition twice. See Figure 56 for an example of each of these change conditions.

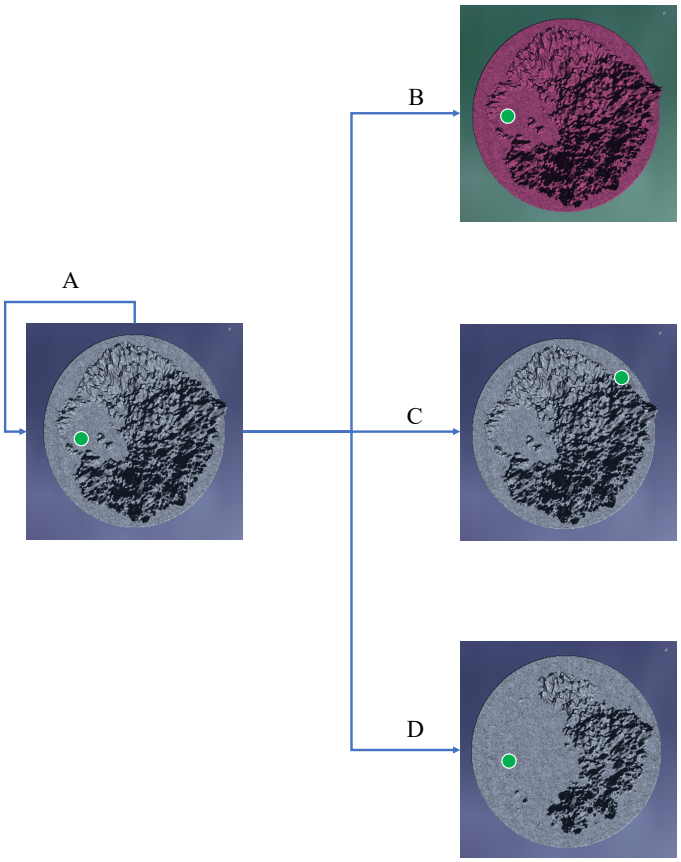

Figure 56: Visual representation of the four possible conditions within each block of trials. Green circle represents goal location. At beginning of each block, a random topography, goal location, and environment appearance are selected. After 10 trials a change takes place. A: no change. B: visual change. C: goal location change. D: topography change.

In addition to a different change condition, each block of trials contains a randomly selected seed and fractal dimension (D) with which to generate the terrain of the virtual

island. We used a set of 30 random seeds, and three different values of D, 1.2, 1.4, and 1.6. These were chosen based on previous research which demonstrated that humans were exceptionally poor at navigating environments consisting of a $D > 1.6$ (Juliani et al., 2016), and as such we would expect that they would be equally poor at this task. We also excluded environments with $D = 1$, as these would consist of a flat ground, and not be amenable to the manipulations required to impose the terrain change conditions. See Figure 57 for examples of the effect of varying the fractal dimension in three different random seeds.

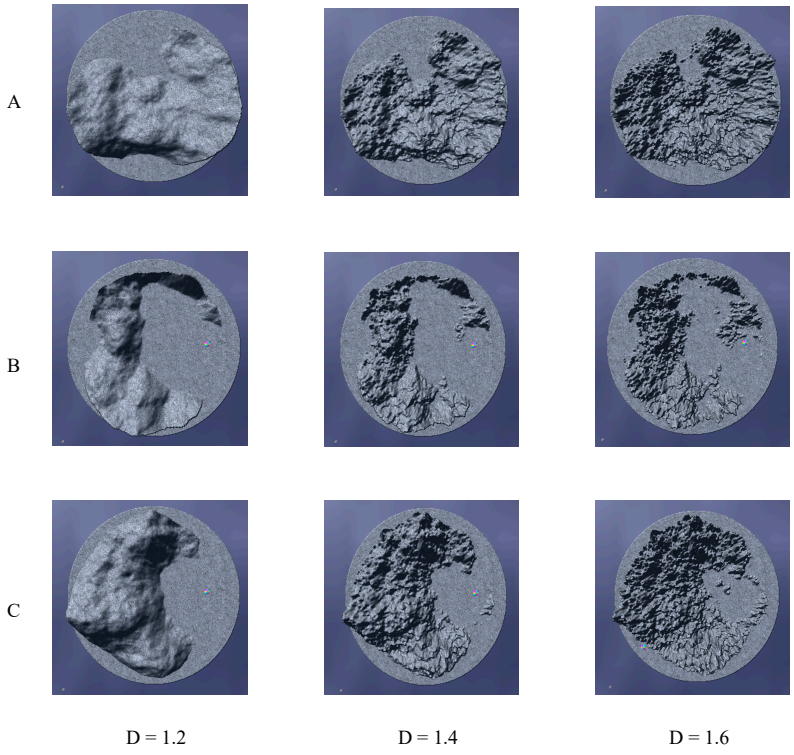

Figure 57: Examples of different seed used to generate three environment topographies each with different complexity levels. Rows: Different random initialization seeds. Columns: different values of D used to generate topographies.

Lastly, in addition to the fractal dimension and seed, the terrain is generated using a specific threshold value to determine the point at which there is flat ground as opposed to unnavigable terrain. This point is either 0.4 or 0.6, corresponding to more terrain and more ground, respectively. The terrain height for a given block of trials is selected such that in

terrain-less condition blocks, it is 0.4 in the first half of trials, and 0.6 in the second half. Likewise, in terrain-more condition blocks, it is 0.6 in the first half and 0.4 in the second half. In other condition blocks the height is randomly selected at the beginning of the block and held constant.

VI.2 Environmental Complexity and Human Navigation

The complexity of the environment has a meaningful impact on how humans are able to skillfully navigate. This disparity has both been demonstrated in Euclidean environments (O'Neill, 1992), and those composed of fractal topographies (Juliani et al., 2016). Specifically, in the case of fractal topographies, humans demonstrate relative optimal performance in environments with low-to-mid fractal dimension ($D = 1.2$ to $D = 1.4$), or complexity. One interpretation of these results is part of a fractal fluency theory whereby the human visual system shows improved information processing for patterns within this range. In addition to navigational performance, this preference has been demonstrated in aesthetic judgments (Taylor, Spehar, Hagerhall, & Van Donkelaar, 2011; Bies, Blanc-Goldhammer, et al., 2016), and discrimination and sensitivity (Spehar et al., 2015).

As part of a larger study exploring human navigational strategies during environmental change, we attempt to replicate the finding that humans are able to best navigate environments with a low-to-mid complexity. In the original work from Juliani et al. (2016), two navigation tasks were used, an object finding task and a map reading task. In the first case humans were able to most quickly find the goal object in the low-to-mid complexity environments, and in the second case they were able to make the most accurate judgments of goal location within the same range.

Here we use a slightly different task than map reading or object discovery. We employ a task inspired by the canonical Morris Water Maze (Morris et al., 1982). In this task, the participant must find a hidden goal location within the environment. Once they do so, they then are moved to a random location within the environment, and must return to

the location. The speed at which they return in subsequent trials determines the navigational performance. Participants complete a number of these sets of trials on environments with different fractal topographies consisting of varying fractal dimension. We find that participants are able to learn and remember the goal location best in environments with low-to-mid fractal complexity, thus providing additional evidence for the fractal fluency theory.

VI.2.1 Results

Overall, sixty-six participants completed the study. We removed five participants results from the analyzed data due to insufficient successful completion rates of the task, resulting in a total of sixty-one participants data being analyzed to compile results. We defined insufficient task completion as a failure to locate the goal in 25% or more trials. We believe that such participants were likely distracted or failed to properly attend to the task in the absence of a controlled experimental environment, as the median failure rate was 6%, and 90% of participants had a failure rate of 20% or less.

Among the remaining participants, we first turn to the question of understanding their ability to find and remember a goal location in the environment as a function of that environment itself. We find that there is a significant difference between participant performance in each of the three fractal dimensions ($F(2,6097) = 55.263, p < 0.001$). Measuring performance in time-to-goal (lower is better), we find that performance is best in the $D = 1.2$ environments ($Mean = 12.763, Std = 6.208$), followed by $D = 1.4$ environments ($Mean = 13.670, Std = 6.861$), followed by the $D = 1.6$ environments ($Mean = 14.998, Std = 7.182$). Figure 58 presents these results graphically.

To better understand the impact of the fractal dimension on the learning process over time, we further compare performance by fractal dimension at various stages of a given block of trials. We divide each block of 20 trials into four evenly distributed stages (1-5, 6-10, 11-15, 16-20). This allows us to examine how performances changes over time in

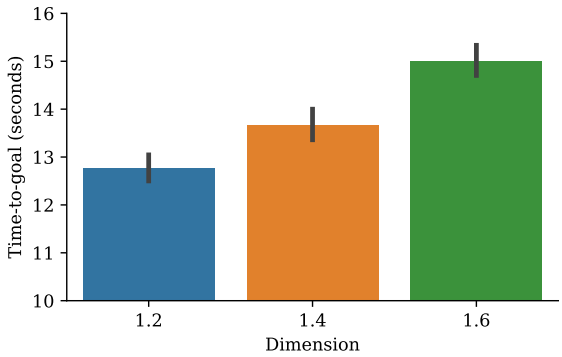

Figure 58: Mean human performance by fractal dimension. Lower time to goal corresponds to better navigation performance. Error bars correspond to standard error.

the environment, by compare early in a block to later in a block. See Figure 59 for these results.

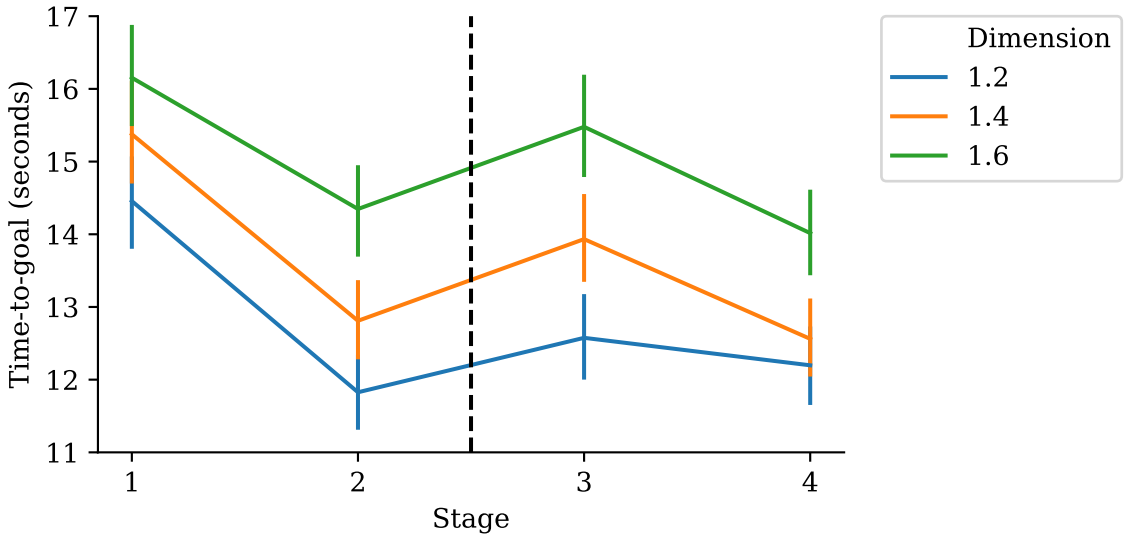

Figure 59: Mean human performance by fractal dimension in four stages of a single block. Lower time to goal corresponds to better navigation performance. Stage 1: Trials 1 - 5. Stage 2: Trials 6 - 10. Stage 3: Trials 11 - 15. Stage 4: Trials 16 - 20. Error bars correspond to standard error.

We find that the main effect of relative performance with respect to fractal dimension holds true. However, we additionally find that the differences between the two lower fractal dimensions ($D = 1.2$ and $D = 1.4$) appears to diminish throughout the block, and by Stage 4 is in fact no longer significantly different ($p = 0.339$). We expect that by Stage 4 the

participant will have the most experience with a given fractal topography, and with this experience participants are able to navigate both $D = 1.2$ and $D = 1.4$ environments with similar proficiency. This aligns with the findings of (Juliani et al., 2016) and conforms to the prediction of the fractal fluency theory that a value of $D = 1.3$ would correspond to optimal performance. In contrast, participants show additional difficulties with environments of $D = 1.6$ even at the end of a full block of trials.

We next examine an additional property of the fractal topographies, the threshold used to determine the ground level. As mentioned above, this level was either set to 0.4 or 0.6 depending on the condition of the block as well as a randomization process which ensured equal exposure to both levels for participants. We ask whether this value has an impact on participant performance in the task, and find that participants are significantly faster at completing a given trial in the 0.6 threshold level trials ($Mean = 13.438, Std = 6.805$) compared to the 0.4 level trials ($Mean = 14.198, Std = 6.836$) ($t(6098) = -4.352, p < 0.001$). We present these results graphically in Figure 60. This result suggests that the additional open space afforded by the higher threshold allowed participants to better localize themselves and the goal location, and navigate between the two.

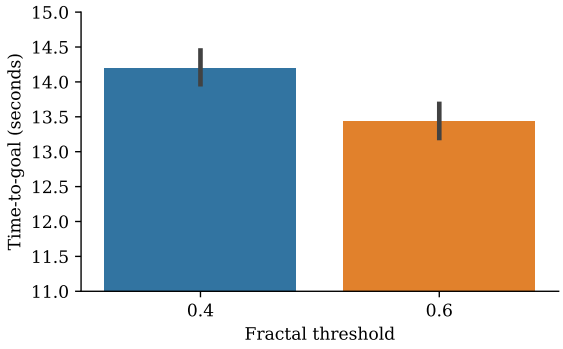

Figure 60: Mean human performance per trial by fractal height threshold. Lower time to goal corresponds to better navigation performance. Error bars correspond to standard error.

VI.3 Evidence for a Hybrid Behavioral Strategy in Humans

One approach to understanding human decision making has been to classify the 'algorithm' humans use to make decisions as being either a model-free or model-based strategy (Daw et al., 2005). A model-free strategy is one which conditions the current action on only the current state, whereas a model-based strategy would take additional information into account, typically information present in predicted future states, or explicit memory of past states (Niv, 2009; Sutton & Barto, 2018).

In recent years a third strategy has been proposed, a so-called hybrid decision making strategy, where key information about future states is cached and re-used, but a model of the entire environment need not be learned (Momennejad & Haynes, 2012). A popular instantiation of this hybrid approach has been the successor representation, and its applicability has been demonstrated both in a simple two-step decision making task (Momennejad et al., 2017), as well as a more complex navigational task (de Cothi, 2020).

The mark of such a successor-based decision making strategy is the dissociation between adaptation to changes in the goal state versus changes to the structure of the environment. In a successor learning paradigm, the reward function and successor representation are learned separately, and as a result an agent utilizing such a representation can adapt to changes to goal and structure separately. In comparison, a model-free agent would learn a joint value function, from which it is not possible to dissociate these two things. On the other end of the spectrum, a model-based learning agent would dissociate reward and structure, and would be able to adapt to both very rapidly, whereas a successor based agent would adapt more quickly to changes in goal than to changes in structure. This is due to the underlying successor representation being a statistical estimate, rather than a complete model as in the model-based case.

Here we utilize the experimental design consisting of blocks of trials with different change conditions to determine what kind of decision making strategy best matches human

behavior in a visually rich virtual navigational task. We find that humans are able to near-instantly adapt to changes in superficial visual content, but adapt to both changes in goal location and environment structure over a longer time course. Critically, we find that adaptation to goal location takes place faster, and with better final performance than adaptation to changes in environmental structure, suggesting that a successor-like representation may be guiding human behavior in this task.

VI.3.1 Results

We first compare the overall learning trend to validate that participants are indeed able to find the goal location, remember it, and deploy a successful navigation strategy for returning to it from multiple different locations. This trend is presented visually in 61. We find that participants indeed show signs of learning over the course of each block of trials.

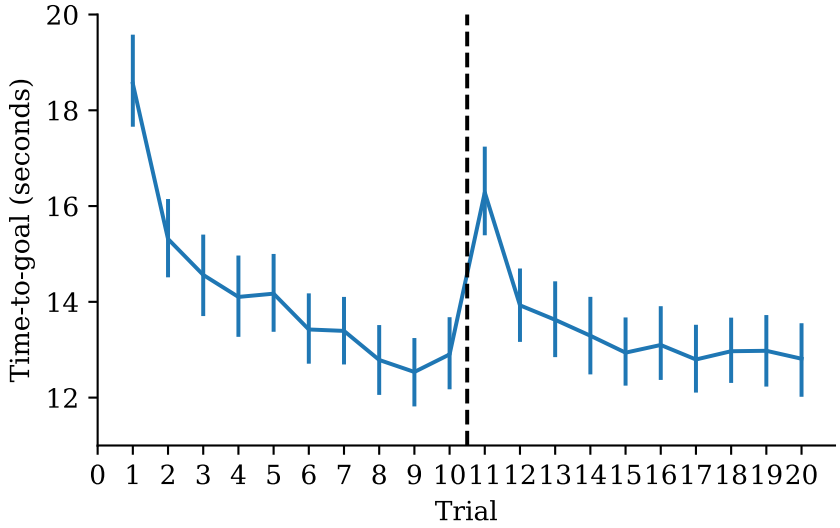

Figure 61: Mean human performance over time within a single block. Error bars correspond to standard error. Trial 11 corresponds to change trial.

We find that when distributing the trials into four stages (1-5, 6-10, 11-15, and 16-20), there is a significant decrease ($p < 0.001$) in time-to-goal between the first ($Mean = 15.34, Std = 7.52$) and second stages ($Mean = 13.0, Std = 6.36$). We furthermore find that the change in the environment halfway through the block disrupts performance, with

the third stage ($Mean = 14.01, Std = 6.96$) performance being significantly worse than the second stage ($p < 0.001$), but not as bad as the first stage ($p < 0.001$), suggesting that environmental knowledge is retained. Finally, we find that in the fourth stage ($Mean = 12.93, Std = 6.10$) performance is not significantly different from that of the second stage ($p = 0.75$), suggesting that participants are able to adapt to the changes.

We next turn our attention to the individual block conditions. We find that the participant's performance was significantly impacted by the condition of the trial block ($F(4, 6095) = 5.29, p < 0.001$). See Figure 62 for a graphical presentation of relative performance in each condition. Between these conditions, we find only four significant differences. The first is between the goal-change and no-change conditions ($p = 0.041$). The second is between the terrain-less and no-change conditions ($p = 0.015$). The last two are between terrain-more and visual-change ($p = 0.023$) and no-change ($p = 0.001$) conditions.

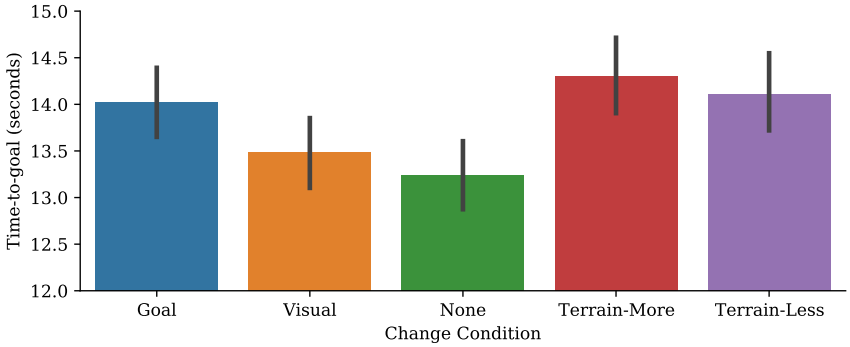

Figure 62: Mean human performance by block change condition. Lower time to goal corresponds to better navigation performance. Error bars correspond to standard error.

These results provide the following insights into the initial question regarding human decision making strategies. Due to the lack of difference between the visual-change and no-change conditions, we see that at the very least humans are not using an entirely reactive model-free policy, since they are on the whole able to ignore the superficial visual changes in the environment. Secondly, we find that the goal-change and terrain-change conditions do indeed disrupt performance compared to the no-change condition. This analysis alone however is not enough to provide evidence for either a hybrid or model-based strategy.

In order to determine that, we next turn to a more fine-grained analysis of the impact of condition on each stage of a block of trials.

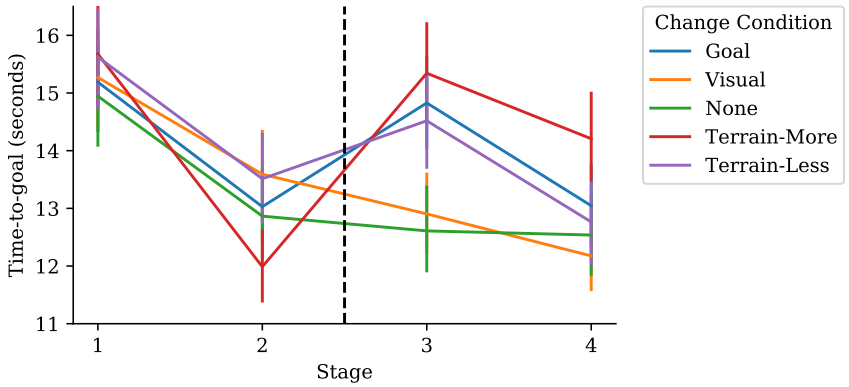

Figure 63: Mean human performance by block change condition. Lower time to goal corresponds to better navigation performance. Stage 1: Trials 1 - 5. Stage 2: Trials 6 - 10. Stage 3: Trials 11 - 15. Stage 4: Trials 16 - 20. Error bars correspond to standard error.

Figure 63 presents the participant performance over time for each of the block conditions. We find that in the first stage, there is no significant difference between conditions ($F(4,1520) = 0.50, p = 0.73$). In the second stage, we indeed find a significant difference between conditions ($F(4,1520) = 3.142, p = 0.013$), with the participants in the terrain-more condition being significantly better at finding the goal location than in the goal-change ($p = 0.04$), visual-change ($p = 0.001$), or terrain-less ($p = 0.003$) conditions. These results are perhaps not surprising, given that in the previous section we found that participants performed better in environments where the fractal threshold was set higher. As such, participants are able to better learns the task in the condition where all environments contain a high terrain threshold.

We next turn our attention to the second half of the block, and the second two stages. It is in this set of trials in which the environment change has taken place, that we expect greater effects. Indeed, we find a significant difference between conditions in stage three ($F(4,1520) = 9.49, p < 0.001$), with two distinct groups of conditions emerging. The first group consists of the no-change and visual-change conditions. The second consists of the goal-change and terrain-less and terrain-more conditions. All p-values between groups are

less than $p = 0.01$ and all p-values within groups are greater than $p = 0.1$. This suggests that changes in both the terrain and goal location disrupt the participants performance, whereas a change to the visual appearance of the environment does not.

Next, we examine the relative performance within each condition in the fourth stage of a given set of trials. We find that there is a significant difference between conditions ($F(4, 1520) = 4.97, p < 0.001$). Comparing the conditions, this difference comes primarily from the terrain-more condition, which participants performed significantly worse on this stage than all other conditions ($p < 0.05$). This result is again not surprising, since in the second half of trials in a terrain-more condition block, the environment terrain will have a lower threshold, which participants performed worse on overall.

We finally turn our attention to the original question regarding evidence for different behavioral strategies. Given the lack of impact from the visual change condition, we can rule out a purely model-free decision making strategy. This leaves two possibilities, a hybrid or model-based strategy. A hybrid strategy based on a successor representation would predict differences in learning between changes in the goal location and changes in the environment structure, with environment structure being more disruptive. We find some evidence for this, with the terrain-more condition resulting in a degraded performance which continues beyond the initial change (stage 3) and persists through the end of the block (stage 4). While we do not find a significant difference between the terrain-less and goal-change conditions in the final two stages, terrain-less should result in an easier environment to navigate, but instead is just as disruptive as the goal change.

Taken together, we believe that these results provide some additional evidence for a hybrid decision making strategy based on a successor-like representation which dissociated goal representation from environment representation. In the next section, we again turn to neural network modeling to provide a set of artificial agents with which to compare with the human decision makers presented here.

VI.4 Artificial Agent Behavior Varies with State Space Type

With an understanding of human behavior within the fractal island environments, we next turn to an examination of the behavior of artificial agents learning the same task. In previous chapters we demonstrated the ability of a series of generative temporal models to adapt to changes in both environmental appearance, content, and structure. Here we evaluate for all these of these together within a single environment, using a goal-directed navigational task as the metric of this performance.

As demonstrated in the previous section, humans can rapidly adapt to changes in environmental appearance, quickly adapt to changes in goal location, and more slowly adapt to changes in environmental structure. We demonstrate similar capabilities of artificial agents trained using the latent space of a TSWM on a task similar to that completed by the human participants. We find that the inferred state space s results in agents with the most human-like adaptation to environmental changes. In contrast, agents with a z state space, or agents using the ground-truth agent position and orientation information show greater disruption from goal-changes, and fail to fully adapt to such a change in goal location.

VI.4.1 Modeling Methods

In order to derive the candidate state spaces, we utilize the Tri-Stream World Model as described in Chapter V. Observations from the environment are rendered as $64 \times 64 \times 3$ color images, and the model utilizes a CNN encoder to infer the z, s, and c latent representations. The s and z state spaces from this model are then used as the input into separate reinforcement learning models. In addition, we define a third state space using the ground-truth spatial information concerning the agent's position and orientation within the island, and refer to this as the 'Spatial Info' state space.

All configurations of the 2D and 3D gridworld environments contained relatively small state spaces on the order of 10s or 100s of states. For example, in the case of a 3D gridworld

of size 7×7, there is a total of 196 states, counting each orientation and position combination. In contrast, even by discretizing the fractal island environment into 64×64 square meters, and discretizing the orientation into eight directions, there are a total of 32768 possible states. In order to address this state space which is orders of magnitude larger than earlier environments, we turn from linear reinforcement learning to deep reinforcement learning, which has been shown to be successful in learning policies even in environments with state spaces many orders of magnitude larger than those studied here (Silver et al., 2016). Specifically, we utilize a two-layer neural network for the policy and value function, and train this model using the popular Proximal Policy Optimization (PPO) algorithm (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). The size of the intermediate layer of this network is set to 128 unuts. While not biologically inspired, PPO is an actor-critic method, which remains a popular method for understanding the hippocampal-striatal axis (O'Doherty et al., 2004; Tessereau et al., 2020).

VI.4.2 Evaluation Methods

In order to tailor the behavioral task to an artificial agent, we implement a series of adjustments to the fractal island environment and task. The observations presented to the agent consist of $64 \times 64 \times 3$ color images representing a 90-degree field of view. The agents action space is simplified compared to that utilized by the human participants. The agent space consists of six possible actions: move forward, rotate left, rotate right, move forward and rotate left, move forward and rotate right, and move backward. This simplification is designed to make the learning problem easier, and to avoid the issue of representing the action space as a set of joint probability distributions. We also increase the effect of each of these actions relative to the result of the human participants pressing the keyboard keys, such that each agent action is equivalent to two consecutive button presses by the human. This is similar to "action repeat" used frequently in agent simulations of ATARI games (Mnih et al., 2015).

We furthermore modify the task itself in order to accommodate the artificial agents. While retaining the hidden goal aspect of the task, we remove the visible goal used in the human version of the task, and simply provide a $+1$ reward when the agent reaches within 4 meters of the goal location, and end the episode. This is done to ensure better consistency with all previous modeling experiments where the goal location was hidden. Furthermore, because the model is initially trained in an environment without any goals, the introduction of a visual goal during policy-learning time would result in disturbed latent representations due to out-of-distribution goal object observations.

Because of the extended time required to train a deep reinforcement learning policy compared to a linear policy, we evaluate on an environment derived from a single initialization seed (seed 0 in this case), and a fractal dimension of $D = 1.2$. We retain the policy of training each agent using five random initialization seeds in order to collect information about the distribution of learned behavior. Due to the agent being initialized with a random behavioral policy, we also provide the agent with the equivalent of double the amount of time each human received per-trial, corresponding to 300 agent time-steps. Finally, due to the inherently less efficient learning in the agents compared to humans, we provide the agents with 500 learning trials, with the change condition taking place after trial 250. Furthermore, a unique agent is trained per change condition. With three agent state spaces, five change conditions, and five repetitions per condition-state-space pair, a total of 75 agents are trained in all.

The z and s state spaces were derived from a TSWM trained for 7500 iterations in a dataset of 250 episodes of 50 time-steps each of a semi-random behavioral policy. The model's s and z latent states each consisted of eight gumbel-softmax distributions of size 16 each. As such, both latent state vectors were in total 128 units each. The TSWM was trained using a learning rate of $\alpha = 5e^{-4}$.

All agents were trained using a learning rate of $\alpha = 0.005$, and an entropy bonus of $\beta = 0.02$, which prevents premature convergence to sub-optimal policies during the learning

process, and a discount factor of $\gamma = 0.99$ to encourage long-term credit assignment. Each agent was trained for 500 episodes of a maximum of 300 time-steps each.

VI.4.3 Results

Before examining the behavior of the agents, it is worthwhile to analyze the learned representations z and s within the virtual fractal environment. Figure 64 presents example activations for these two sets of latent states, gather from an agent performing a random walk around the environment for 100 episodes of 50 time-steps each. We find that there is no local coherence in activation for units within the z space. In contrast, we find that there is high coherence for units in the s space, many with activation profiles consistent that of with place cells. The nature of these response profiles will be of relevance for interpreting the behavioral results presented below.

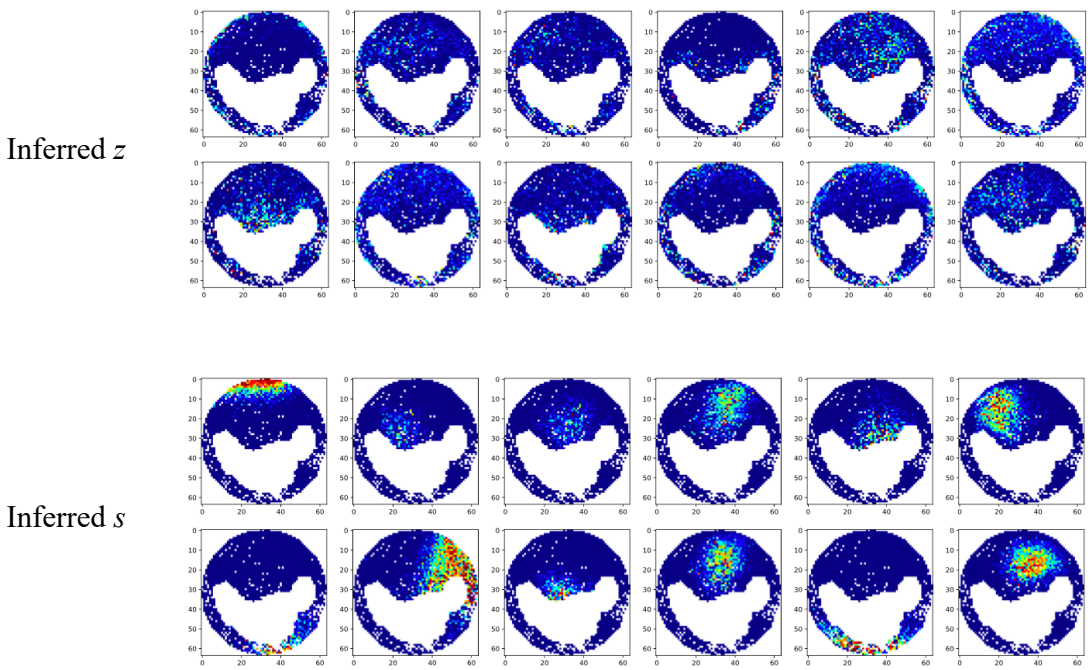

Figure 64: Activation profiles of first sixteen units of inferred latent s and z spaces in the TSWM model trained on a single fractal island topography.

Turning to the behavior of the trained agents, we first examine the impact of the state space type on agent performance over time during the learning process. We find that agents utilizing the 'Spatial Info,' 'Inferred z' and 'Inferred s' state space types all support learning the task, with each showing a significant decrease in time-to-goal between Stages 1 (Trials 1-125) and 2 (Trials 126-250) of the learning process (all $p < 0.001$). See Figure 65 for the relative performance of each set of agents during learning.

Having verified that learning does indeed take place for all agents in this task, we turn our attention to the second question, which is whether there are significant differences between agents with different state space types in the extent to which the learn to perform the hidden-goal task. We find that in both pre-change stages, the agents using an 'Inferred z' state space (Stage 1: $Mean = 111.26, Std = 45.47$; Stage 2: $Mean = 59.08, Std = 20.31$) significantly outperforms agents with either the 'Spatial Info' (Stage 1: $Mean = 154.80, Std = 54.01$; Stage 2: $Mean = 95.61, Std = 47.08$) or the 'Inferred s' (Stage 1: $Mean = 150.35, Std = 45.60$; Stage 2: $Mean = 86.63, Std = 24.61$) state space ($p < 0.001$).

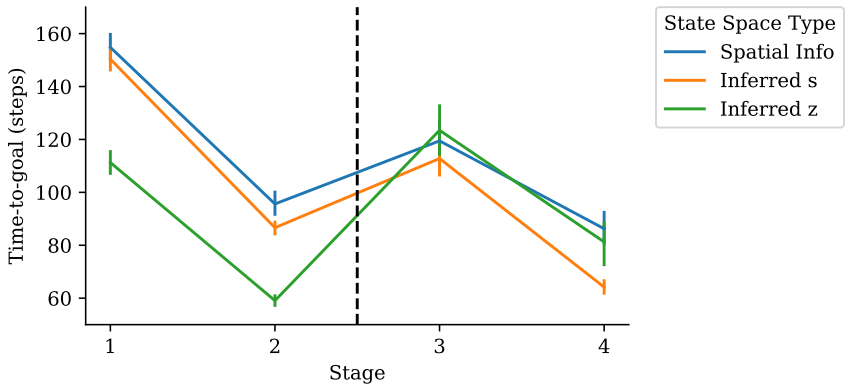

Figure 65: Mean agent performance with three different state spaces. Lower time to goal corresponds to better navigation performance. Stage 1: Trials 1 - 125. Stage 2: Trials 126 - 250. Stage 3: Trials 251 - 375. Stage 4: Trials 376 - 500. Error bars correspond to standard error.

Next we turn to the post-change conditions, first examining the result of environment changes on agent performance in Stage 3. We find that averaged over all change conditions,

agents with each of the three state types (Stage 3. Inferred s: $Mean = 112.78, Std = 69.76$; Inferred z: $Mean = 123.45, Std = 105.53$; Spatial Info: $Mean = 119.45, Std = 86.94$) are disrupted by the change, measured as a significant difference between Stage 2 and Stage 3 performance (all $p < 0.001$). Furthermore, we find no significant differences between the amount of disruption experienced by each set of agents ($F(2,372) = 0.46, p = 0.63$).

Finally, we analyzed the artificial agent's ability to recover their performance from the disruption caused by the environmental change, measured as a significant difference between Stage 2 and Stage 4 performance. We find that agents utilizing the 'Inferred s' (Stage 4. $Mean = 64.19, Std = 25.28; p < 0.001$) and 'Inferred z' (Stage 4. $Mean = 81.26, Std = 87.30; p = 0.016$) state spaces are able to recover in performance after the change, while the 'Spatial Info' (Stage 4. $Mean = 86.25, Std = 63.33; p = 0.25$) agents are not. We find however that the effect in the case of agents utilizing 'Inferred z' is relatively small, and further analysis shows that agents utilizing the 'Inferred s' state space outperform both agents with either the 'Inferred z' ($p < 0.035$) or 'Spatial Info' ($p < 0.006$) state spaces.

We next turn to a more in-depth analysis of the performance of the trained agents within each of the five different change conditions. Doing so allows us to better examine the source of the difference between the 'Inferred s' and other two state spaces in their ability to recover their performance after the environment change. Figure 66 presents the per-condition learning curves for agents utilizing each state type.

We find that qualitatively, the overall trends for the no-change and visual-change conditions are the same for agents with all three state spaces. In the case of both the no-change and the visual-change, there is no disruption from the change (or lack thereof), and likewise no need to recover from a disruption in performance. In both cases, the mean time-to-goal actually decreases between Stage 2 and Stage 3. We find that this trend matches that of the human participants, where there was no significant disruption from the change in visual appearance, or in the no-change condition.

Turning to the terrain change conditions, we find different response patterns for each

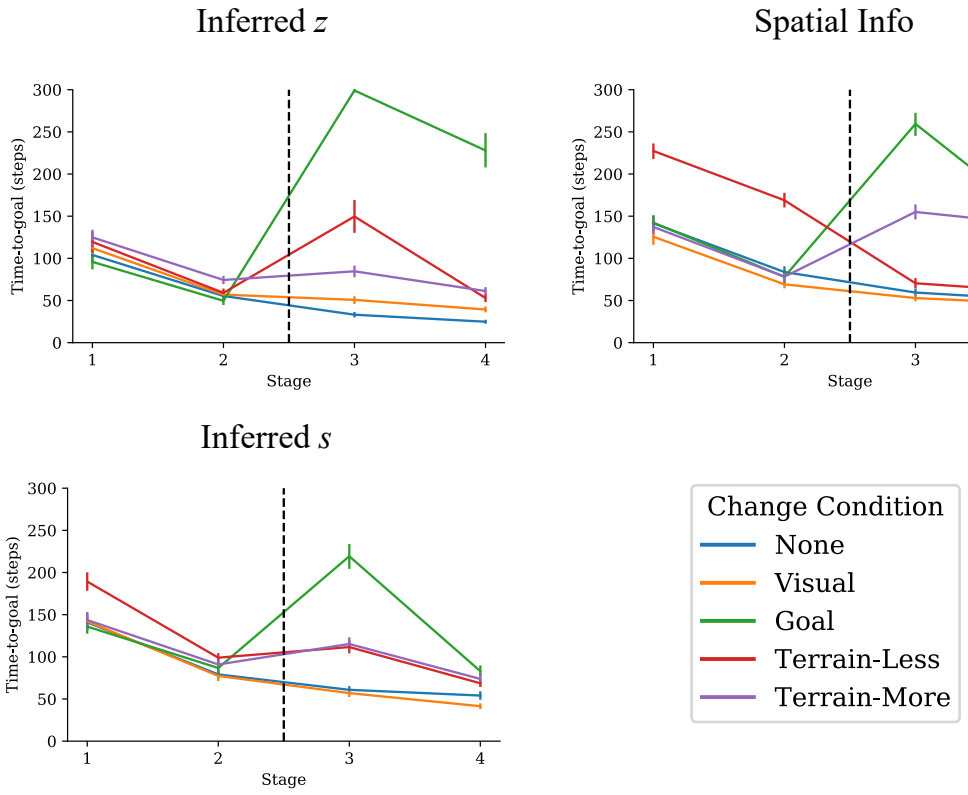

Figure 66: Mean agent performance within each change condition, and utilizing one of three different state spaces. Lower time to goal corresponds to better navigation performance. Stage 1: Trials 1 - 125. Stage 2: Trials 126 - 250. Stage 3: Trials 251 - 375. Stage 4: Trials 376 - 500. Error bars correspond to standard error.

of the three sets of agents. Returning to the results presented from human participants, disruption was found for both terrain-more and terrain-less conditions, with terrain-more being more disrupted overall. In the set of 'Spatial Info' state space agents, the terrain-more condition ($Mean = 155.024, Std = 37.35$) is more disruptive than the terrain-less condition ($Mean = 70.44, Std = 22.82$) ($p < 0.001$). In the 'Inferred s' set of agents, the terrain-more ($Mean = 115.22, Std = 30.32$) and terrain-less ($Mean = 111.60, Std = 31.73$) conditions are equivalently disruptive ($p = 0.737$). In the 'Inferred z' set of agents, the terrain-less condition ($Mean = 149.64, Std = 89.39$) is more disruptive than the terrain-more condition ($Mean = 84.64, Std = 26.85$) ($p < 0.001$). As such, none of the three sets of agents clearly resembles the human performance profile.

Finally, we turn to the goal-change condition. In this condition, human participants were significantly disrupted by the goal location changing, but recovered to pre-change performance levels by the end of the block of trials. We find that among the artificial agents, the goal-change condition results in significantly greater disruption than all other conditions for all three groups of agents ($p < 0.001$). Furthermore, we find that the agents utilizing the 'Inferred s' state space are able to fully recover from this disruption (Stage 2: $Mean = 86.76, Std = 30.24$; Stage 4: $Mean = 82.99, Std = 26.69$; $p = 0.764$), whereas agents utilizing either the 'Inferred z' (Stage 2: $Mean = 49.61, Std = 18.11$; Stage 4: $Mean = 227.95, Std = 99.14$; $p < 0.001$) or 'Spatial Info' (Stage 2: $Mean = 78.06, Std = 34.30$; Stage 4: $Mean = 136.81, Std = 90.68$; $p < 0.001$) state spaces are not. We can interpret these results as providing evidence that the agents utilizing the 'Inferred s' state space best match the behavior of the human participants.

VI.5 Discussion

In this chapter we sought to understand the behavior of both humans and artificial agents when performing a memory-based navigation task in a complex virtual environment. We found broadly that both humans and agents are able to learn to consistently navigate to a hidden goal location, doing so from a continuous stream of high-dimensional visual images presented to them.

We demonstrated that the structure of the environment has a significant impact on human performance, with lower fractal dimension topographies corresponding to participants reaching the goal location faster and more consistently. This can be seen as a partial replication of the results of Juliani et al. (2016), who found a similar trend on a set of topographies generated using the same methods described here. It can also be interpreted within the context of a larger body of work suggesting that humans respond to various visual stimuli of differing fractal dimensional with a general processing preference for stimuli consisting of a low-to-mid dimensional fractal (Spehar et al., 2015; Bies, Blanc-Goldhammer, et al.,

2016).

Using the results from the human participants, we also examined the effect various changes on the environment had to the learning process. We did this in order to find evidence for either a model-based, model-free, or hybrid decision making strategy. Previous work has found that humans navigate using a hybrid strategy (Daw et al., 2005; Momennejad & Haynes, 2012), which manifests as selective disruption to various kinds of environmental changes. This hybrid strategy has been modeled in the past using a successor representation learning algorithm (Momennejad et al., 2017; de Cothi, 2020). We find partial evidence for a hybrid strategy, with humans showing no disruption for visual changes, apparent disruptions for changes to the terrain and goal location, but importantly a consistent recovery from changes in goal location, but a less consistent recovery from changes in terrain.

Finally, we trained a set of artificial agents to perform a modified version of this task using the PPO algorithm. We proposed three different states spaces to use as input into these agents, one based on the pre-computed location and orientation of the agent, one based on the inferred 'where' state space s from a TSWM model, and one based on the inferred 'what' state space z from the same model. We found that all agents were able to perform the task well, but that only the agents trained using the inferred s latent state showed adaption to changes in goal consistent with human behavior.

While not directly analogous to the traditional means of classifying model-free, model-based, and hybrid behavioral strategies, there is a connection which can be made between these state spaces and these behavioral strategies. Rather than interpreting decision making strategy as being the result of an algorithm, we can interpret it as being the result of the representations utilized in a learning process. Here we compared three separate representations, each conveying different kinds of information to the agent.

The inferred z state space consists of an auto-encoded compressed representation of the observation, and thus can be interpreted as providing the basis for a purely reactive

policy mapping in the agent. As such, in order to account for changes to goal location, a complex series of mappings from visual features of the observation to predicted value need to be re-aligned. In contrast, the inferred s state contains spatial information, but unlike the 'spatial info' state, it contains information which is adapted to the structure of the environmental topography. Recall that these learned representations contain place-like firing properties, which show signs of a 'geodesic' representation, known to be found in place cells (Stachenfeld et al., 2017). We believe that it is this structural accommodation within the representation which enables the agents utilizing the state space to better adapt to all change conditions.

CHAPTER VII

GENERAL DISCUSSION AND CONCLUSION

Humans and other mammals are able to quickly make sense of their environment, and in doing so skillfully navigate their surroundings. The capacity to do so has long been connected to the notion of a cognitive map (Tolman, 1948), which has been proposed to be a major role of the hippocampus (O'Keefe & Nadel, 1978). In parallel, research into human episodic memory led to an understanding of the central role that the hippocampus also plays in memory encoding and retrieval (Tulving, 2002).

Recent theories connect these two functions under the notion of an experience-construction system (Hassabis & Maguire, 2009). In such a system, the dynamics of an environment are learned through experience, and then used to aid in both planning future actions in that environment, as well as in memory recall and imagination. All three of these abilities rest on the capacity of the hippocampus to spontaneously generate coherent trajectories of experience, a phenomenon referred to as replay (Foster, 2017), or preplay in the case of novel sequences (Dragoi & Tonegawa, 2011).

Simply referring to the hippocampus as an experience construction system however is insufficient, if we fail to define what an experience actually is. In most cases, experiences can be thought of as being tied to the perceptual, affective, and cognitive phenomena at a given delineated period of time. These phenomena are largely associated with the cortex, with an experience of visual perception being associated with the visual cortex, for example. It has been proposed that the role of the hippocampus is to provide a low-dimensional

index to these high-dimensional cortical states corresponding to phenomenal experiences (Teyler & DiScenna, 1986). Rather than learning the transition dynamics between entire cortical states, the hippocampus needs only to learn the transition dynamics which govern the indices, which correspond to a kind of grammar (Liu et al., 2018).

The theories of cognitive maps, episodic memory, experience-construction, and memory indexing provide a blue-print for the potential function of the hippocampus, and the broader medial temporal lobe within mammals. These theories also collectively describe a hypothetical system which bears a strong resemblance to a class of recent neural network models referred to as generative temporal models (Gemici et al., 2017; Ha & Schmidhuber, 2018). In their simplest form, generative temporal models contain a system by which observations from the environment are compressed into a latent state (indexing of episodic memories), a dynamics model is learned over these latent states (experience-construction), and these states and dynamics model is then used to guide goal-directed action (cognitive map). This work has provided a series of demonstrations by which such capacities can be realized by generative temporal models.

VII.1 Maps, Memories, and Models

This work has attempted to empirically demonstrate the connection between generative temporal models and the medial temporal lobe by presenting a series of models, and demonstrating their properties with respect to the theories outlined above. Starting with a simple world model, we demonstrated that place and time cells can be learned in an unsupervised fashion, and that these cells show activity patterns which are biased by the behavioral policy of the learning agent. We then demonstrated that dynamics models can be learned using these latent representations, and that the learned model displays temporal community, a key element of hippocampal representation (Schapiro et al., 2016). Furthermore, we showed that the process of latent state inference and generation within a generative temporal model can be connected to pattern separation and completion within

the hippocampus.

We next turned to the question of goal-directed navigation. Building from the actor-critic theory of learning in the dorsal and ventral striatum (O'Doherty et al., 2004), we demonstrated that the learned latent representations from a generative temporal model are useful as a basis function for performing reinforcement learning. Then, taking inspiration from more contemporary theories of striatal-hippocampal axis function (van der Meer et al., 2010), as well as evidence for a successor representation in CA1 of the hippocampus (Stachenfeld et al., 2017), we demonstrated that the learned latent space from a world model also enables successful learning using the successor representation in a goal-switch task. We next introduced a simple extension to the successor representation algorithm which enables it to be used with an extended class of basis functions, thus speeding up the learning process. Finally, we showed that the dynamics model of the world model additionally improves performance when used to provide Dyna-like updates (Sutton, 1991), which can be seen as a form of experience replay, similar to that which takes place spontaneously within the hippocampus (Pezzulo et al., 2014).

In the following chapter, we turned to an important aspect of cognitive maps, the ability to learn representations which are based on the structure of an environment, and invariant to that environment's content. In the visual system, context and content information are separated into separate streams, the dorsal and ventral streams, respectively. Within the medial temporal lobe, this separation takes place largely within the entorhinal cortex (Knierim, Neunuebel, & Deshmukh, 2014), with the lateral entorhinal cortex containing content information in the form of object-detecting cells (Deshmukh & Knierim, 2011), and the medial entorhinal cortex containing contextual information in the form of spatially selective grid cells (Hafting et al., 2005).

In order to capture this separation of content and structure, we presented a novel architecture, the Dual Stream World Model (DSWM), which separately encoded incoming observations from the environment into different latent representations. By separating the

streams, the ‘what’ latent representation was trained only to auto-encode the observation, while the ’where’ representation was trained to extract relevant spatial information from the observation. The ’where’ latent states were then used as keys, and the ’what’ latent states as values in a dictionary-based storage and retrieval system. Additionally, a forward model was learned over the ’where’ information, enabling generalization between environments with shared structure, but varying content. We evaluated this model on a set of 2D and 3D environments, demonstrating both the ability to generate more coherent trajectories than a single-stream model, but also a more useful representation for goal-directed navigation.

After demonstrating the capacity for content generalization with a DSWM, we next turned to the question of structural generalization. We first introduced a context latent variable into the world model, and demonstrated various methods for training this representation. First, we showed that the representation could learn to identify the environment index when trained on a fixed set of environment topographies, and that this representation was then useful for modeling the dynamics within the environments. We then demonstrated that the context representation could be trained to predict a 2D image of the environment topography, and that this led to greater performance when predicting the transition dynamics of environments.

With an understanding of the role that a contextual representation learned using a supervised loss signal could produce, we next introduced a fully unsupervised loss function to train the contextual representation, and demonstrated that it outperformed both supervised learning signals. We then extended this contextual representation to the DSWM model, introducing the Tri-Stream World Model (TSWM). We showed that this additional contextual representation can be interpreted as playing a similar role to that of the parahippocampal area, providing spatial context information useful for understanding transition dynamics in novel environments (R. A. Epstein, 2008).

With a fully realized generative temporal model, capable of generalization over changes in environment content, structure, and goal location, we then turned out attention back to

the biological systems which inspired this model, humans. We examined human navigation ability in a complex hidden goal navigation task in a visually rich 3D virtual environment. We first demonstrated that human performance in this task was impacted by the statistical structure of the environment topography in a way consistent with fractal fluency theory (Bies, Blanc-Goldhammer, et al., 2016; Juliani et al., 2016), with participants performing best in environments with low-to-mid level fractal complexity.

We next used the virtual navigation task to assess whether there was evidence for a hybrid decision making strategy when adapting to environment changes, as recently proposed by (Momennejad & Haynes, 2012; Momennejad et al., 2017; de Cothi, 2020). We found that humans are able to near-instantly adapt to changes in environment appearance, quickly adapt to changes in goal location, and more slowly adapt to changes in environment topography. Due to the difference between the disruption and adaptation profiles in the goal-change and terrain-change conditions, we can interpret these results as providing some evidence for a hybrid strategy.

In order to better understand these behavioral trends, and their relationship to various learning algorithms, we trained a series of artificial agents using Deep Reinforcement Learning to perform a modified version of the hidden-goal navigation task. We compared three different state space types, one based on the inferred z from a TSWM, one based on the inferred s from the same model, and one based on pre-computed location and orientation of the agent. We found that only the agents utilizing the inferred s representation showed signs of full adaption to the goal-change condition, showing a similar performance profile to that of humans.

Given the similarity of the inferred s latent space and the place cells found in mammals, along with the hypothesized role of place cells in guiding navigation, we believe that the specific properties of this representation may be essential to some of the findings suggesting that humans follow a hybrid decision making strategy. This is especially the case when we consider that a key property of both the inferred s latent state and the successor

representations utilized to model hybrid decision making strategies is their conformity to the topographical structure of an environment (Stachenfeld et al., 2017). We believe that this analysis can serve as the starting point for a novel approach to determining the kind of behavioral strategy being employed in a task. The nature of the representation being utilized to guide a decision making policy contains important priors about the environment which are just as important, if not more-so than the learning algorithm being used on top of these representations.

VII.2 Connections to Contemporary Modeling Research

The generative temporal models presented in the preceding chapters can be seen as a small subset of a growing class of models within the literature. While we largely focused on the popular World Model, introduced by Ha and Schmidhuber (2018), there are a number of other relevant models within the field. We chose the World Model for its popularity in the field of machine learning, its simplicity, and because the original work by Ha and Schmidhuber contained the basic building blocks of encoding into a latent state, learning the dynamics of the state, and then using those learned dynamics to learn a behavioral policy.

Since the introduction of the World Model, there have been a number of relevant advancements which have improved the adaptability and scalability of generative temporal models. As mentioned in the introduction, these fall into a few categories, depending on the nature of the task being learned. Two major themes include the introduction of memory augmentation, and the separation of the latent state into multiple separate variables.

In memory augmentation, an additional differentiable memory mechanism is used to store and retrieve latent states. This allows the model to quickly adapt to changes in the environment without the need for backpropagation to update the weights of the network, an often slower and more data intensive process. Within the literature, the nature of this memory mechanism has varied, with some model architectures adopting a simple differential

neural dictionary (Pritzel et al., 2017), such as the Generative Temporal Model with Spatial Memory (GTM-SM) (Fraccaro et al., 2018). Others have opted for more complex storage and retrieval mechanisms, such as the Memory-Based Predictor (MBP), which utilizes a differentiable memory store with multi-headed storage and retrieval mechanisms (Wayne et al., 2018). Still other models have sought to rely on more biologically plausible mechanisms such as a Hopfield Network for storage and retrieval of latent states, such as the Tolman-Eichenbaum Machine (TEM) (Whittington et al., 2019).

In the Dual-Stream and Tri-Stream world models, we chose a straightforward implementation of the differentiable neural dictionary (DND), described by Pritzel et al. (2017). We made this choice in order to avoid the more complex storage and look-up mechanisms used in the MBP, as well as to avoid the capacity limitations inherent in Hopfield networks. As such, the DSWM bears a resemblance to the GTM-SM, however we use a more structured latent representation for the key state, whereas in their work a simple two-dimensional vector is used which corresponds to the x and y coordinates of the agent location. By using an arbitrary learned latent state for the key, our model can be applied to both spatial and non-spatial environments, as well as be applied to downstream linear RL tasks. By not using more complex storage mechanisms with multiple read and write heads, such as the MBP, our model stores more redundant information, and can only access one relevant experience at a time. In the experiments presented in this work, this limitation does not present an issue, but in more realistic environments, where many different memories need to be stored corresponding to different events which take place in the same location, our retrieval mechanism would likely under-perform relative to these other models.

The second major theme has been the separation of the latent state into multiple separate latent variables. Doing so enables each latent variable to represent a unique subset of the entire latent state, and enables novel model architectures which can deal with each aspect of the state in a unique way. For example, the Recurrent State Space Model (RSSM) (Hafner et al., 2018) introduces both a discrete and stochastic latent state, enabling

the model to separately model known and unknown aspects of the environment dynamics independently. Multiple latent states have also been utilized within hierarchical models, where lower-level latent states help to condition higher-level states, such as in the Stochastic Latent Actor Critic Model (SLAC) (A. X. Lee, Nagabandi, Abbeel, & Levine, 2019). Aside from separating the variables based on hierarchy or stochasticity, the latent states can also be separated based on the type of environmental information being stored, such as in the TEM (Whittington et al., 2019), and GTM-SM (Fraccaro et al., 2018), where content and context variables are modeled separately, in both cases enables content-based generalization.

We chose to focus on the separation of latent states based on content (what), context (how), and location (where). As such, our model is similar to the TEM and GTM-SM models. Doing so enables the model to separately learn to encode each of these three variables from the stream of incoming observations, and as such to generalize over changes within the distributions of each of them. This generalization takes the form of adaptability to changes in the content of an environment with the same structure, as well as the ability to adapt to changes in the structure of the environment itself. Both the TEM and GTM-SM models demonstrate content generalization, but not structural generalization, which is a more difficult problem. While Chapter V presents initial results toward structural generalization, we note that the improvements from the contextual latent state are relatively modest, and the problem remains not fully solved. While both our TSWM and the GTM-SM model demonstrate learning from high-dimensional egocentric visual observations, TEM was demonstrated using only low-dimensional "toy" problems. Despite this, TEM has been shown to match real biological data in terms of both the presence of grid cells as well as place cells, something not shown in our work.

Lastly, we want to address the choice of latent distribution in all of the models presented in this work. All related contemporary models which we have discussed up to this point have either utilized a gaussian or deterministic latent state. This likely follows because

of the great initial success demonstrated by the use of variational auto-encoders with a gaussian latent distribution (Kingma & Welling, 2013). This choice is not unfounded, as it has recently been demonstrated that representations in primate inferotemporal cortex can be modeled using a gaussian VAE trained with a disentanglement loss function (Higgins et al., 2020).

Instead of gaussian distributions, we chose to utilize gumbel-softmax distributions for all latent states within our models. We were motivated in this decision by recent theoretical work which proposed that the medial temporal lobe is involved in clustering high-dimensional state information in the cortex (Mok & Love, 2019). While Mok and Love (2019) utilize a non-differentiable k-means clustering algorithm, we opted for the gumbel-softmax distribution due to its capacity to be used as a latent distribution within a fully differentiable neural network (Jang et al., 2016). The efficacy of a gumbel-softmax distribution for learning a latent state space has also been previously demonstrated in a generative temporal model on a simple T-Maze navigation task (Corneil, Gerstner, & Brea, 2018). We hypothesized that such a distribution would enable a "soft" probabilistic form of state grouping, similar to the potential functional role of time and place cells.

From a practical perspective, we also found that for simple auto-encoding tasks, models utilizing the gumbel-softmax distribution outperform those utilizing a gaussian distribution, as described in Chapter II. More importantly, the use of this distribution allows for the natural development of time or place cells, depending on the nature of the observation stream being learned by the model. We find that this is an inherent property of the distribution, with such cell types always developing under a variety of conditions. This is in contrast to recent modeling work showing the development of grid cells, where highly specific hyperparameters and activation functions are needed, and are difficult to reproduce (Banino et al., 2018; Sorscher, Mel, Ganguli, & Ocko, 2019).

VII.3 Biological Implications and Open Questions

The connection between the medial temporal lobe and the class of neural networks known as generative temporal models presented here is a starting point for a much more in-depth set of potential future analyses. The connections drawn in this work raise a number of relevant questions regarding the biological plausibility of the models presented here, as well as pose potential research questions which could be explored within the context of empirical biological research.

The first question which can be asked is regarding the nature of the gumbel-softmax distribution as a basic building block of hippocampal representation. We have demonstrated here that this distribution induces place and time-like cells when used in a model trained to perform auto-encoding of spatial and temporal information, respectively. This auto-encoding process can be interpreted as being part of the hippocampal indexing system (Teyler & DiScenna, 1986). In particular, there is evidence that the dentate gyrus within the medial temporal lobe contains sparse connections, from the entorhinal context and the CA3 region of the hippocampus (Leutgeb et al., 2007). These sparse connections have been referred to as performing pattern separation, and we find evidence of this in induced gumbel-softmax representations in the experiments presented here. It would be possible through empirical research to verify whether the induced representation by this pathway matches more precisely the properties of a gumbel-softmax or similar distribution.

The next relevant question extends from our specific implementation of the hippocampal indexing theory within the context of a generative temporal model. By utilizing such a probabilistic model, we inherently arrive at an interpretation of the hippocampal representations within the context of inferred and generated latent variables. We have proposed that this can be seen to map onto the dentate gyrus, CA3, and CA1 regions of the hippocampus. The inference process thus taking place is as follows: latent state is inferred from information within the entorhinal cortex, made sparse (and "pattern separated") by dentate

gyrus, represented in CA3, with a prediction of the future latent state ("pattern completion") generated within CA1. The properties of place cells in CA3 and CA1 provide some evidence for this, as they match the induced distributions from the inferred and generated latent states in the models we have presented here. Indeed, this theory has been recently proposed by simultaneous other work (Sanders, Wilson, & Gershman, 2020). Further biological recording could be done within DG, CA3, and CA1 regions to determine whether the place cells within this region best match those of inferred and generated latent variables from a generative temporal model.

Directly related to the question of hippocampal inference is the phenomena of remapping (Fyhn, Hafting, Treves, Moser, & Moser, 2007), whereby large changes in the structure of appearance of the environment induce a new set of place cells to fire. Within the context of the theoretical models discussed here, this can be seen as a unique state space being instantiated. While we did not directly address remapping in the work presented, there are potential extensions which would make the study of this phenomena possible. We believe that an unsupervised loss function for the s latent space (as opposed to the supervised spatial loss function demonstrated) within the DSWM could lead to remapping of s. Furthermore, we explore inferring s using only integration within a recurrent neural network. It is a promising avenue of research to explore the extent to which neural networks which allow for greater conditioning, such as hypernetworks or networks with fast-weights would better adapt to this problem (Ha et al., 2016). Indeed, models such as those presented in (Whittington et al., 2019) and (Sanders et al., 2020), thus demonstrating its possibility within a generative temporal model.

Through our analysis of context-augmented generative temporal models, we drew a connection between a latent representation developed to extend the expressibility of the dynamics model and the parahippocampal area. This connection was based on the evidence that the parahippocampal area responds preferentially to stimuli which provide spatial contextual information (R. A. Epstein, 2008). The hypothesis we put forth is that this

area integrated these contextual spatial cues from sensory information in order to aid and modify the dynamics model represented within the hippocampus itself. This would imply a dissociation between state information represented within the hippocampus itself, and contextual information represented within the parahippocampal area. The implicit contextual model we presented here can be seen as one possible implementation of this system, and serve as the basis for a prediction of biological function.

Moving beyond the medial temporal lobe, we also presented a potential novel model of the hippocampal-striatal axis. The traditional interpretation of this system has been within the context of an actor-critic model, where the hippocampus provided the state representation, the ventral striatum the value estimation, and the dorsal striatum the policy (O'Doherty et al., 2004). Using recent work suggesting that the outgoing CA1 representation from the hippocampus is best modeled using a successor representation (Stachenfeld et al., 2017), we proposed that the ventral striatum may act as a reward representation as opposed to a value representation. Thus, the hippocampal-striatal axis could be thought of as implementing a successor learning algorithm, as opposed to an actor-critic algorithm. We find some additional biological evidence for this in empirical work showing that the ventral striatum learns a more local representation of value which may be more in-line with reward identification or prediction than a traditional notion of value as predicted discounted reward (van der Meer et al., 2010). To fully test this hypothesis, a more detailed study of the role of the dorsal and ventral striatum in learning is necessary, as a successor-based theory of policy learning would make specific predictions about the nature of the induced policy. For example, we have utilized a cosine similarity metric to measure state similarity, and thus determine the reward and value function values. Such a mathematical operation can be directly tested for.

The hippocampus has a number of additional downstream connections beyond the striatum. One group of particular interest are the more frontal regions. Both the medial prefrontal cortex (mPFC) and the orbital frontal cortex (OFC) have been studied in their re-

lationship to the hippocampus. In the case of the former, it has been demonstrated that mPFC provides a goal-like signal to the hippocampal region (Ito, Zhang, Witter, Moser, & Moser, 2015). Such a signal could help determine the specific nature of hippocampal replay events, biasing the generated trajectories towards states known to currently be salient, or of interest to higher-level attention. Such a system was formalized in a model by Erdem and Hasselmo (2012). In the case of the latter region, it has been shown that the OFC is involved in value estimation, and represents states at a more abstracted level than that of the hippocampus (Wikenheiser & Schoenbaum, 2016). The interaction of both regions, while different in their purpose, both point to the mutual notion of hierarchical representation. By representing state spaces at higher levels of abstraction than what is possible within the hippocampus, animals are able to reason over longer spans of time, and do so in ways more generalizable to diverse circumstances. We see modeling these dynamics within the context of generative temporal models presented here as an intriguing future direction.

A last major question is the plausibility of a differentiable neural dictionary and recurrent neural network for representing the latent state generation ("pattern completion") process induced by the CA3 region of the hippocampus. We recognize that this modeling choice was made largely for computational convenience, rather than biological plausibility. One alternative used in related work is the Hopfield network, which was originally inspired by the hippocampus (Hopfield & Tank, 1985; Whittington et al., 2019). While the original Hopfield networks had restrictive computational limitations with respect to the number of possible patterns which they could store and retrieve, modern versions of these networks have been able to enable the storage and retrieval of orders of magnitude more patterns, and the successful application to real-world problems such as natural language text generation (Ramsauer et al., 2020). We believe that such networks represent a promising future direction, both for modeling the hippocampus, but also as a potential source of additional hypotheses regarding the computational properties of the hippocampus itself.

VII.4 Conclusion

The function of the hippocampus and medial temporal lobe can seem miraculous. As humans we are able to not only recall a seemingly vast amount of memories from our childhood to today, but we are also able to put these memories into contexts, and create narratives out of them. These narratives are both the re-telling of the past, but also serve to help create new stories and to plan out possible future events. The coherence of these plans then makes possible skillful navigation and action within our ever-changing world in order to realize them. The models presented here represent a modest attempt at formalizing the system which makes this possible. It is our hope that this formalism can help to provide a common language for future developments within the fields of both neuroscience and machine learning, as they both continue to develop, providing reciprocal insights into both the nature of biological and artificial intelligence.

REFERENCES CITED

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and brain sciences*, *22*(3), 425–444.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., & O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nature neuroscience*, *10*(1), 126–131.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... others (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., & Silver, D. (2017). Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems* (pp. 4055–4065).

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Bellemare, M., Dabney, W., Dadashi, R., Taiga, A. A., Castro, P. S., Le Roux, N., ... Lyle, C. (2019). A geometric perspective on optimal representations for reinforcement learning. In *Advances in neural information processing systems* (pp. 4358–4369).

Bies, A. J., Blanc-Goldhammer, D. R., Boydston, C. R., Taylor, R. P., & Sereno, M. E. (2016). Aesthetic responses to exact fractals driven by physical complexity. *Frontiers in human neuroscience*, *10*, 210.

Bies, A. J., Boydston, C. R., Taylor, R. P., & Sereno, M. E. (2016). Relationship between fractal dimension and spectral scaling decay rate in computer-generated fractals. *Symmetry*, *8*(7), 66.

Bliss, T. V., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, *361*(6407), 31.

Burgess, N., Barry, C., & O'keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, *17*(9), 801–812.

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron*, *87*(3), 507–520.

Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology*, *20*(6), 544–547.

Chadwick, M. J., Jolly, A. E., Amos, D. P., Hassabis, D., & Spiers, H. J. (2015). A goal direction signal in the human entorhinal/subicular region. *Current Biology*, *25*(1), 87–92.

Climer, J. R., Newman, E. L., & Hasselmo, M. E. (2013). Phase coding by grid cells in unconstrained environments: two-dimensional phase precession. *European Journal of Neuroscience*, *38*(4), 2526–2541.

Corneil, D., Gerstner, W., & Brea, J. (2018). Efficient model-based deep reinforcement learning with variational state tabulation. *arXiv preprint arXiv:1802.04325*.

Cueva, C. J., & Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*.

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron*, *63*(4), 497 - 507. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0896627309005820` doi: https://doi.org/10.1016/j.neuron.2009.07.027

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

de Cothi, W. J. (2020). *Predictive maps in rats and humans for spatial navigation* (Unpublished doctoral dissertation). UCL (University College London).

Deshmukh, S. S., & Knierim, J. J. (2011). Representation of non-spatial and spatial information in the lateral entorhinal cortex. *Frontiers in behavioral neuroscience*, *5*, 69.

Deuker, L., Bellmund, J. L., Schröder, T. N., & Doeller, C. F. (2016). An event map of memory space in the hippocampus. *Elife*, *5*, e16534.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657.

Dragoi, G. (2020). Cell assemblies, sequences and temporal coding in the hippocampus. *Current Opinion in Neurobiology*, *64*, 111–118.

Dragoi, G., & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, *469*(7330), 397.

Dragoi, G., & Tonegawa, S. (2013). Distinct preplay of multiple novel spatial experiences in the rat. *Proceedings of the National Academy of Sciences*, *110*(22), 9100–9105.

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), 732.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, *12*(10), 388–396.

Erdem, U. M., & Hasselmo, M. (2012). A goal-directed spatial navigation model using forward trajectory planning based on grid cells. *European Journal of Neuroscience*, *35*(6), 916–931.

Fiete, I. R., Burak, Y., & Brookings, T. (2008). What grid cells convey about rat location. *Journal of Neuroscience*, *28*(27), 6858–6871.

Foster, D. (2017). Replay comes of age. *Annual review of neuroscience*, *40*, 581–602.

Foster, D., Morris, R., & Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, *10*(1), 1–16.

Foster, D., & Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680.

Fraccaro, M., Rezende, D. J., Zwols, Y., Pritzel, A., Eslami, S., & Viola, F. (2018). Generative temporal models with spatial memory for partially observed environments. *arXiv preprint arXiv:1804.09401*.

Frank, L. M., Stanley, G. B., & Brown, E. N. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., & Moser, E. I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, *446*(7132), 190–194.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *Elife*, *6*, e17086.

Gemici, M., Hung, C.-C., Santoro, A., Wayne, G., Mohamed, S., Rezende, D. J., ... Lillicrap, T. (2017). Generative temporal models with memory. *arXiv preprint arXiv:1702.04649*.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553–1568.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... others (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695–705.

Ha, D., Dai, A., & Le, Q. V. (2016). Hypernetworks. *arXiv preprint arXiv:1609.09106*.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2018). Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104*(5), 1726–1731.

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1263–1271.

Hasselmo, M. E. (2009). A model of episodic memory: mental time travel along encoded trajectories using grid cells. *Neurobiology of learning and memory*, *92*(4), 559–573.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2020). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. *arXiv preprint arXiv:2006.14304*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. *International conference on learning representations*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hollup, S. A., Molden, S., Donnett, J. G., Moser, M.-B., & Moser, E. I. (2001). Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *Journal of Neuroscience*, *21*(5), 1635–1644.

Hopfield, J. J., & Tank, D. W. (1985). "neural" computation of decisions in optimization problems. *Biological cybernetics*, *52*(3), 141–152.

Horner, A. J., Bisby, J. A., Zotow, E., Bush, D., & Burgess, N. (2016). Grid-like processing of imagined navigation. *Current Biology*, *26*(6), 842–847.

Howard, L., Javadi, A., Yu, Y., Mill, R., Morrison, L., Knight, R., ... Spiers, H. (2014). The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. *Current Biology*, *24*(12), 1331 - 1340. Retrieved from http://www.sciencedirect.com/science/article/pii/S0960982214005260 doi: https://doi.org/10.1016/j.cub.2014.05.001

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological review*, *112*(1), 75.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I., & Moser, M.-B. (2015). A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation. *Nature*, *522*(7554), 50.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience*, *10*(1), 100.

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, *18*(9), 1163–1171.

Juliani, A. W., Berges, V.-P., Vckay, E., Gao, Y., Henry, H., Mattar, M., & Lange, D. (2018). Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*.

Juliani, A. W., Bies, A. J., Boydston, C. R., Taylor, R. P., & Sereno, M. E. (2016). Navigation performance in virtual environments varies with fractal dimension of landscape. *Journal of environmental psychology*, *47*, 155–165.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature neuroscience*, *10*(12), 1625.

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature neuroscience*, *12*(7), 913.

Kay, K., Chung, J. E., Sosa, M., Schor, J. S., Karlsson, M. P., Larkin, M. C., ... Frank, L. M. (2020). Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell*, *180*(3), 552–567.

Kim, T., Ahn, S., & Bengio, Y. (2019). Variational temporal abstraction. In *Advances in neural information processing systems* (pp. 11570–11579).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Knierim, J. J., Neunuebel, J. P., & Deshmukh, S. S. (2014). Functional correlates of the lateral and medial entorhinal cortex: objects, path integration and local–global reference frames. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1635), 20130369.

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), 217.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS biology*, *7*(8), e1000173.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194.

Lee, A. X., Nagabandi, A., Abbeel, P., & Levine, S. (2019). Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*.

Lehn, H., Steffenach, H.-A., van Strien, N. M., Veltman, D. J., Witter, M. P., & Håberg, A. K. (2009). A specific role of the human hippocampus in recall of temporal sequences. *Journal of Neuroscience*, *29*(11), 3475–3484.

Leutgeb, J. K., Leutgeb, S., Moser, M.-B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and ca3 of the hippocampus. *science*, *315*(5814), 961–966.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, *29*(31), 9771–9777.

Liu, K., Sibille, J., & Dragoi, G. (2018). Generative predictive codes by multiplexed hippocampal neuronal tuplets. *Neuron*, *99*(6), 1329–1341.

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737–749.

Mandelbrot, B. B. (1983). *The fractal geometry of nature* (Vol. 173). WH freeman New York.

Marr, D., Willshaw, D., & McNaughton, B. (1991). Simple memory: a theory for archicortex. In *From the retina to the neocortex* (pp. 59–128). Springer.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, *7*(8), 663.

McNaughton, B. L., Chen, L., & Markus, E. (1991). "dead reckoning," landmark learning, and the sense of direction: a neurophysiological and computational hypothesis. *Journal of Cognitive Neuroscience*, *3*(2), 190–202.

Mehta, M. R., Quirk, M. C., & Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*(3), 707–715.

Mittelstaedt, M.-L., & Mittelstaedt, H. (1980). Homing by path integration in a mammal. *Naturwissenschaften*, *67*(11), 566–567.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature communications*, *10*(1), 1–9.

Momennejad, I., & Haynes, J.-D. (2012). Human anterior prefrontal cortex encodes the 'what'and 'when'of future intentions. *Neuroimage*, *61*(1), 139–148.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*, e32548.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, *13*(1), 103–130.

Morris, R. G., Garrud, P., Rawlins, J. a., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*(5868), 681.

Muller, R. U., & Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, *7*(7), 1951–1968.

Muller, R. U., Kubie, J. L., & Ranck, J. B. (1987). Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *Journal of Neuroscience*, *7*(7), 1935–1950.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences*, *97*(20), 11120–11124.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, *304*(5669), 452–454.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental neurology*, *51*(1), 78–109.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife*, *4*, e06063.

O'Neill, M. J. (1992). Effects of familiarity and plan complexity on wayfinding in simulated buildings. *Journal of Environmental Psychology*, *12*(4), 319–327.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*(5894), 1322–1327.

Peng, J., & Williams, R. J. (1993). Efficient learning and planning within the dyna framework. *Adaptive Behavior*, *1*(4), 437–454.

Pennartz, C., Ito, R., Verschure, P., Battaglia, F., & Robbins, T. (2011). The hippocampal–striatal axis in learning, prediction and goal-directed behavior. *Trends in neurosciences*, *34*(10), 548–559.

Peters, J., & Büchel, C. (2010). Neural representations of subjective reward value. *Behavioural brain research*, *213*(2), 135–141.

Pezzulo, G., Kemere, C., & Van Der Meer, M. A. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Annals of the New York Academy of Sciences*, *1396*(1), 144–165.

Pezzulo, G., van der Meer, M. A., Lansink, C. S., & Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in cognitive sciences*, *18*(12), 647–657.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74.

Poucet, B., & Hok, V. (2017). Remembering goal locations. *Current Opinion in Behavioral Sciences*, *17*, 51 - 56. Retrieved from `http://www.sciencedirect.com/science/article/pii/S2352154616302832` (Memory in time and space) doi: https://doi.org/10.1016/j.cobeha.2017.06.003

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764–R773.

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., ... Blundell, C. (2017). Neural episodic control. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 2827–2836).

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., ... others (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, *13*(9), e1005768.

Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*(15), 5900–5920.

Sanders, H., Wilson, M. A., & Gershman, S. J. (2020). Hippocampal remapping as hidden state inference. *Elife*, *9*, e51140.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M.-B., & Moser, E. I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, *312*(5774), 758–762.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience*, *16*(4), 486–492.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11.

Silva, D., Feng, T., & Foster, D. J. (2015). Trajectory events across hippocampal place cells require previous experience. *Nature neuroscience*, *18*(12), 1772.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484.

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, *322*(5909), 1865–1868.

Sorscher, B., Mel, G., Ganguli, S., & Ocko, S. (2019). A unified theory for the origin of grid cells through the lens of pattern formation. In *Advances in neural information processing systems* (pp. 10003–10013).

Spehar, B., Wong, S., van de Klundert, S., Lui, J., Clifford, C. W. G., & Taylor, R. (2015). Beauty and the beholder: the role of visual sensitivity in visual preference. *Frontiers in human neuroscience*, *9*, 514.

Spiers, H. J., & Maguire, E. A. (2007). A navigational guidance system in the human brain. *Hippocampus*, *17*(8), 618-626. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.20298` doi: 10.1002/hipo.20298

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643.

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nature Neuroscience*, *23*(5), 651–663.

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tanji, J., & Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current opinion in neurobiology*, *11*(2), 164–170.

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, *10*(2), 420–435.

Taylor, R., Spehar, B., Hagerhall, C., & Van Donkelaar, P. (2011). Perceptual and physiological responses to jackson pollock's fractals. *Frontiers in human neuroscience*, *5*, 60.

Tessereau, C., O'Dea, R., Coombes, S., & Bast, T. (2020). Reinforcement learning approaches to hippocampus-dependant flexible spatial navigation. *bioRxiv*.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral neuroscience*, *100*(2), 147.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., & Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*(7721), 57–62.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual review of psychology*, *53*(1), 1–25.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, *8*(3), 198–204.

van der Meer, M. A., Johnson, A., Schmitzer-Torbert, N. C., & Redish, A. D. (2010). Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron*, *67*(1), 25–32.

Van Essen, D. C., & Maunsell, J. H. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, *6*, 370–375.

Van Hoesen, G. W. (1982). The parahippocampal gyrus: new observations regarding its cortical connections in the monkey. *Trends in neurosciences*, *5*, 345–350.

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., ... Daw, N. D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron*.

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

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., ... others (2018). Unsupervised predictive memory in a goal-directed agent. *arXiv preprint arXiv:1803.10760*.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2019). The tolman-eichenbaum machine: Unifying space and relational memory through generalisation in the hippocampal formation. *bioRxiv*, 770495.

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513.

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation*, *2*(4), 490–501.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676–679.

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in neurosciences*, *34*(10), 515–525.

Zaehle, T., Jordan, K., Wüstenberg, T., Baudewig, J., Dechent, P., & Mast, F. W. (2007). The neural basis of the egocentric and allocentric spatial frame of reference. *Brain research*, *1137*, 92–103.