Computational Modeling of the Interactions Between Episodic Memory and Cognitive Control

Hugo Chateau-Laurent

Computational Modeling of the Interactions Between Episodic Memory and Cognitive Control

Hugo Chateau-Laurent

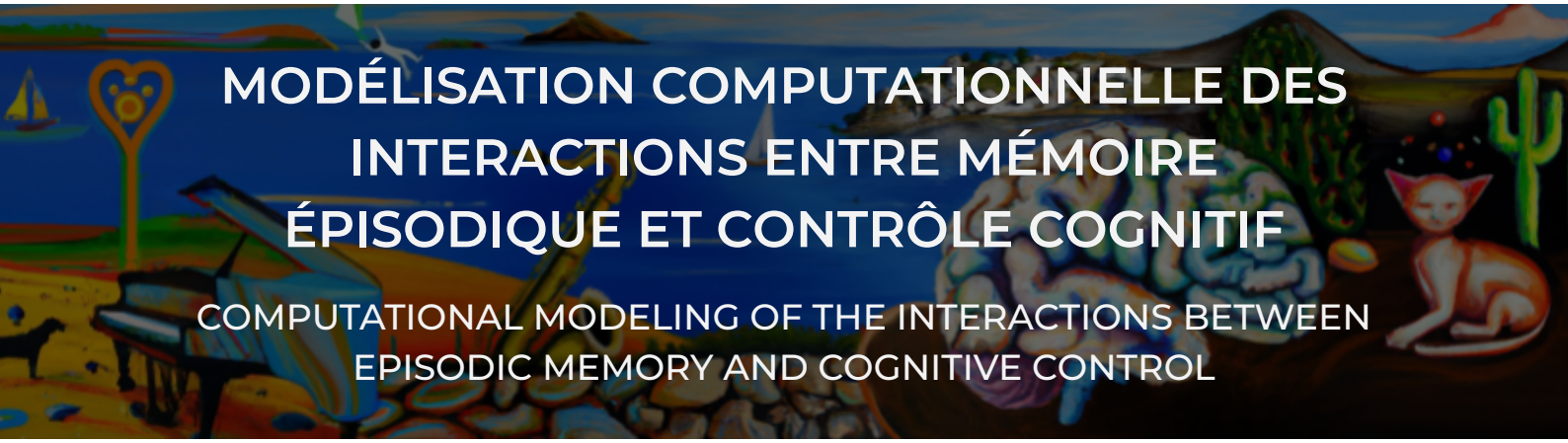
THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE
L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE
SPÉCIALITÉ INFORMATIQUE

Par Hugo CHATEAU-LAURENT



MODÉLISATION COMPUTATIONNELLE DES INTERACTIONS ENTRE MÉMOIRE ÉPISODIQUE ET CONTRÔLE COGNITIF

COMPUTATIONAL MODELING OF THE INTERACTIONS BETWEEN EPISODIC MEMORY AND COGNITIVE CONTROL

Sous la direction de : Frédéric ALEXANDRE

Soutenue le 8 février 2024

Membres du jury :

M. Frédéric ALEXANDRE	Directeur de recherche	Université de Bordeaux	Directeur de thèse
M. Mehdi KHAMASSI	Directeur de recherche	Sorbonne Université	Rapporteur
M. Randall O'REILLY	Professeur	UC Davis	Rapporteur
Mme. Anna SCHAPIRO	Professeure adjointe	Université de Pennsylvanie	Examinatrice
M. Rufin VANRULLEN	Directeur de recherche	CNRS	Examinateur
Mme. Emmanuelle ABISSET-CHAVANNE	Professeure	Arts et Métiers ParisTech	Examinatrice et Présidente du Jury

Acknowledgements

I am profoundly grateful to my supervisor, Frédéric Alexandre, whose unwavering support and guidance have been crucial in shaping my PhD journey. Finding a mentor whose philosophy aligns so closely with mine has been a unique and invaluable experience.

I am also grateful to Dolton Fernandes for conducting preliminary experiments on episodic control and introducing me to the universal Hopfield networks paper. I want to express appreciation to Thomas Vitry for his ongoing efforts in fine-tuning the model for the emergence of prettier place cells. I extend my sincere appreciation to Dmitry Krotov for his crucial assistance in understanding modern Hopfield networks. Despite our limited interactions on Twitter, his expertise has significantly enriched my research, and I am grateful for his valuable insights. I want to express special gratitude to Simon Lecomte, who, in moments when he believed he was not contributing, played a significant role in helping me understand the intricacies of the hippocampus. I would like to express my sincere gratitude to all my other colleagues and friends at the Mnemosyne team and Inria for the wonderful working environment.

I want to extend my appreciation to my family and friends for their unconditional love and support. Their belief in me has been a driving force behind my perseverance. Special thanks to Fandi, Sigrid and Thomas for their patience, encouragement, and unwavering belief in me. Their support has been a constant source of strength. I would also like to acknowledge Xavier Hinaut for his guidance and perpetual flow of project ideas, akin to my own, which proved both frustrating when convergence was essential at the end and exhilarating during other times.

Lastly, my sincere apologies to the thousands of proto-brains I brought into existence and promptly terminated in my simulations. May they rest in digital peace! These experiments were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see `https://www.plafrim.fr`).

Computational Modeling of the Interactions Between Episodic Memory and Cognitive Control

Abstract: Episodic memory is often illustrated with the madeleine de Proust excerpt as the ability to re-experience a situation from the past following the perception of a stimulus. This simplistic scenario should not lead into thinking that memory works in isolation from other cognitive functions. On the contrary, memory operations treat highly processed information and are themselves modulated by executive functions in order to inform decision-making. This complex interplay can give rise to higher-level functions such as the ability to imagine potential future sequences of events by combining contextually relevant memories. How the brain implements this construction system is still largely a mystery. The objective of this thesis is to employ cognitive computational modeling methods to better understand the interactions between episodic memory, which is supported by the hippocampus, and cognitive control, which mainly involves the prefrontal cortex. It provides elements as to how episodic memory can help an agent to act. It is shown that neural episodic control, a fast and powerful method for reinforcement learning, is in fact mathematically close to the traditional Hopfield network, a model of associative memory that has greatly influenced the understanding of the hippocampus. Neural episodic control indeed fits within the universal Hopfield network framework, and it is demonstrated that it can be used to store and recall information, and that other kinds of Hopfield networks can be used for reinforcement learning. The question of how executive functions can control episodic memory operations is also tackled. A hippocampus-inspired network is constructed with as little assumption as possible and modulated with contextual information. The evaluation of performance according to the level at which contextual information is sent provides design principles for controlled episodic memory. Finally, a new biologically inspired model of one-shot sequence learning in the hippocampus is proposed. The model performs very well on multiple datasets while reproducing biological observations. It ascribes a new role to the recurrent collaterals of area CA3, that is to disambiguate overlapping sequences by making retrospective splitter cells emerge. Implications for theories of the hippocampus are discussed and novel experimental predictions are derived.

Research unit
Inria centre at the University of Bordeaux, 200 Avenue de la Vielle Tour,
33405 Talence, France.

Modélisation Computationnelle des Interactions Entre Mémoire Épisodique et Contrôle Cognitif

Résumé : La mémoire épisodique est souvent illustrée par la madeleine de Proust comme la capacité à revivre une situation du passé suite à la perception d'un stimulus. Ce scénario simpliste ne doit pas mener à penser que la mémoire opère en isolation des autres fonctions cognitives. Au contraire, la mémoire traite des informations hautement transformées et est elle-même modulée par les fonctions exécutives pour informer la prise de décision. Ces interactions complexes donnent lieu à des fonctions cognitives supérieures comme la capacité à imaginer de futures séquences d'événements potentielles en combinant des souvenirs pertinents dans le contexte. Comment le cerveau implémente ce système de construction reste un mystère. L'objectif de cette thèse est donc d'employer des méthodes de modélisation cognitive afin de mieux comprendre les interactions entre mémoire épisodique reposant principalement sur l'hippocampe et contrôle cognitif impliquant majoritairement le cortex préfrontal. Elle propose d'abord des éléments de réponse quant au rôle de la mémoire épisodique dans la sélection de l'action. Il est montré que le contrôle épisodique neuronal, une méthode puissante et rapide d'apprentissage par renforcement, est en fait mathématiquement proche du traditionnel réseau de Hopfield, un modèle de mémoire associative ayant grandement influencé la compréhension de l'hippocampe. Le contrôle épisodique neuronal peut en effet s'inscrire dans le cadre du réseau de Hopfield universel, il est donc montré qu'il peut être utilisé pour stocker et rappeler de l'information et que d'autres types de réseaux de Hopfield peuvent être utilisés pour l'apprentissage par renforcement. La question de comment les fonctions exécutives contrôlent la mémoire épisodique est aussi posée. Un réseau inspiré de l'hippocampe est créé avec le moins d'hypothèses possible et modulé avec de l'information contextuelle. L'évaluation des performances selon le niveau auquel le contexte est envoyé propose des principes de conception de mémoire épisodique contrôlée. Enfin, un nouveau modèle bio-inspiré de l'apprentissage en un coup de séquences dans l'hippocampe est proposé. Le modèle fonctionne bien avec plusieurs jeux de données tout en reproduisant des observations biologiques. Il attribue un nouveau rôle aux connexions récurrentes de la région CA3, qui est celui de distinguer les séquences se chevauchant en faisant émerger des cellules de séparation rétrospective. Les implications pour les théories de l'hippocampe sont discutées et de nouvelles prédictions expérimentales sont dérivées.

Unité de recherche
Centre Inria de l'Université de Bordeaux, 200 Avenue de la Vielle Tour,
33405 Talence, France.

Contents

1 Introduction Générale

> *Et tout d'un coup le souvenir m'est apparu. Ce goût c'était celui du petit morceau de madeleine que le dimanche matin à Combray (parce que ce jour-là je ne sortais pas avant l'heure de la messe), quand j'allais lui dire bonjour dans sa chambre, ma tante Léonie m'offrait après l'avoir trempé dans son infusion de thé ou de tilleul.*
>
> *Marcel Proust, Du côté de chez Swann (1913)*

Comment la simple saveur d'une madeleine peut-elle nous transporter à travers le temps et l'espace, nous permettant de revivre un moment spécifique de notre passé ? Qu'est-ce qui fait que cet événement particulier se distingue de la myriade d'autres ? En quoi ce souvenir est-il censé guider notre comportement ? Ce sont les questions auxquelles cette thèse vise à répondre.

La mémoire est une composante essentielle de notre identité. Nos expériences façonnent nos représentations et guident nos actions. Se souvenir d'une situation précise fait appel à une forme de mémoire spécifique : la mémoire épisodique. C'est la forme à laquelle nous pensons spontanément car elle fait partie de la mémoire explicite, en ce sens qu'elle a un accès privilégié à la conscience.

Un élément crucial de la mémoire épisodique est sa capacité à apprendre très rapidement des informations qui n'ont parfois été perçues qu'une fois. Pensez au jour où vous avez eu votre baccalauréat. Cela ne s'est produit qu'une fois et pourtant le souvenir peut encore être présent. C'est un défi pour les modèles d'apprentissage automatique qui s'appuient souvent sur des exemples répétés pour ajuster méticuleusement les paramètres.

Avec ces méthodes, un élément vu une fois a très peu d'impact sur les paramètres et sera presque certainement oublié. Lorsque les modèles sont ajustés plus rapidement, les nouvelles informations écrasent les souvenirs existants, un problème appelé oubli catastrophique. Cela se produit principalement parce que les modèles d'apprentissage automatique sont conçus pour généraliser leurs connaissances afin de classer, estimer et agir dans de nouvelles situations. Ce n'est pas le cas de la mémoire épisodique qui est distincte d'autres fonctions telles que la mémoire sémantique. La généralisation est le rôle de la mémoire sémantique tandis que la mémoire épisodique peut intégrer rapidement de nouvelles informations en les stockant séparément pour éviter l'interférence avec les anciennes informations (McClelland et al. 1995).

Un autre aspect important de la mémoire épisodique est sa capacité à retenir l'ordre des événements, ce qui est difficile dans le cas de séquences qui se chevauchent. Considérons le scénario suivant. Lorsque vous arrivez sur votre lieu de travail, vous prenez l'ascenseur pour accéder à votre bureau et, à l'heure du déjeuner, vous prenez le même ascenseur pour vous rendre à la cafétéria. Il est utile de se souvenir de l'itinéraire menant à la cafétéria lorsque l'on prend l'ascenseur depuis son bureau. En revanche, lorsque vous arrivez au travail depuis l'extérieur, il est moins utile de s'en souvenir. L'étude de la navigation spatiale est liée à l'étude de la mémoire épisodique, car les représentations spatiales sous-tendent l'encodage des épisodes dans la région de l'hippocampe du cerveau. Certains neurones de cette région, appelés cellules de lieu, codent explicitement des variables spatiales. Dans notre exemple, l'ascenseur est un passage commun vers différentes destinations, ce qui pose problème pour distinguer les souvenirs pertinents. Certaines cellules de lieu représentent en fait les positions spatiales différemment selon l'endroit d'où nous venons (cellules séparatrices rétrospectives) ou l'endroit où nous avons l'intention d'aller (cellules séparatrices prospectives), comme l'explique Duvelle et al. (2023).

La mémoire épisodique nous joue parfois des tours. Nous pouvons nous rappeler malgré nous d'un souvenir intrusif désagréable, oublier des informations utiles ou nous souvenir d'une version déformée de la réal-

ité. Ces effets sont exacerbés dans des pathologies telles que la maladie d'Alzheimer, le stress post-traumatique ou les troubles obsessionnels compulsifs. Le cerveau sain évite cela grâce aux interactions entre la mémoire épisodique et des systèmes qui vérifient la véracité des informations rappelées et contrôlent le moment et le contenu de la récupération en mémoire. Ces systèmes sous-tendent les fonctions exécutives, en ce sens qu'ils contrôlent d'autres fonctions cognitives pour exécuter des plans comportementaux. Le cortex préfrontal en est le principal acteur. Il est soupçonné de provoquer la division prospective des cellules de lieu mentionnée ci-dessus afin d'orienter la recherche de mémoire vers des objectifs. En retour, il utilise les informations envoyées par la mémoire épisodique pour guider le comportement. C'est ainsi, par exemple, que nous pouvons choisir d'acheter un disque d'un artiste dont nous avons apprécié la musique par le passé. Des fonctions encore plus complexes naissent de l'interaction entre la mémoire et les fonctions exécutives. C'est le cas, par exemple, de l'imagination de situations futures, qui semble faire intervenir à la fois les structures impliquées dans la mémoire du passé et les régions dédiées au contrôle cognitif (Schacter, Addis, and Buckner 2007). Cela suggère que l'imagination anticipe le futur en recombinant des souvenirs.

Les progrès récents de l'intelligence artificielle générative ont eu un impact sans précédent sur la société, avec l'utilisation emblématique des transformateurs dans ChatGPT (Vaswani et al. 2017). Comparées à la vitesse d'apprentissage de notre mémoire épisodique, ces méthodes nécessitent des quantités astronomiques de données. Il est intéressant de noter que certains aspects des transformateurs sont similaires à des éléments de modèles de la mémoire épisodique (Ramsauer et al. 2020; Whittington, Warren, et al. 2021). Ce parallèle mérite d'être approfondi pour comprendre leurs besoins divergents en matière de données. En outre, les transformateurs manquent de contrôle cognitif, ce qui les rend susceptibles de produire des informations inexactes.

L'objectif principal de cette thèse est de caractériser les interactions entre la mémoire épisodique et le contrôle cognitif avant de poursuivre le parallèle avec les transformateurs pour développer des modèles computation-

nels biologiquement plausibles de ces interactions. L'objectif est double : d'une part, participer à la compréhension de ces fonctions dont l'altération sous-tend des pathologies sévères, et d'autre part, le développement de meilleurs modèles d'apprentissage automatique. La thèse est structurée comme suit. Elle commence par une revue détaillée de la littérature, fournissant une vue d'ensemble de l'état de l'art sur le sujet. Le chapitre 3 présente la mémoire épisodique et le contrôle cognitif en relation avec d'autres fonctions cognitives. Le chapitre 4 décrit l'anatomie, les représentations, la dynamique et les modèles informatiques de la mémoire épisodique, en mettant l'accent sur l'apprentissage d'une séquence vécue une seule fois. Enfin, le chapitre 5 présente les aspects anatomiques, représentationnels, dynamiques et informatiques des interactions entre le préfrontal et l'hippocampe, qui permettent un comportement guidé par la mémoire et une mémoire orientée vers un but. L'analyse est suivie de trois chapitres de contribution. Dans le chapitre 6, une équivalence mathématique est trouvée entre les modèles de mémoire épisodique et un module d'un algorithme d'apprentissage profond pour la sélection des actions, renforçant le lien entre cet algorithme et les interactions préfrontal-hippocampique. Il est ensuite démontré que des éléments de l'algorithme améliorent les performances des modèles de mémoire épisodique de pointe dans certaines tâches, et que des éléments des modèles de mémoire épisodique pourraient à leur tour améliorer l'algorithme d'apprentissage profond. Le chapitre 7 explore la modulation contextuelle des modèles de mémoire épisodique. Les réseaux neuronaux émergents comprennent des neurones modulés par le contexte, à l'instar des cellules séparatrices. Les résultats remettent également en question le cadre standard de la mémoire épisodique dans l'hippocampe. À la suite de ces études exploratoires, un nouveau modèle d'apprentissage de séquences en une seule fois dans l'hippocampe est présenté au chapitre 8. Ce modèle surpasse les autres modèles testés et reproduit divers résultats biologiques. Enfin, les contributions sont résumées et discutées au chapitre 9.

2 General Introduction

> *And suddenly the memory returns. The taste was that of the little crumb of madeleine which on Sunday mornings at Combray (because on those mornings I did not go out before church time), when I went to say good day to her in her bedroom, my aunt Leonie used to give me, dipping it first in her own cup of real or of lime-flower tea.*
>
> *Marcel Proust, Swann's Way (1913)*

How can the simple flavor of a madeleine transport us through time and space, allowing us to vividly revisit a specific moment from our past? What makes this particular event stand out amidst the myriad of others? In what way is this memory intended to guide behavior? These are the questions this thesis aims to address.

Memory is a key component of our identity. Experiences shape our representations and guide our actions. Remembering a specific experience calls on a very specific form of memory: episodic memory. This is the memory we think about spontaneously as it is part of explicit memory, in that it has privileged access to consciousness.

A crucial element of episodic memory is its capacity to learn information very quickly, having sometimes only been perceived once. Think back to when you graduated from high school. It only happened once and yet the memory can still be vivid. This is a challenge for machine learning models, which often rely on millions of repeated examples to meticulously adjust parameters. With these methods, an element seen once has very little impact on the parameters and will almost certainly be forgotten. When

models are adjusted more strongly, new information overwrites existing memories, a problem called catastrophic forgetting. This happens mainly because machine learning models are designed to generalize their knowledge to classify, estimate and act in new situations. This is not the case with episodic memory, which is distinct from other functions such as semantic memory. Generalization is the role of the complementary semantic memory, while episodic memory can rapidly integrate new information, storing it separately to avoid interference with old information (McClelland et al. 1995).

Another important aspect of episodic memory is its ability to retain the order of events, which is challenging in the case of overlapping sequences of information. Consider the following scenario. Upon reaching your workplace, you use the elevator to access your office, and during lunchtime, you take the same elevator to reach the cafeteria. It is beneficial to remember the route to the cafeteria when entering the elevator from your office. However, when arriving at work from outside, there is no need to remember this. The study of spatial navigation is closely linked to the study of episodic memory, as spatial representations underpin the encoding of episodes in the hippocampus of the brain. Some neurons in this region, called place cells, explicitly encode spatial variables. In the example scenario, the elevator serves as a common pathway to different destinations, posing a challenge in distinguishing relevant memories. Certain place cells in fact represent spatial positions differently depending on where we come from (retrospective splitter cells) or where we intend to go (prospective splitter cells), as reviewed by Duvelle et al. (2023).

Sometimes episodic memory plays tricks on us. We may recall an unpleasant intrusive memory in spite of ourselves, forget useful information, or remember a distorted version of reality. These effects are exacerbated in conditions such as Alzheimer's disease, post-traumatic stress disorder or obsessive-compulsive disorder. The healthy brain prevents this through interactions between episodic memory and systems that verify the veracity of recalled information and control the timing and content of memory retrieval. These systems underlie executive functions, in that they con-

trol other cognitive functions to execute behavioral plans. The prefrontal cortex is the principal cognitive control region, and is suspected to cause the prospective splitting of place cells mentioned above to direct memory search according to goals. In return, it uses the information sent by episodic memory to guide behavior. This is how, for example, we may choose to buy a record by an artist whose music we have enjoyed in the past. Even more complex functions arise from the interaction of memory and executive functions. This is the case, for example, of imagining future situations, which seems to involve both the structures involved in remembering the past and regions dedicated to cognitive control (Schacter, Addis, and Buckner 2007). This suggests that imagination is a future-oriented construction process based on the recombination of past memories.

Recent advances in generative artificial intelligence have had an unprecedented impact on society, with the emblematic use of transformers in ChatGPT (Vaswani et al. 2017). Compared to the learning speed of our episodic memory, these methods require astronomical amounts of data before they can function properly. Interestingly, some aspects of transformers are similar to elements of episodic memory models (Ramsauer et al. 2020; Whittington, Warren, et al. 2021). This parallel deserves to be explored in greater depth to understand their divergent needs for data. In addition, transformers lack cognitive control, which makes them susceptible to hallucinations, that is to produce inaccurate information.

The main goal of this thesis is to characterize the interactions between episodic memory and cognitive control before extending on the parallel with transformers to develop efficient biologically plausible computational models of these interactions. The objective are twofold: first, to participate in the understanding of these functions whose impairment underlies severe pathologies, and second, the development of better machine learning models. The thesis is structured as follows. It begins with a comprehensive review of the literature, providing an overview of the state of the art on the subject. Chapter 3 presents episodic memory and cognitive control in relation to other cognitive functions. Chapter 4 describes the anatomy, representations, dynamics and computational models of episodic memory,

with a particular focus on one-shot sequence learning. Finally, Chapter 5 introduces the anatomical, representational, dynamical and computational aspects of prefrontal-hippocampal interactions enabling memory-driven behavior and goal-directed memory. The review is followed by three contribution chapters. In Chapter 6, a mathematical equivalence is found between episodic memory models and a module of a deep learning algorithm for action selection, strengthening the link between this algorithm and prefrontal-hippocampal interactions. It is then shown that elements of the algorithm improve the performance of state of the art episodic memory models in some tasks, and that elements of episodic memory models could in turn improve the deep learning algorithm. Chapter 7 explores the contextual modulation of episodic memory models. The emergent neural networks comprise neurons modulated by context similar to splitter cells. The results also question the standard framework of episodic memory in the hippocampus. Following these exploratory studies, a novel model of one-shot sequence learning in the hippocampus is introduced in Chapter 8. It outperforms other models tested and reproduces various biological findings. Finally, the contributions are summarized and discussed in Chapter 9.

Part I

Memory & Control

3 Cognition & Memory

An agent with cognition and memory

Generative AI

Following fundamental laws of physics and the contingency of events, time naturally leaves a trace. Humans grow out of embryos, become adults, live their life and eventually die, but these events cannot happen in a different order (except perhaps for Benjamin Button). Humans and other animals can leverage information left by a chain of events to make decisions. They can for example hunt by following a track to a prey, as prints reveal its path. Time can also be detrimental. No matter how helpful they can be,

traces of time are sometimes not sufficient to inform the agent as to what to do. Information can be lost when subsequent events further modify the environment. On a more radical note, agents must "fight" against the effect of time for self-preservation. For these reasons, organisms have developed strategies for actively changing their environment to their own benefit. Organisms maintain homeostasis, they build shelters and most importantly to our concern they form memories.

In cognitive science, memory has indeed become an umbrella term referring to the general ability to learn: encode, store and retrieve information (Squire 2009). Forming a memory is the action of modifying a substratum (encoding) in such a way that the modification persists in time (storage) with the aim of extracting information from that modification (retrieval) and benefit from it to act in the future. As a result of being general, Zlotnik and Vansintjan (2019) described how this definition extends the traditional boundaries of memory to processes outside the brain and mind. For example, it can be argued that immunological processes such as the triggering of a cascade of reactions as a result of the recognition of a previously identified threat can be interpreted as memory processes. Furthermore, books and hard drives are devices that are typically used to store information outside the body. As a consequence, theories of the extended mind (Francisco et al. 1991; Clark and Chalmers 1998; Clark 2008) have been proposed to account for dynamic interactions between the mind, body and environment.

This thesis is largely confined to memory processes operating within the brain (while being compatible with theories of extended cognition) but this introduction is meant to emphasize how general the notion of memory has become. Cognitive scientists have thus tried to refine it over the last decades by proposing criteria for classifying memory systems and other cognitive functions. Below are detailed some of these criteria in order to better lay down the specifics of episodic memory and cognitive control, which underlie higher-level cognition and are the main focus of this thesis.

3.1 Associative and Nonassociative Memory

A common criterion that is used to classify memory forms is whether the information being stored is an association between elements. Most memories formed are associative, like when you learn that *2+2* is associated with the result *4*, when you learn the association between food and pleasure (or disgust) or when you know which song comes next while listening to an album, to cite a few examples. Canonical forms of associative learning are classical and operant conditioning in which a first stimulus is paired to a second stimulus and a consequence respectively. All memories, however, are not associative. Habituation and sensitization are common forms of nonassociative learning. Habituation refers to the progressive attenuation of a response to a stimulus after repetitive presentation without necessarily learning the association between stimulus and outcome (e.g. lack of punishment). A well known instance of habituation is the desensitization of hunting dogs to loud sounds. While most dogs innately blench when hearing loud stimuli, hunting dogs learn to ignore auditive stimuli like rifle shots. In the case of sentization, the response strength increases. It can happen when a stimulus is presented several times above some tolerance threshold (which determines whether habituation of sensitization is induced). Sensitization can be characteristic of pathologies such as post-traumatic stress disorder and phobias.

3.1.1 Autoassociation and Heteroassociation

A particular case of associative memory is the memorization of individual elements which might not seem to involve association at first glance. Memorization can indeed be framed in associative terms as learning the interconnections between parts and aspects of the whole. The remembrance of an item thus typically occurs following the perception of an altered version of that item, whether being incomplete (Figure 3.1b) or noisy (Figure 3.1c), as remaining bits of information are used as cues for recalling the original memory. This particular case where the cue enabling recall is part of the

same whole is called auto-association. It contrasts with hetero-association, the cue of which belongs to a different item.

It is worth noting that in some cases, an autoassociation might be recalled from a cue that does not really belong to the same item but is not considered heteroassociative if the cue was not present during encoding. In the case of pareidolia[1], the recalled association can be considered spurious (Figure 3.1e). It is even harming in pathologies with intrusive memory recalls. However, such remote associations can be helpful for generalising attributes of an instance to a broader class (Figure 3.1d). Likewise, this process can support creative behavior such as the naming of brain structures (Figure 3.1f).

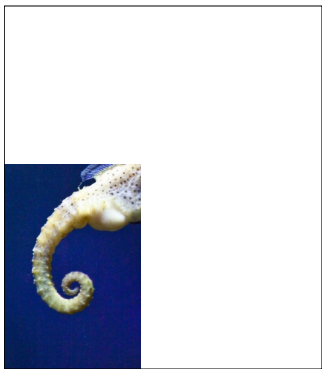
This way of storing information about individual and multiple things through associations radically contrasts with the way typically used in a computer, where a memory is located at an address, the content of which can be addressed with a pointer to that address. In the computer case, memories are thus stored locally and very susceptible to data loss. In the case of associative memory the content is distributed and can be addressed with cues of many forms which makes the system much more resilient.

3.1.2 Associative Space and Forms

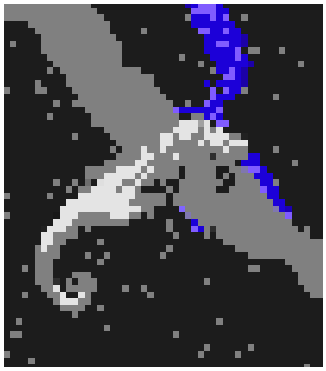
Elements of memory can be associated in many dimensions. The space in which associations are formed includes but is not limited to spatial (e.g. *X is on top of Y*), semantic (e.g. *X is a duck*), emotional (*X triggers emotion Y*) and cognitive dimensions (e.g. *X helps me reach goal Y*). Coming back to a previous example, when learning the contiguity of songs on an album, an extra dimension is involved: time. Memories concern which songs follow one another. Note how associations can thus be multidimensional, as songs can be described by semantic, emotional and cognitive dimensions, and that the order of songs constitutes an additional temporal component. The cognitive function enabling the learning of event order is called se-

[1] Pareidolia is an optical illusion involving the perception of an object (e.g. a face) in a totally unrelated object (e.g. a cloud)

(a) A seahorse.

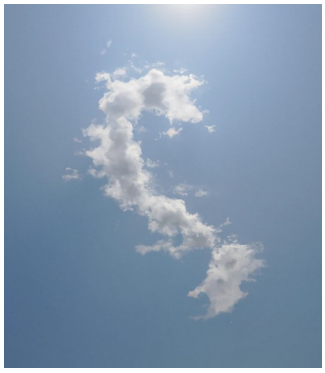


(b) Part of Figure 3.1a.



(c) Corrupted version of Figure 3.1a.

(d) A different kind of seahorse.



(e) Clouds in the shape of a seahorse (AI generated).

(f) Preparation of a human hippocampus and fornix. Adapted from Seress (2010).

Figure 3.1: Various visual representations related to the concept of seahorse.

quence memory, which is heteroassociative in essence because it concerns the temporal relationship between multiple events. An interesting characteristic of music is that time is also inherent to individual songs. Music is a temporal combination of many sounds and sounds themselves are a variation of air pressure over time. When practicing, a musician learns to play each note for a specific duration. Here, it is not only the order of events that is learned but also the timing of individual elements. A related example is the acquisition of songs by birds. HVC (used as a proper name) is the avian brain structure responsible for sequencing songs, the output of which is associated with downstream muscle activity for song production (Hahnloser et al. 2003; Long and Fee 2008; Sankar 2022). When this region is cooled down, the produced song is slowed down, revealing the causal role of HVC in motor timing (Long and Fee 2008).

In mammals, the hippocampus is a crucial brain region for learning the temporal unfolding of events. So-called time cells can be found whose activity is modulated by elapsed time (Itskov et al. 2011; MacDonald et al. 2011). Recently, it has been proposed that the hippocampus generates sequences of content-free pointers that get associated with sequences of cortical representations for learning event chronology (Buzsáki and Tingley 2018). After encoding this associative mapping, replaying sequences in the hippocampus would replay activity in the cortex corresponding to experienced events. This supposedly underlies the retrieval of memories and their temporal structure.

Howard Eichenbaum (2017a) proposed that sequential memory is actually distinct from the memory of associations. In his review, he describes three forms of memory: associative, sequential and schematic memory. While "associations" are necessarily empty of semantic meaning in his view, sequential and schematic memories involve meaningful relations. Sequential memory specifically relates to the temporal ordering of events, whereas what he calls schematic memory stores relations of hierarchy (e.g. A is greater than B). Temporal dynamics of the hippocampus play a major role for representing, storing and recalling sequences (see Section 4.4).

3.2 Implicit and Explicit Processes

In everyday language, the term memory typically refers to declarative knowledge related to events, concepts or facts. The type of memory concerned with content amenable to such conscious recall and verbal report is referred to as explicit (or declarative) memory. A lot of what is learned however cannot be accessed and reported so easily. In cognitive science, the notion of memory also encompasses implicit knowledge. For example, procedural memory is the ability to acquire skills without necessarily being able to communicate or reflect on this knowledge (think about how difficult it is to explain how to ride a bike). Another example is semantic priming, the facilitatory effect in the processing of a word following the presentation of a related word. This process occurs automatically without one's awareness.

Historically, the distinction between implicit and explicit memory traces its origins back to the famous case of patient H.M. who suffered from epileptic seizures. A large portion of his left medial temporal lobe was thus surgically removed. As a consequence of that operation, H.M. became severely amnesic, unable to recall events he had experienced in the few years preceding the surgery, as described in Scoville and Milner (1957) (see also Milner et al. 1968). The patient's ability to form lasting explicit memories was also severly impaired. Scoville and Milner (1957) noted that "once he had turned to a new task the nature of the preceding one could no longer be recalled, nor the test recognized if repeated" (p. 108). It was rapidly discovered however that patient H.M. could learn novel information of a certain nature. In the so-called mirror tracing task, H.M. successfully learned to draw an object using the image reflected from a mirror as sole feedback (Milner 1962). His performance improved over time, demonstrating preserved motor skill learning. In a more recent study, it was reported that the acquired skills could be assessed as many as 359 days after exposure to the task (Gabrieli et al. 1993). Follow-up investigations revealed that H.M. and other amnesic patients can acquire other implicit learning tasks, as reviewed by Hannula and Greene (2012).

The case of patient H.M. reveals a marked distinction between acquired information amenable to conscious access and verbal report (i.e. explicit), the learning of which involves the medial temporal lobe (where the hippocampus is located), and information learned that cannot be accessed consciously and reported verbally (i.e. implicit). This thesis is focusing on episodic memory and cognitive control which are traditionally considered explicit cognitive functions. Episodic memory indeed concerns the storage and explicit retrieval of experiences. Similarly, cognitive control was originally assumed to always employ explicit representations for modulating action selection contextually. However, the classification of cognitive processes on the basis of whether they happen with or without consciousness has been questioned, especially in the case of episodic memory (Henke 2010; Hannula and Greene 2012) and cognitive control (Lau and Passingham 2007; Gaal et al. 2012). In the following sections, these notions will be introduced and their relationship to consciousness will be discussed.

3.3 Semantic and Episodic Memory

Tulving (1972) further divided explicit long-term[2] memory into episodic and semantic subsystems. Episodic memory is characterized by its fast acquisition and the nature of information it stores. So-called episodes typically comprise multimodal sensory information along with emotional content. Each episode corresponds to a specific experience and refers to the individual experiencing it (i.e. autobiographical). Another key characteristic of episodic memory is the storage of how events unfold through time. It is a temporal associative memory (see Section 3.1).

On the other hand, what is stored in semantic memory is general facts and concepts about the world, such as *Paris is the capital of France* or *1+1=2*. Characteristically, these pieces of knowledge are invariantly true regardless of the sensory, emotional and temporal context. They are indeed not stored in association with their context of acquisition, as the context may

[2]See Section 3.4 for a discussion on the notion of long-term memory.

not be invariantly associated with the semantic information. For example, while the association between *Paris* and *capital of France* is likely to hold in most situations, the association between *Paris is the capital of France* and *My geography teacher is wearing a blue shirt* is far less likely to hold in the future. The web of associative, schematic and sequential relations (see Section 3.1.2) forms so-called schemas, which constitute the scaffolding of semantic memory.

3.3.1 Complementary Learning Systems Theory

Why are there multiple systems instead of one playing the roles of both episodic and semantic memory? This is the question addressed by McClelland et al. (1995) and the complementary learning systems theory. The main idea is that the hippocampus and the cortex support episodic and semantic memory respectively, and that this division of labor avoids the problem of catastrophic forgetting introduced by McCloskey and N. J. Cohen (1989). They illustrate this problem in the context of the AB-AC paradigm (J. M. Barnes and Underwood 1959) in which a first word is presented (stimulus) and the goal is to produce the paired word (response). The correct answer is given at the end of each trial. Once a first set of associations (e.g. *locomotive-dishtowel*, *table-street*, and *carpet-idea*) has been learned after multiple trials of each pair, a second list involving the same stimuli but different responses is presented (e.g. *locomotive-banana*, *table-basket*, and *carpet-pencil*). Given the stimulus and the context of which list to use, participants must provide the correct association. Rumelhart (1990) trained a connectionist network to perform the task by providing the stimulus and context (first or second list) as input and assessing the output. They showed that the ability of the model to recall the first set of associations drops catastrophically as soon as the second set is presented, whereas this interference effect is much weaker in humans. Human subjects are indeed able to acquire the new set of associations and the drop of performance for the first set is much slower and less important than in the model. Catastrophic interference can be attenuated by interleaving trials of the first set and trials involving the

new associations. When such a scheduling is used, the network acquires the new associations at a slower pace (since old information is presented along) but the interference effect is greatly mitigated as old associations are reminded continually (McClelland et al. 1995).

While the interleaved training regime appears to solve the problem of catastrophic interference, an agent cannot always control the order in which information to learn is presented in its environment. The key proposal of the complementary learning systems is that the hippocampus teaches the cortex by replaying old information interleaved with new information to acquire. The hippocampus thus plays the role of an internally simulated environment with a more optimal training regime than the actual environment for learning in the cortex. This raises the question of how the hippocampus realises this function and does not suffer itself from catastrophic interference. McClelland et al. (1995) proposed that the hippocampus stores experiences separately from each other and that the storage of one does consequently not interfere with the retention of another. In order to do so, the hippocampus uses sparse coding (C. A. Barnes et al. 1990; Quirk et al. 1990) such that representations of slightly different events have little overlap (David Marr and Thach 1991; Bruce L McNaughton and Morris 1987; O'Reilly and McClelland 1994). This way of representing and storing information is well suited for fast acquisition without interference, yet comes at the cost of not being able to discover common structure among several examples and generalize across situations. Interestingly, the nature of episodic memory supposedly results from this solution to the catastrophic interference problem: as experiences are stored separately, the multisensory context must not be discarded. In comparison, the cortex employs much denser representations and slowly aggregates information, which makes it much better at finding patterns in data and generalizing over multiple situations (McClelland et al. 1995). In the process of generalization, the context of acquisition is not imprinted as only invariants are kept. Finally, the cortex is much more modular and each cortical region is much more specialised as compared to the hippocampus which virtually integrates information from

the whole cortex. These elements explain the contrast between semantic and episodic memory.

It is worth to note that novel information can be acquired rapidly in the neocortex when it is consistent with an existing schema (Tse et al. 2011; McClelland 2013). The complementary learning systems theory has thus been refined to account for these findings Kumaran, Hassabis, et al. (2016).

3.3.2 Implicit Episodic Memory

In Section 3.2, the distinction between explicit and implicit memory systems was introduced with the notable case of patient H.M. While he retained the ability to acquire skills, his episodic memory was severely impaired. Episodic memory is typically considered an explicit memory process, as one can be conscious of and verbally report one's episodic memories. However, there have been reports of hippocampus learning operations, which are typically associated with episodic memory, that happen without conscious awareness (Hannula and Greene 2012; Henke 2010). In Hannula and Ranganath (2009), hippocampal activity predicted whether participants successfully made eye movements to the face that was previously associated with the current scene. This was also the case in trials in which participants could not consciously remember which one was the correct face. By contrast, amnesic patients with hippocampal damage did not exhibit the eye movement effect Hannula, Ryan, et al. (2007). This suggests that hippocampus-dependent recall of associations can occur without affecting conscious experience. Similarly, subliminal information can be encoded in the hippocampus and influence future conscious retrieval (Henke, Mondadori, et al. 2003; Henke, Treyer, et al. 2003). In these studies, pairs of faces and occupations (e.g. plumber) were presented very briefly (17 ms) with visual noise preceding and following the stimulus, making it inaccessible to conscious perception. During both encoding (subliminal presentation) and retrieval (supraliminal presentation), hippocampal activity changed in the experimental versus control condition. In a follow-up study (Degonda et al. 2005), face-occupation pairs were presented supraliminally after present-

ing incongruent pairs subliminally (same faces were paired with different occupations). The hippocampus was also found to be modulated by the presentation of stimuli independently of conscious awareness. Most interestingly, implicit pairs impaired the retrieval of incongruent explicit pairs, suggesting an influence of implicit memory on explicit memory.

These findings challenge the consciousness-based division of memory systems. An alternative classification was proposed based instead on processing modes (Henke 2010), as illustrated in Figure 3.2. Episodic memory enables rapid encoding of flexible associations (or bindings) such as face-occupation pairs, and it relies on the hippocampus and parts of the neocortex. Alternatively, familiarity and priming involve rapid encoding of single or unitized items and do not require the hippocampus but rather surrounding medial temporal lobe structures. Slow encoding of rigid associations supporting procedural memory, classical conditioning and semantic memory is totally independent from the medial temporal lobe (at least after consolidation ; see Sections 3.3.1 and 3.4).

3.4 Short-Term and Long-Term Memory

Another common way to classify memory systems is to distinguish them on the basis of their temporal capacity. While information stored in working memory is susceptible to distractions and is typically not kept for longer than seconds or minutes, the content of semantic, episodic, procedural and other memory systems can last for years or be lifelong. There is no clear division between short-term and long-term memory which is why these terms can be misleading. In fact, there is a continuum of temporal capacities among memory systems, and there can also be differences between how much time individual items are stored in the same system. However, there seems to be a relationship between processing stage and system capacity as suggested by the Multi-Store Model of Atkinson and Shiffrin (1968), in which each system has a smaller capacity than the preceding one, as it selects which information to store, but can store selected information for a

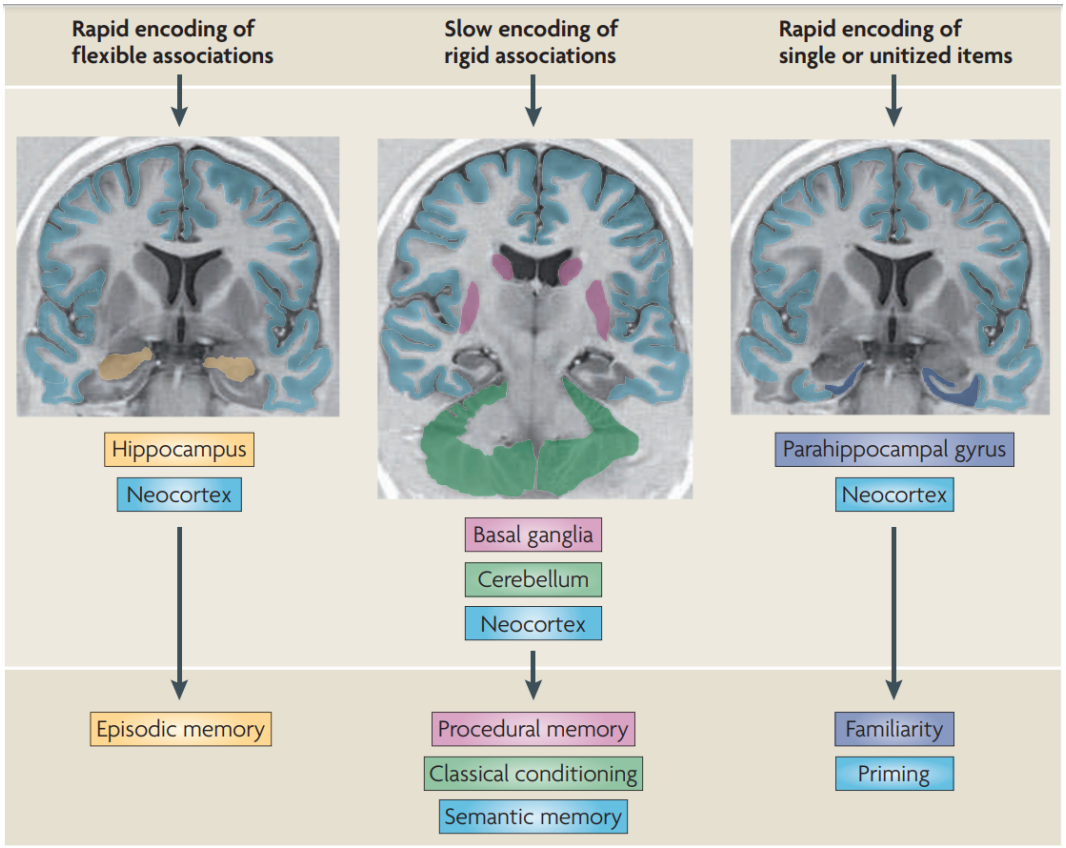


Figure 3.2: Memory systems classification based on processing modes. Taken from Henke (2010).

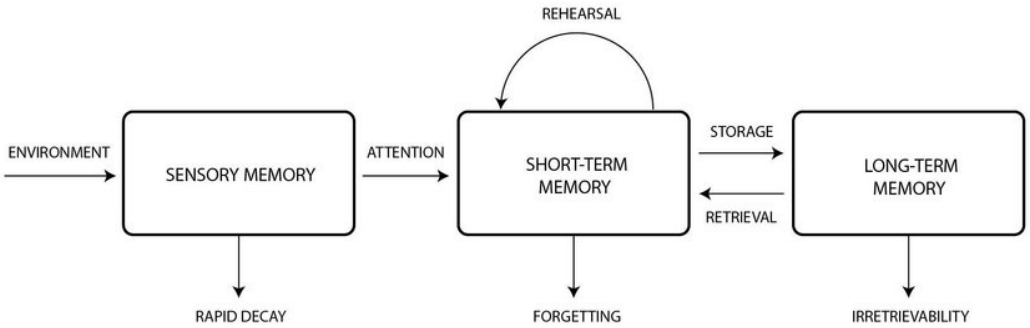


Figure 3.3: Multi-Store Model of Atkinson and Shiffrin (1968). Taken from Reser (2022).

longer time (see Figure 3.3). Sensory memory is at the earliest stage and has biggest capacity but information transiting through it rapidly decays unless it is attended to and stored in working memory.

There is a commonly agreed difference between how short-term (e.g. sensory and working memory) and long-term (e.g. episodic and semantic) memory systems store information. Short-term systems typically rely on the reverberation of neural activation for maintaining information across time. Since activity can be perturbed, this mode of storage is less robust than long-term memory systems that store information by changing the structure of the network (typically the synapses). Stokes (2015) reviewed some data suggesting an alternative synaptic-based and activity-silent way of storing information in working memory. According to Beukers et al. (2021) a more parsimoniously explanation is that episodic memory (synaptic-based) interacts with working memory (activity-based). This illustrates that so-called “long-term" memory systems cannot only play a role in the long term but can also support cognition in the short term.

3.4.1 Standard Consolidation Theory

There is also a disagreement on the time span of hippocampal involvement in memory that further blurs the line between short-term and long-term memory systems. In the hippocampal memory indexing theory of Teyler and DiScenna (1986), the hippocampus flexibly stores novel combinations of

neocortical activations and can recall memories by reinstantiating the activation of the corresponding cortical areas. Eventually, some of these memories are transferred to the cortex through a process called consolidation, while others are forgotten. The standard theory of consolidation (Squire and Alvarez 1995; McClelland et al. 1995), which is similar to the complementary learning systems theory mentioned in Section 3.3.1, builds upon the observation that patients with hippocampal damage can recall remote memories better than recent ones, a phenomenon called temporally-graded retrograde amnesia. The idea is that once memories have initially been stored in the hippocampus, the consolidation process underlies a transfer of memories from the hippocampus to the neocortex for longer term storage, as originally proposed by Marr (1971), and that memories are no longer stored in the hippocampus. Then, when information perceived from the environment is consistent with information consolidated in the cortex (i.e. with schemas), the hippocampus is inhibited by the prefrontal cortex to prevent it from learning redundant information (Frankland and Bontempi 2005; Van Kesteren et al. 2012).

3.4.2 Transformative Accounts of Consolidation

The standard theory of consolidation was later challenged by findings of remote memory recall impairment following hippocampal lesions when memory are truly episodic/autobiographical. The hippocampus thus seems necessary for recalling detailed episodic memories vividly (Viskontas et al. 2002; Rosenbaum et al. 2000; Cipolotti et al. 2001). The multiple trace theory was proposed by Nadel and Moscovitch (1997) as an alternative to the standard consolidation theory in order to account for these findings. As its name implies, the theory states that memories are stored with multiple traces (i.e. in both the hippocampus and neocortex) and that the hippocampus is recruited for recalling episodic memories, regardless of how remote they are. This theory is in agreement with the broader hypothesis that memories are transformed during transfer, such that they take a more semantic form in the neocortex while remaining truly episodic in

the hippocampus. The recently proposed CRISP theory of the hippocampus (Cheng 2013) also departs from the standard view and emphasizes the temporal sequence nature of episodic memory and hippocampal processing. The author states that semantic memory stores information similar to episodic memory, but that the recall of temporal associations necessarily involves the hippocampus. This rightfully highlights the temporal nature of episodic memory, however this claim can be contrasted with the observation that language (and especially syntax) also contains a strong temporal component that is largely preserved following hippocampal impairment (but see H.M.'s linguistic impairments reported by MacKay et al. 1998).

The complementary learning systems (see Section 3.3.1) is often associated with the standard theory but also suggests a mechanism for the transformation of episodic memories into semantic ones. Hippocampal convergence of cortical connections as well as the differing learning rates and modes of representation are good candidates for explaining why semantic memory forgets the details of individual situations and discards the context of acquisition. In sum, memory systems have different temporal capacity but the relationship is much more complex than the simple distinction between short-term and long-term memory, and the temporal component might not be a very relevant criterion to distinguish them.

3.5 Cognitive Control

3.5.1 The Need for Cognitive Control

In Section 3.2, the notion of procedural memory was introduced as the ability to acquire skills. In the motor domain, this boils down to the selection and execution of the appropriate motor commands at the right time. A straightforward strategy for motor control is to learn how rewarded each action is and to always select the most rewarded candidate. The problem is that this outcome might be a stochastic function of the action, not necessarily because the world is intrinsically stochastic but rather because the agent only has partial information. In that case, individual outcomes

might not be predictable, but an agent could still learn statistics related to the distribution of possible outcomes. Typically, this is realized by keeping track of intermediate statistics and updating them with the stream of experience. This implies a trade-off between precision and flexibility: updates should be slow to converge to the real statistics but slowness impairs adaptation in the case where the distribution is not stationary (i.e. changes over time). This happens for similar reasons as those justifying complementary episodic and semantic systems (see Section 3.3.1). A case of non stationarity is when the action leads to a state that is no longer rewarded, as captured by reward devaluation paradigms (Dickinson and Balleine 2002). This can happen if an animal is full after eating or if it has exhausted a food source. Another case is a modification of state contingency. What changes in that case is not the reward, but rather the transitions between states. The travel distance between an agent and a rewarding state might for example increase as a consequence of the apparition of an obstacle. In these cases, the commonly chosen actions are likely to result in suboptimal outcomes and the strategy must therefore be revised.

3.5.2 The Many Forms of Control

Humans and many other animals are capable of adapting to these situations. They can change their behavior by integrating information from different sources. An obvious strategy is not to learn absolute action-outcome associations but rather associations between stimulus, response and outcome: choosing to eat will result in a much better outcome when facing food than nothing. Beyond this caricatural example, the relationship between action and outcome can be arbitrarily complex. It might be that the right action is conditional on a past observation of the world, on a stimulus that was perceived in the past. For example, foraging honeybees are thought to store memories of food sources and communicate information about the direction and distance of these sources to their hive with the waggle dance (Menzel 2023). It is critical for the survival of the colony that the dance conveys correct information regarding the position of actual food

sources. Likewise, an agent might decide to take a detour after retrieving the perception of an obstacle which is changing state contingency.

It might also be that the right thing to do depends on a combination or sequence of stimuli. This is especially the case with language, note how different *My cat chased the dog* is from *My dog chased the cat*, and from *chased cat My the dog*. Similarly, the sequence of actions can be of importance, as individually sub-optimal actions can be executed in pursuit of a long-term plan. An agent incapable of executing a coherent sequence of actions to reach a goal would be analog to Buridan's ass which died midway between a stack of hay and a pail of water because its hunger equaled its thirst.

The many layers of decision making often appear to coexist within animals like mammals. While actions generally leading to positive outcomes are automatized, they can be flexibly inhibited in favor of less salient responses at will. This is realized through a process called cognitive control, which monitors incoming outcomes, stores relevant stimuli in working memory and biases actions accordingly by deploying attention to modulate salience in the premotor cortex (Posner et al. 2004). This function is mainly dependent on the integrity of the prefrontal cortex. In contrast to H.M. whose hippocampal lesion specifically impaired episodic memory, patient K.M. who had prefrontal damage showed cognitive control deficits.

3.5.3 Implicit Cognitive Control

Cognitive control, attention and working memory are intuitively (and historically) associated (and often confused) with consciousness. Once again, the consensus has shifted away from this view following evidence of these functions operating in absence of conscious awareness (Lau and Passingham 2007; Van Gaal, Ridderinkhof, Fahrenfort, et al. 2008; Van Gaal, Ridderinkhof, Scholte, et al. 2010; Gaal et al. 2012, but see Persuh et al. 2018). A taxonomy of conscious processing was proposed by Dehaene et al. (2006). They emphasize on the distinction between consciousness and attention: attending to a stimulus is a necessary condition for it to access consciousness but a stimulus can be attended to and remain unconscious if it is too weak. The

stimulus is then referred to as subliminal. In a similar way as for episodic memory, it seems that cognitive control is intimately linked to explicit representations, but that a better defining characteristic is its processing mode: the contextual modulation of automatic responses.

3.6 Metacognition and Metamemory

Both the hippocampus and prefrontal cortex are thought to underlie flexible cognition, the former binding novel neocortical combinations for memory storage (episodic memory) and the latter biasing responses when context undermines habitual actions (cognitive control), as described by J. D. Cohen and O'Reilly (1996). Combining these two structures can enable even more flexible behavior, each one executing its function on the other. Together, they are thought to support what has been called the construction system of the brain (Hassabis and Maguire 2009), that is a system that combines information form memory to predict the future through imagination (Schacter, Addis, and Buckner 2007; D. T. Gilbert and T. D. Wilson 2007; Spreng et al. 2009; Hassabis and Maguire 2009; Schacter, Addis, Hassabis, et al. 2012).

In Chapter 5, prefrontal-hippocampal interactions will be described (Figure 3.4). Hippocampus-to-prefrontal projections will be introduced as providing episodic memory context to cognitive control. The hippocampus will also be shown to encode prefrontal activity like that of other cortical areas. Consequently, the hippocampus can also replay prefrontal information for consolidating memories and improve the ability of the prefrontal cortex to perform cognitive control. Conversely, the prefrontal influence on the hippocampus will be described. In addition to modulating action selection contextually, the prefrontal cortex will be shown to control episodic memory operations. The control of encoding has been mentioned in Section 3.4.1. The contextual control of episodic memory retrieval will also be described thoroughly.

These functions perhaps lie at the highest levels of our cognition. They are functions that exert influence on other functions and can thus be re-

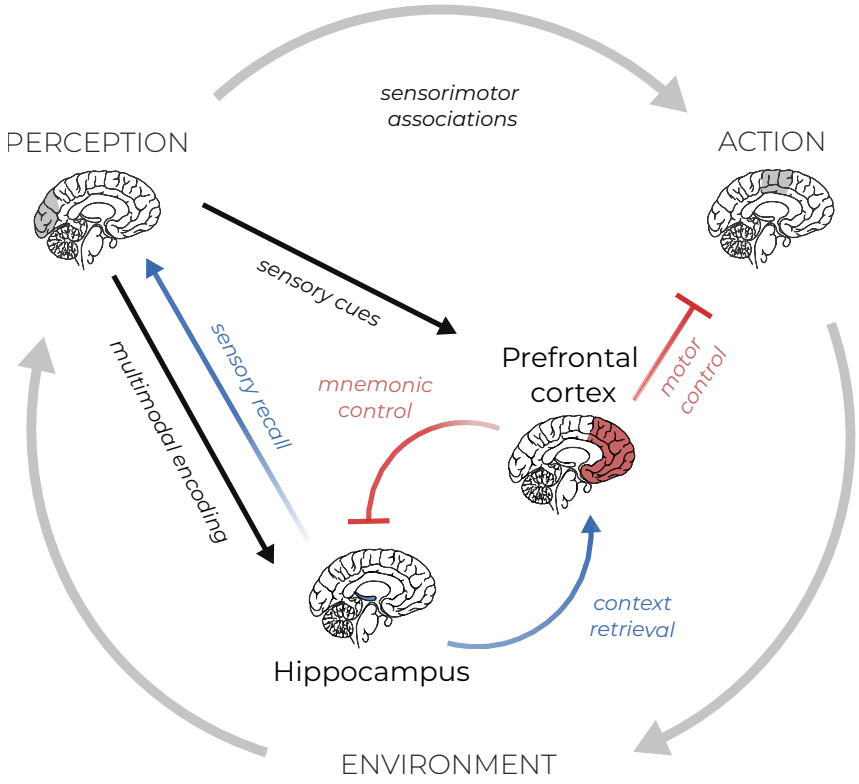


Figure 3.4: Proposed interactions between the prefrontal cortex (cognitive control) and the hippocampus (episodic memory). The prefrontal cortex controls action selection and hippocampal recall. In return, the hippocampus provides context to the prefrontal cortex in addition to augmenting perceptual representations with memory.

ferred to as metacognitive or metamemory. A central goal of this thesis is to argue that neuroscience, cognitive science and AI are now mature enough to tackle these topics, and that these functions could augment machine learning Chateau-Laurent and Alexandre (2021a), account for creative processes Chateau-Laurent and Alexandre (2021b) and underlie the construction system of the brain.

4 Episodic Memory

An agent remembering an episodic memory

Generative AI

4.1 Introduction

As presented in the preceding chapter 3, episodic memory can be distinguished from other memory systems using multiple criteria. These criteria can be questioned individually, but collectively shed light on the nature of episodic memory. The content of episodic memory typically influences

explicit processes and can often be reported verbally, which contrasts with implicit systems like procedural memory. It can potentially be stored for a lifetime and provide context to other cognitive functions like decision making and cognitive control. It can also be consolidated into more perennial semantic forms.

What perhaps best characterizes episodic memory is its mode of processing, that is its ability to encode and store novel relations between known objects flexibly and rapidly (Figure 3.2) by binding previously unrelated cognitive representations. It is a memory form that stores sequences of multimodal information corresponding to specific situations. In order to do that, it must have access to information of a wide range of cognitive modules and encode their conjunction. It must also be able to learn these conjunctive bindings rapidly, as similar information might not be experienced multiple times, and also how this information evolves through time. In addition of disambiguating different memories, it will become clear in this chapter that the brain regions underlying episodic memory are also able to quickly extract statistical regularities among similar situations. When necessary, the memory must be able to retrieve a memory by filling cognitive modules with information absent at the time of retrieval, and perhaps replace information of the current situation. Episodic memory then conducts cognition for the organism to re-experience all aspects of the episode in the right order. This perhaps idealistic view corresponds to the current understanding of how episodic memory works, interacts with other cognitive functions and supports memory consolidation.

It is common to regard the episodic content as a what-where-when conjunction, but it is hard to apprehend the complete potential of binding the content of cognitive functions as diverse as those exhibited by humans. The "what" can indeed include semantic, emotional, social, abstract and even metacognitive aspects that are encoded at multiple spatiotemporal scales. Another aspect of episodic content is that it refers to the self, contrasting with the allocentric nature of semantic memory. It follows from the multimodal contextualization of information and underlies autobiographical memory (i.e. memory of one's life events) and autonoetic consciousness,

that is the ability to place oneself in the past and future. On the other hand, the allocentric nature of semantic memories can be explained by abstracting away details of individual experiences.

This chapter introduces the anatomy (Section 4.2), representations (Section 4.3) and dynamics (Section 4.4) underlying episodic memory in the medial temporal lobe. Then, computational models for rapid pattern and sequence learning, as well as models for the progressive extraction of statistical regularities, are presented in Section 4.5.

4.2 Anatomy

The medial temporal lobe is the seat of episodic memory (Scoville and Milner 1957). It is mainly composed of the hippocampus and parahippocampal regions like the entorhinal cortex. These regions collectively support the encoding, storage and retrieval of episodes.

Region	Count
EC2	110,000
EC3	250,000
DG granule cells	1,200,000
Hilar mossy cells	50,000
CA3	250,000
CA1	390,000
EC5	330,000

Table 4.1: Number of principal neurons in different regions of the rat's medial temporal lobe, as retrieved from Cutsuridis et al. (2019).

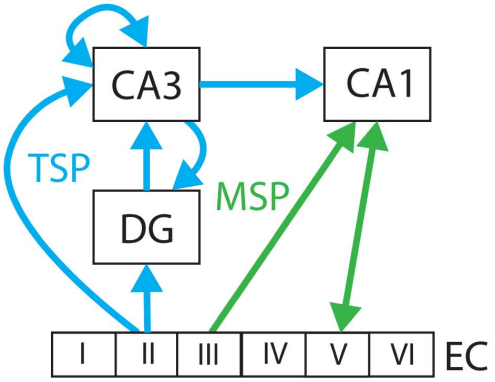


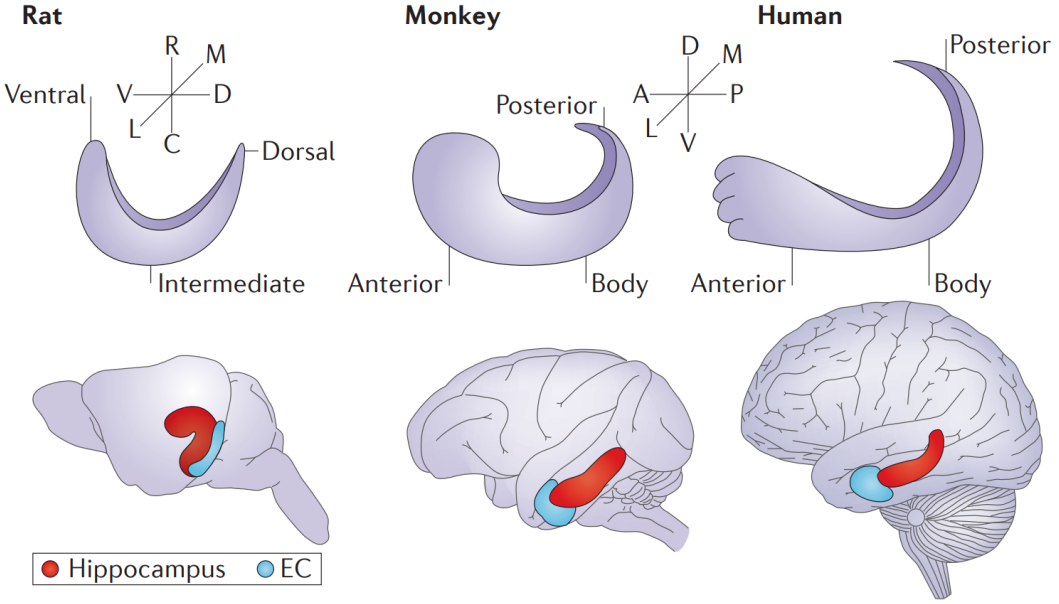
Figure 4.1: Schematic illustration of the trisynaptic (blue) and monosynaptic (green) pathways connecting the entorhinal cortex and hippocampus. Adapted from Ketz et al. (2013) with the CA3-to-DG backprojection through the hilus.

4.2.1 Entorhinal Cortex

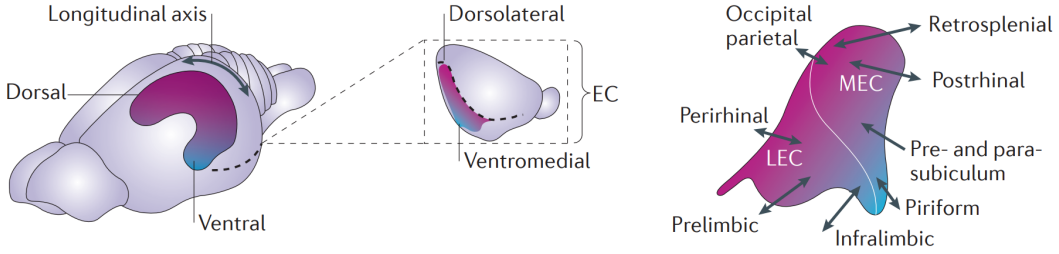
The entorhinal cortex (EC) is part of the medial temporal lobe and is located next to the hippocampus. In terms of connectivity, the entorhinal cortex is the main interface between the neocortex and the hippocampus. Thus, the entorhinal cortex is the place of convergence of many different sources. It is divided into medial and lateral regions that provide inputs of different kinds to the hippocampus (Howard Eichenbaum, Yonelinas, et al. 2007; Manns and Howard Eichenbaum 2006). The lateral portion connects to the "what" pathway, a neocortical circuit involving mainly the temporal lobe that conveys semantic information about the perceived scene (e.g. features of an object). On the other hand, the medial portion of the entorhinal cortex connects to the "where" pathway that conveys information about the spatial context (e.g. where the object is located) through the parietal lobe. Both entorhinal portions connect to the hippocampus which is thought to bind information from the two streams (Howard Eichenbaum, Yonelinas, et al. 2007; Whittington, T. H. Muller, et al. 2020).

Pretty much orthogonal to the medial-lateral divide of the entorhinal cortex is the longitudinal axis (Figure 4.2), as reviewed by Strange, Menno P Witter, et al. (2014). In rats it extends from the dorsolateral parts to more ventromedial portions. These parts preferentially connect to the dorsal and ventral hippocampus ends respectively (Figure 4.2b). In macaque and humans this connectivity gradient is oriented along the anteroposterior axis of the hippocampus (Figure 4.2a).

Like in other cortical regions, neurons of the entorhinal cortex are grouped into layers (Figure 4.3). Six layers compose the entorhinal column, layers II, III, V and VI containing most neurons. Telencephalic inputs converge to neurons of layers II and III. Layer II then connects to the dentate gyrus and CA3 of the hippocampus, while neurons from layer III project to CA1. CA1 and the subiculum in turn send the output of the hippocampus to layer V of the entorhinal cortex, which is then transmitted back to telencephalic structures. Layers II, III and V of the rat entorhinal cortex are composed



(a) Top: The rat hippocampus is oriented such that the longitudinal axis corresponds to the ventral-dorsal direction. In macaque monkeys and humans, the longitudinal axis is described as anteroposterior. Bottom: Position of the entorhinal cortex relative to the hippocampus in these species.



(b) Topographical arrangement of the rat entorhinal-hippocampal connectivity. The dorsolateral portion of the entorhinal cortex (magenta) is preferentially connected to the dorsal hippocampus while more ventromedial parts connect to more ventral ends of the hippocampus.

Figure 4.2: Schematic illustrations of the entorhinal–hippocampal longitudinal axis in different species. Taken from Strange, Menno P Witter, et al. (2014).

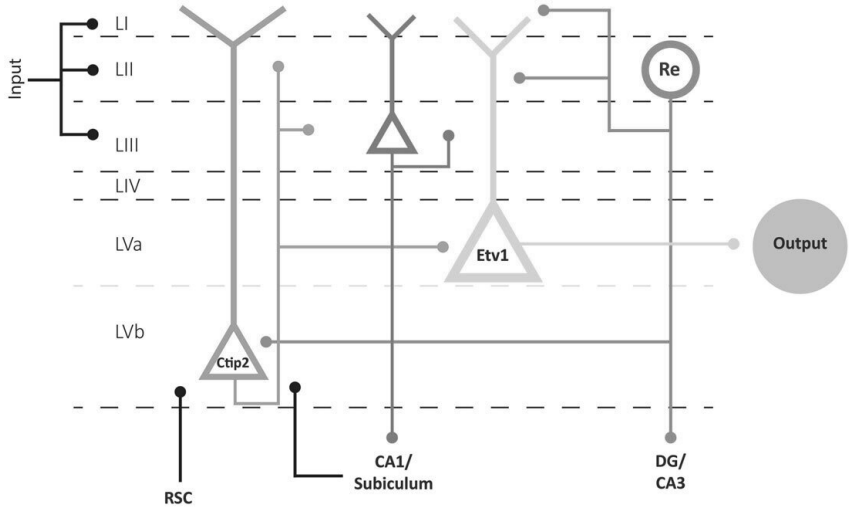


Figure 4.3: Connectivity of the entorhinal column. Abbreviations: Re, reelin-expressing neurons; RSC, retrosplenial cortex. Adapted from Menno P Witter, Doan, et al. (2017).

of approximately 110,000, 250,000 and 330,000 neurons respectively (Table 4.1).

4.2.2 Dentate Gyrus

The dentate gyrus (DG) is part of the trisynaptic circuit of the hippocampal formation, that is a pathway with three synapses separating the entorhinal cortex from CA1 (Figure 4.1). The dentate gyrus is mainly populated by granule cells and mossy cells located in the hilus. Granule cells of the dentate gyrus receive projections from the entorhinal cortex through the perforant path, originating mainly from layer II. Granule cells in turn project to both mossy cells of the dentate gyrus and CA3 pyramidal cells. Mossy cells receive inputs from both granule cells and CA3 pyramidal cells, and in turn project back to granule cells. The circuit is thus two-fold recurrent, with a small intra-DG loop, and a bigger loop that involves CA3.

Granule cells have unique properties. First, they are the most abundant neurons of the hippocampal formation, as indicated in Table 4.1. Secondly, their activity appears to be among the most sparse in the brain (M. Jung and B. McNaughton 1993; Diamantaki et al. 2016). In a study of freely mov-

ing rats, M. Jung and B. McNaughton (1993) found only 12% of granule cells that fired faster than 0.5Hz. This sparsity is thought to be mediated by inhibitory basket cells (Andersen 1975). Thirdly, new granule cells are produced throughout life in fish, reptiles, birds, rodents, nonhuman primates and humans (Augusto-Oliveira et al. 2019) while such adult neurogenesis has not been documented elsewhere besides the subventricular zone. Newborn granule cells integrate into the network of preexisting granule cells (Toni et al. 2008). Furthermore, they exhibit high excitability and enhanced synaptic plasticity (Schmidt-Hieber et al. 2004; Trinchero et al. 2017), unlike mature granule cells which show low excitability and plasticity to the extent that they are sometimes considered functionally retired (as critically reviewed by Lopez-Rojas and Kreutz 2016). The role of newborn neurons in learning is controversial but experiments suggest that neurogenesis is necessary to reduce memory interference between similar events (Yau et al. 2015). Finally, granule cells make contact with CA3 cells with very low probability and uniquely powerful "detonator" synapses, such that a single action potential of a granule cell is sufficient to generate an action potential in the CA3 neurons it targets (Henze et al. 2000; Vyleta et al. 2016). Combined with the steep sigmoidal response of the dentate gyrus to entorhinal inputs (Renata Bartesaghi et al. 1995), the detonator synapse participates in making the hippocampus a region with an almost all-or-none response profile (Bartesaghi et al. 2006).

4.2.3 CA3

Properly speaking, the hippocampus is the small curved formation called Cornu Ammonis, which is divided in four portions and does not include the dentate gyrus[1]. The third portion, CA3, is mainly populated by pyramidal neurons receiving inputs from granule cells of the dentate gyrus (second synapse of the trisynaptic circuit). Pyramidal cells of CA3 are known for their axons composed of recurrent collaterals that branch and connect

[1] However, the dentate gyrus is part of the hippocampal formation and it is very common to include the dentate gyrus when speaking about the hippocampus.

to other CA3 pyramidal cells, and Schaffer collaterals that project to CA1 pyramidal cells, completing the trisynaptic pathway (EC $\rightarrow_1$ DG $\rightarrow_2$ CA3 $\rightarrow_3$ CA1). A connection often overlooked is the afference of the second entorhinal layer. An even lesser known connection is the backprojection from CA3 pyramidal cells to hilar mossy cells of the dentate gyrus (Scharfman 2007).

4.2.4 CA1

CA1 is the largest region of the hippocampus proper, it comprises approximately 390,000 pyramidal neurons in rats (Table 4.1). CA1 receives inputs from both CA3 Schaffer collaterals (completing the trisynaptic pathway) and layer III of the entorhinal cortex (monosynaptic pathway). These two major sources can be distinguished in terms of destination. While entorhinal afferents connect to the stratum lacunosum-moleculare of CA1 pyramidal neurons, Schaffer collaterals make contact at the stratum radiatum. CA1 neurons in turn project to layer V of the entorhinal cortex. They also project to the subiculum which is not central to this thesis but potentially plays a role in spatial navigation, memory and stress regulation (see the reviews of S. M. O'Mara et al. 2001; S. O'Mara 2005).

4.3 Representation: the Hippocampus as a Cognitive Map

4.3.1 Grid Cells

Hafting, Fyhn, Molden, et al. (2005) found that cells of the rat medial entorhinal cortex are tuned to represent a map of the spatial environment. They revealed the hexagonal activity patterns of what they called "grid cells". These patterns are called grid fields and can be described by three parameters that vary from cell to cell: orientation, spacing and spatial phase. Grid cells are topographically organized such that neighboring neurons are tuned with a similar orientation and spacing (Hafting, Fyhn, Molden, et al.

2005). This is also the case for phase at a much finer scale (Heys et al. 2014). On top of that, spacing increases along the dorsomedial-to-ventrolateral anatomical axis of the medial entorhinal cortex (Hafting, Fyhn, Molden, et al. 2005), with discrete steps suggesting that grid cells are organized in functional modules (Stensola et al. 2012). Grid cells and an axis along which grid spacing increases have also been found in monkeys (Killian et al. 2012).

Perhaps even more interesting is the persistence of their tuning in the absence of visual cues, in total darkness. According to the authors, this property suggests that grid cells play an important role in path integration, which is the capacity of an agent to maintain a representation of its spatial location, and update it with self-motion information. This theory was later supported by the emergence of grid-like cells in artificial recurrent neural networks trained to path integrate (Banino et al. 2018; Cueva and Wei 2018; Sorscher et al. 2019; Whittington, T. H. Muller, et al. 2020; Nayebi et al. 2021). These findings, however, have been reviewed and challenged by Schaeffer et al. (2022) who did not find grid fields unless post-hoc implementation choices were made. They proposed that the path integration objective is perhaps necessary but not sufficient for grid fields to emerge. Beyond this controversy, it is very common for computational models of grid cell formation to assume the medial entorhinal cortex is integrating velocity information. Some of them rely on the notion of continuous attractors (e.g. Burak and Fiete 2009; Fuhs and Touretzky 2006; Bruce L McNaughton, Battaglia, et al. 2006; Guanella et al. 2007) and predict that network dynamics lie on a toroidal manifold. A recent study has provided support for this theory by recording from 7.671 entorhinal neurons simultaneously and showing their activity span a two-dimensional ring that also persists during sleep and across environments (Gardner et al. 2022). Another normative theory of interest is that grid fields emerge as a result of the eigendecomposition of a matrix predicting future state occupancy (Stachenfeld et al. 2017). This relates to the successor representation framework described in Section 4.5.5.

4.3.2 Place Cells

The hippocampus also comprises so-called place cells modulated by spatial location, but whose firing fields (named place fields) do not have the hexagonal structure of grid cells (O'Keefe and Dostrovsky 1971). The discovery of these place cells supported the old theory that the hippocampus encodes a cognitive map of the environment (Tolman 1948).

When the environment changes, the activity of place cells changes such that their field is reallocated to a different position, they remap (R. U. Muller and Kubie 1987). The continuous morphing of the environment leads to a discontinuous global remapping, which is reminiscent of attractor dynamics (Wills et al. 2005).

An intuitive explanation as to why these place fields emerge is that hippocampal neurons receive information from grid cells of variable tuning, and that place fields naturally emerge as the conjunction of their respective grid. This idea has been formulated and mathematically validated by Solstad et al. (2006). Another demonstration of this phenomenon was recently provided by Santos-Pata et al. (2021) who trained a network to simply output its input in an autoencoder fashion, with inputs resembling entorhinal activity. Once trained, place cells could be observed in intermediate hippocampal layers and exhibited physiological phenomena such as rate modulation and global remapping. The network however received not only medial entorhinal input but also a lateral entorhinal source. While the theory of Solstad et al. (2006) only relies on grid cells and thus does not ascribe a role to the lateral entorhinal cortex, it has been noted that the hippocampus is the place of convergence of the "what" and "where" cortical pathways (Howard Eichenbaum, Yonelinas, et al. 2007; Manns and Howard Eichenbaum 2006; Whittington, T. H. Muller, et al. 2020), and that place fields might emerge as the conjunction of these two sources of information. The hippocampus thus would encode items in their context by associating semantic information provided by the lateral entorhinal cortex with position information encoded by grid cells. A prediction of this alternative theory is that semantic information exerts an influence on the activity of place cells,

which has been confirmed by the observation of landmark-vector cells that respond to object manipulations and signal distance to landmarks (Deshmukh and Knierim 2013).

Unlike grid cells, place cells do not seem to be topographically arranged (M. A. Wilson and Bruce L McNaughton 1993; M.-B. Moser et al. 1995 ; but see Nakamura et al. 2010). However, their scale also increases in the dorsal-to-ventral direction in rodents (Min W Jung et al. 1994; Maurer et al. 2005; Kjelstrup et al. 2008). Strange, Menno P Witter, et al. (2014) reviewed a set of results describing this axis both anatomically and functionally. The multiscale representations employed by grid and place cells are thought to support the hierarchical organization of memory and navigation. A good example is the task of transitive inference, in which participants must infer that A is associated to C after being informed that A is associated to B and B is associated to C. The latter associations are of a lower order than the association between A and C. Studies showed that this inference is dependent on the ventral hippocampus of rats (Komorowski et al. 2013) and the anterior hippocampus of humans (Heckers et al. 2004; Preston, Shrager, et al. 2004). Since the size of place fields in these regions is larger, their activity covers more spatial locations and more items which are potentially not overlapping in dorsal/posterior regions, a mechanism which probably underlies the aforementioned transitive inference (Strange, Menno P Witter, et al. 2014).

Mehta et al. (1997) showed that place fields are shaped by experienced spatial trajectories. When rats travel across a linear track multiple times in the same direction, their place fields shift their center of mass in the opposite direction and increase in size. They thus become asymmetric. Furthermore, some place cells not only fire according to the position in space, but also depending on the temporally close past and future trajectory (Wood et al. 2000; L. M. Frank et al. 2000). L. M. Frank et al. (2000) indeed found cells that only fire when the animal occupies its place field and comes from a specific part of the environment. Some cells fired when the rat reached their field and was about to go in a specific direction. Results related to these so-called splitter cells (their firing is split by a prospective or retrospective signal) have recently been reviewed by Duvelle et al. (2023).

The name place cell intimately links this category of neurons to spatial navigation and blurs the relationship with memory. While these cells obviously code for spatial variables, they are modulated by other variables too as suggested by the remapping and splitter cells phenomena discussed above. Recently, it has been found that place fields emerge in a model whose aim is to compress sensory experiences, and that place cells might simply be memory cells, in that their activity reflects an efficient strategy for storing correlated patterns (Benna and Fusi 2021). Their model for example accounts for history dependent splitter cells which decorrelate similar representations on the basis of the trajectory. This is in line with the older proposal of Howard Eichenbaum, Dudchenko, et al. (1999) that the observation of place cells does not suffice to conclude that hippocampal neurons primarily encode space.

In the same vein, many studies have found hippocampal cells modulated by nonspatial variables such as sound (Sakurai 2002; Aronov et al. 2017), odors (Eichenbaum et al. 1987), faces and objects (Fried et al. 1997). These findings further nourish the long legacy of the cognitive map idea by Tolman (1948). A modern generalization is the Tolman-Eichenbaum machine model of Whittington, T. H. Muller, et al. (2020) suggesting that the entorhinal-hippocampal system integrates any cognitive variable with a common mechanism, including the most abstract ones. They generalize spatial path integration to abstract tasks typically associated with relational memory (e.g. social hierarchies and transitive inference) which have been found experimentally to rely on the hippocampus (N. J. Cohen et al. 1999; Konkel and N. J. Cohen 2009; Dusek and Howard Eichenbaum 1997; Kumaran, H. L. Melo, et al. 2012). In this model, place cells emerge as the conjunction of sensory and structural information sent by the lateral and medial entorhinal cortices respectively, as proposed by Manns and Howard Eichenbaum (2006), giving rise to the landmark cells observed by Deshmukh and Knierim (2013). A prediction of the Tolman-Eichenbaum machine that the authors validate with existing experimental data is that place cells do not remap randomly but rather stay consistent with the entorhinal grid code,

such that a place cell co-firing with a grid cell will still co-fire with it following remapping.

4.4 Dynamics: Oscillations as Memory Operations

Brain oscillations refer to rhythmic electrical activity generated by populations of neurons. They are prominent in the hippocampus and are thought to reflect and support cognitive operations (Lisman and O. Jensen 2013; Buzsáki and Watson 2012).

4.4.1 The Entorhinal Buffer

Egorov et al. (2002) described how neurons from layer V of the entorhinal cortex (EC5) in slices of rat brains respond to their input. They demonstrated that EC5 (CA1->EC5) cells could exhibit stable patterns of activity following stimulation for an apparently indefinite period of time (up to 13 min). The likelihood of exhibiting such persisting activity was a nonlinear function of the amplitude (nA) and duration (s) of the stimulation, parameterized differently for each neuron. Furthermore, the authors described multiple stable levels of persistent firing frequency. They showed that up and down transitions in firing frequency could be induced by excitatory and inhibitory stimulations respectively. In a subsequent study, Fransén et al. (2006) intended to reproduce and extend these results. Beyond multiple levels of stability, they observed a continuum of stable states. These results suggest that EC5 neurons implement what is called an integrator, in that they integrate their input over time. Persistent activity has also been observed in the second and third layers of the entorhinal cortex (Tahvildari et al. 2008).

Hasselmo and Stern (2006) reviewed these results and proposed that the entorhinal cortex uses these mechanisms to implement a working memory for novel information. In short, a stimulus can cause persistent activity in the entorhinal cortex, making stimulus information available to that region in the short term after stimulus removal. Since mechanisms

for generating persistent activity in the entorhinal cortex occur at the level of individual cells, this memory does not depend on synaptic connectivity and can thus store stimuli that were never encountered without synaptic plasticity. Contrastingly, the prefrontal and parietal cortices also exhibit persistent activity that is thought to underlie working memory, but this is thought to depend on the recurrent synaptic excitation of cells. Hasselmo and Stern (2006) therefore argued that this connectivity-based implementation of working memory is insufficient for storing novel stimuli that did not shape the recurrent circuitry beforehand. The authors related their theory to A. Baddeley (2000)'s addition of a multimodal episodic buffer to his earlier working memory model (A. D. Baddeley and Graham 1974).

A working memory is typically expected to exhibit a load effect, that is a modulation of its activity depending on the number of stimuli stored. In a fMRI study, Schon et al. (2016) employed a delayed matching-to-sample task to probe whether medial temporal lobe regions exhibit such load effect. Participants were shown a sequence of two (low load) or four (high load) novel landscape pictures, maintained the stimuli in memory during a delay period, and were asked to determine whether a test picture matched one of the previous landscapes of that trial or not. As a result, the activity of the entorhinal cortex was greater during maintenance of four stimuli than two, consistent with the theory of Hasselmo and Stern (2006).

A related line of work comes from the study of the role of oscillatory dynamics in sequence learning (O. Jensen and Lisman 2005; Lisman and Idiart 1995; O. Jensen, Lisman, et al. 1996; O. Jensen and Lisman 1996). In the review of O. Jensen and Lisman (2005), the authors proposed that the entorhinal cortex is a short-term memory oscillatory buffer that holds information about the recent sequence of events and sends it to the hippocampus at a pace optimal for long-term memory storage. This will be described in more details in the context of the Socratic model (Section 4.5.5).

We now turn to provide a description of the information sent from the entorhinal cortex to the hippocampus. These two regions closely interact, as the entorhinal cortex constitutes both the major input source and major output destination of the hippocampus. This loop suggests possible

feedback effects and therefore renders the characterization of hippocampal input challenging. Studies have attempted to study the nature of efferent entorhinal information by describing the entorhinal activity in isolation, under conditions of hippocampal inactivation. In one such study, Hafting, Fyhn, Bonnevie, et al. (2008) reported theta oscillations and phase precession in entorhinal grid cells. A cell exhibits phase precession relative to a population-wide oscillation when it fires earlier and earlier in each successive oscillation cycle. During periods of activity, the hippocampus and the entorhinal cortex are indeed dominated by a strong theta oscillation (4-10Hz). Place cells of the hippocampus typically exhibit phase precession with respect to this rhythm. When entering its place field, a cell will fire at the end of the theta cycle, and as the animal progresses the cell will fire at earlier phases until the animal exits its field and the neuron stops firing. Thus, the population of place cells not only encodes position through the identity of cells firing, but also through the precise timing of their activity with respect to theta. Since Hafting, Fyhn, Bonnevie, et al. (2008) showed that entorhinal cells exhibit phase precession in absence of the hippocampus, phase precession in the hippocampus is at least partially inherited from afferent regions. Interestingly, phase precession was absent in entorhinal neurons of layer III, the layer that connects to CA1, suggesting that the monosynaptic pathway does not convey such timing information. Neurons in entorhinal layer II, CA3 and CA1 do exhibit precession, and CA1 probably inherits it from CA3. Theta sequences are sequences of behavioral states that can be decoded from the activity of hippocampal populations of neurons during individual theta cycles. While this circuit-level phenomenon is probably facilitated by cell-level phase precession, some results suggest that phase precession is already present in the first traversal of a linear track while theta sequences rapidly emerge in subsequent experiences (Feng et al. 2015).

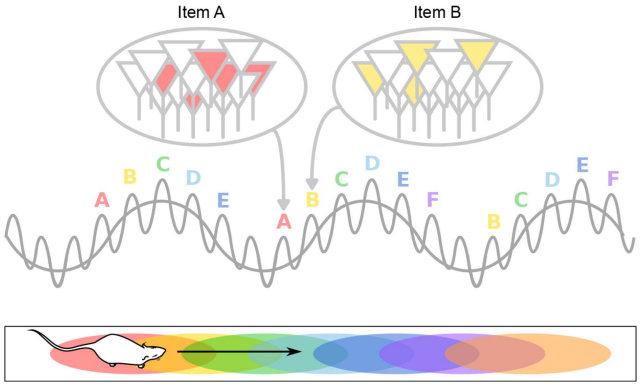


Figure 4.4: Theta-gamma code. As the animal traverses its environment, successive places are represented on successive gamma cycles nested within each theta cycle. Theta cycles represent a rolling window of states, such that spatial positions are represented earlier and earlier in successive theta cycles. Adapted from Drieu and Zugaro (2019).

4.4.2 Theta-Gamma Code

Nested in theta waves of the hippocampus are oscillations of the gamma band (25-140Hz) that are approximately seven times faster than theta oscillations. In fact, increasing the number of bins dividing theta cycles helps decoding the position of the animal up to six bins (O. Jensen and Lisman 2000). Furthermore, place cells in CA1 typically fire during seven successive theta cycles (Skaggs, Bruce L McNaughton, et al. 1996). From these results emerged the view that hippocampal activity is organized with a theta-gamma code when the animal is active (Lisman and O. Jensen 2013). Information is buffered in the entorhinal cortex and sent to the hippocampus in such a way that each state of the ongoing sequence is represented in each successive gamma cycle of the ongoing theta cycle (Figure 4.4). As the animal moves and perceives new information, the new theta starts with a more advanced state and ends with a previously unrepresented state. This is reminiscent of a rolling window of states, or a queue data structure in computer science. Phase precession naturally emerges from this code, as successive states are represented earlier and earlier in the theta cycle.

Hasselmo, Bodelón, et al. (2002) reviewed a set of results describing the role of theta rhythm in scheduling memory operations. Information com-

ing from the entorhinal cortex is sent to the hippocampus at the peak[2] of the theta cycle (Brankačk et al. 1993). Meanwhile, input from CA3 to CA1 is weak. In this same phase, CA3 synapses to CA1 can undergo long-term potentiation (LTP) Hölscher et al.; Wyble et al. (1997; 2000). At the trough of theta, inputs from the entorhinal cortex are weaker, while information in CA3 is transmitted to CA1 Brankačk et al. (1993). Meanwhile, LTP is weak and depotentiation can instead be induced Hölscher et al. (1997). Hasselmo, Bodelón, et al. (2002) proposed that the peak is the phase during which cortical information is sent for synaptic storage in the hippocampus, while stored memories are retrieved at the trough (Figure 4.5). This proposal derives from the standard framework of hippocampal function described in Section 4.5 stating that memories are stored in recurrent collaterals of area CA3.

Gamma oscillations are not uniform and their diversity provides key information regarding how CA1 integrates information from CA3 and the entorhinal cortex. In freely moving rats, Colgin et al. (2009) found that CA1 gamma oscillations could be divided into slow (25-50Hz) and fast (65-140Hz) gamma bands. Within each theta cycle, gamma power was usually concentrated in one of the two bands, and theta cycles with the same gamma bands were typically clustered. The EEG activity of CA3 and layer 3 of MEC were recorded simultaneously. During episodes of slow gamma, gamma activity of CA1 was synchronised with CA3, but not with MEC. On the other hand, episodes of fast gamma were characterized by high synchrony between CA1 and MEC, and not between CA1 and CA3. Furthermore, fast gamma oscillations were strongest at the theta trough, while slow gamma oscillations were strongest at the descending phase of theta.

Based on the observation that place cells sometimes employ prospective coding with theta sequences that represent locations ahead of the animal Battaglia et al. (2004), Bieri et al. (2014) analyzed the spatial coding modes at play during slow and fast gamma rhythms. Consistent with the prediction that prospective coding reflects memory retrieval and retrospective coding

[2] As recorded in the LFP of the stratum pyramidale of CA1.

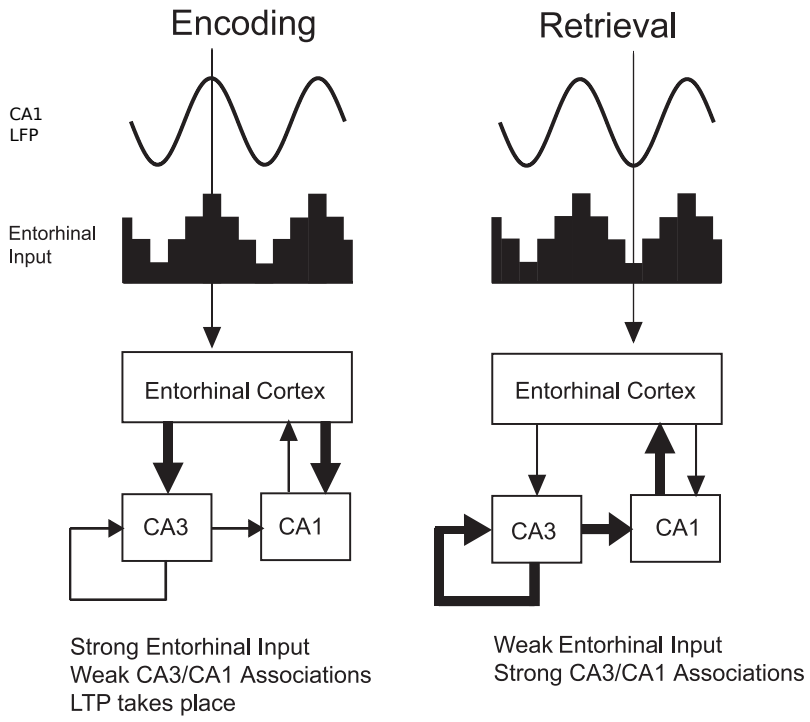


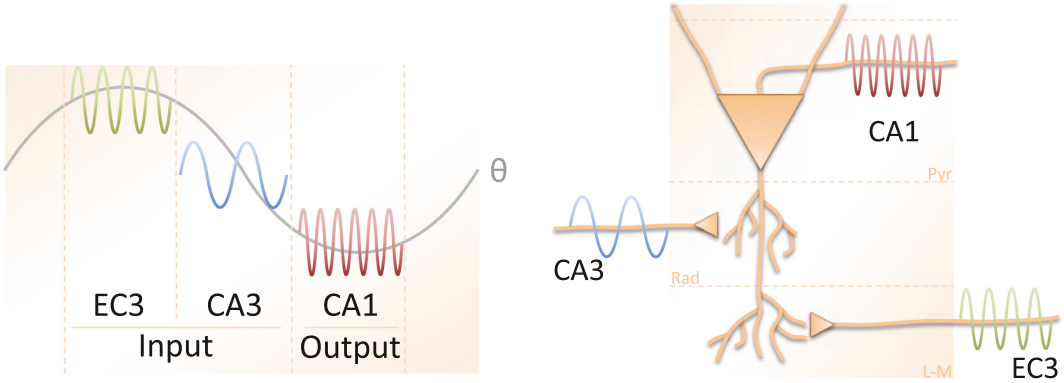
Figure 4.5: Model of Hasselmo, Bodelón, et al. (2002). Left: The peak of theta is the phase of encoding, as input from the entorhinal cortex is strong and LTP is induced. Right: At the trough of theta, input from the entorhinal cortex is weak and CA1 is instead firing in response to CA3, which suggests this phase reflects the recall of past memories. Adapted from Kunec et al. (2005).

reflects encoding, CA3-associated slow gamma increased during prospective coding and EC-associated fast gamma increased during retrospective coding. In the linear track, spikes occurred earlier during slow gamma than during fast gamma, and CA1 cells retrieved upcoming positions during slow gamma and encoded past positions during fast gamma.

If gamma cycles represent successive spatial positions as stated by the theta-gamma code theory, slow gamma should encode shorter trajectories than fast gamma. C. Zheng et al. (2016) in fact showed that slow gamma encodes longer trajectories as multiple locations are encoded in slow (but not fast) gamma cycles. Fast gamma, on the other hand, represented very short sequences. This is potentially related to the absence of phase precession in layer III of the entorhinal cortex (Hafting, Fyhn, Bonnevie, et al. 2008).

While these results shed light on hippocampal information processing, the identified phase of fast gamma (theta trough) is at odds with Hasselmo, Bodelón, et al. (2002)'s view on the relationship between theta phases and memory operations, as pointed out by Aguilera et al. (2022). On top of that, these results suffer from limitations (Aguilera et al. 2022; Colgin 2015; Schomburg et al. 2014). First, the tasks performed by the rats in these studies, open field in Colgin et al. (2009) and linear track in Bieri et al. (2014), do not require allocentric memory. Secondly, Colgin et al. (2009) used single-site recording which prevents the characterization of the topology of CA1 oscillations and the identification of their source.

A finer study was performed by Schomburg et al. (2014) that overcame the limitations of previous investigations. They recorded from up to 256 sites along the transverse axis of the hippocampus while rats were performing navigational tasks with varying memory demands. Using independent component analysis, they found that CA1 oscillations could be divided into not two, but three distinct gamma bands, each of which supporting different modes of communication within the hippocampus (Figure 4.6). At the peak of theta, oscillations were dominated by medium gamma frequencies (60-120Hz) located mainly in the stratum lacunosum-moleculare, the source of which was the firing of EC3 cells. During the descending phase of theta, slow gamma (30-80Hz) power increased, as CA3 sent input



(a) Preferred theta phase of medium, slow and fast gamma oscillations.

(b) Spatial organization of CA1 afferences and efferences with their associated gamma bands.

Figure 4.6: Schematics of the three CA1 gamma bands identified by Schomburg et al. (2014). Adapted from Butler and Paulsen (2014).

to the stratum radiatum. Finally, at the trough of theta, CA1 output firing was phase-locked to fast gamma (> 100Hz), a frequency band dominating the stratum pyramidale. Interestingly, the phases at which CA1 is synchronized with EC and CA3 correspond to the encoding and retrieval phases of Hasselmo, Bodelón, et al. (2002).

Finally, hippocampal dynamics and memory operations are also influenced by neuromodulation. In novel environments, the preferred phase of CA1 firing shifts to the peak of theta (Douchamps et al. 2013; Fernández-Ruiz, Oliva, Nagy, et al. 2017), consistent with Hasselmo, Bodelón, et al. (2002). This effect has been shown to depend on the release of acetylcholine as a novelty response (Douchamps et al. 2013). Dopamine is also thought to play a modulatory role in memory encoding (Lisman and Grace 2005).

4.4.3 Sharp Wave Ripples

Another oscillatory pattern characterizing hippocampal activity is the sharp wave and its nested ripples (140-200Hz). While the theta-gamma state is dominant when the animal is active, ripples are characteristic of immobility, consummatory behaviors, and slow-wave sleep. They are generated in CA3 and transmitted to the dentate gyrus and CA1.

Ripples are associated to the phenomenon of replay, which was first observed independently of these oscillations. Pavlides and Winson (1989) recorded CA1 place cells and found those whose field was visited during waking behavior were more active during following sleep episodes. Skaggs and Bruce L McNaughton (1996) later showed that a place cell firing before another was likely to fire before in subsequent sleep, and conversely. Crucially, no such temporal bias correlation was found between awake activity and activity during the preceding sleep session. These results suggest that spatial trajectories experienced during awake behavior are learned and replayed during sleep, the temporal organization of which being preserved. These replay events happen during sharp-wave ripples at 10 to 20 times the speed of real trajectories (Nádasdy et al. 1999).

First, replay is thought to support memory consolidation. Hippocampal replay indeed co-occurs with replay in the cortex (Peyrache et al. 2009; Ji and M. A. Wilson 2007). Furthermore, suppressing ripple events during rest induces a significant memory deficit (Girardeau et al. 2009; Ego-Stengel and M. A. Wilson 2010). Replay-like events can also happen while awake (Kudrimoti et al. 1999), representing path to a goal or the trajectory leading to that goal once reached in reverse order (Foster and M. A. Wilson 2006; Diba and Buzsáki 2007; Pfeiffer and Foster 2013; Ambrose et al. 2016). Furthermore, replay-like events can represent never experienced trajectories (Gupta et al. 2010). Interestingly, the same phenomena can emerge in artificial agents seeking to maximize expected future rewards (Cazé et al. 2018; Mattar and Daw 2018), providing a normative account of their function beyond simple memory consolidation. These results indeed suggest that the hippocampus plays a role in planning, in line with fMRI studies showing the hippocampus is not only necessary for remembering the past, but also to imagine future-oriented novel situations (Hassabis, Kumaran, Vann, et al. 2007; Hassabis, Kumaran, and Maguire 2007).

Pezzulo et al. (2017) proposed a framework for explaining the many modes of hippocampal sequential activity according to the level of task engagement. In their view, prospective theta sequences support online predictive processing and online replays support planning, prospection and informa-

tion seeking. Offline replays either consolidate memory of recent experience when necessary or reflect a more disconnected sampling process of an internal model of the world for self-consistency, model pruning and homeostasis.

4.5 Computation: The Standard Framework, Limits and Alternatives

Marr (1971) proposed the first influential theory of hippocampal involvement in memory functions. In essence, Marr proposed that the hippocampus temporarily stores sensory information before sending it to the more permanent neocortical store. As noted by Willshaw et al. (2015), Marr thus laid the foundations for the complementary learning systems theory (Section 3.3.1) and theories of consolidation (Section 3.4.1). He was very influenced by a learning rule proposed by Hebb (1949). This rule serves as a core mechanism for storing associative memories in his model. After storing memories with Hebb's rule, the goal of the hippocampus is to complete novel information with information stored in memory.

This model dominates the understanding of hippocampal function so much that it has been termed the standard model (Nadel and Moscovitch 1997; Nadel, Samsonovich, et al. 2000) or standard framework (Cheng 2013). The term framework is preferred considering that many details of the original model of Marr have been discarded. Instead of being used as is, the work of Marr has influenced the development of many models (Bruce L McNaughton and Morris 1987; Buzsáki 1989; Treves and Rolls 1992; Treves and Rolls 1994; O'Reilly and McClelland 1994; McClelland et al. 1995; Becker 2005; Kassab and Alexandre 2018) bringing additional elements, some of which will be described in the following sections.

4.5.1 Main Principles

The standard framework of the hippocampus derives from the need for an initial memory store postulated by the complementary learning systems theory (Section 3.3.1). In order to avoid catastrophic forgetting in the cortex, novel information is first stored in the hippocampus rapidly and progressively incorporated into cortical networks. For this organisation to be useful, the hippocampus must not suffer itself from catastrophic interference. The root of this problem is that a neural network employing dense distributed representations cannot acquire new information without forgetting old information unless new information is presented along old information. The complementary learning systems states that the cortex solves this problem by being sent old information by the hippocampus to avoid forgetting it. The hippocampus is thought to employ a different solution, that is to avoid using dense distributed representations. If memories are represented and stored orthogonally from each other, with little overlap, their separation makes them resilient to the storage of new memories.

As mentioned in Section 4.2.2, the activity of the dentate gyrus is very sparse in the sense that only a few granule cells are active at a time. Sparsity naturally limits the overlap between activity patterns. Patterns are further sparsified when transmitted to CA3, due to the very low probability of synaptic contact from granule cells. Consequently, the dentate gyrus plays of a pattern separator and forces a different representation in CA3 during learning in the standard framework.

In the framework, CA3 is the locus of episodic memory storage. More precisely, memories are thought to be stored in its recurrent collaterals. Like a Hopfield network (J. J. Hopfield 1982 ; Section 4.5.2), these recurrent connections can be modified to encode memory patterns. After storage, a pattern can be retrieved from a partial cue represented by the current state of CA3 neurons, once information has been transmitted through the recurrent synapses to modify the state of CA3. Recurrent collaterals are thus thought to encode autoassociations. In Marr (1971) and most future models, the synaptic modification of recurrent collaterals implementing the

associative memory follows a Hebbian rule, which is captured by Hebb's iconic saying "Cells that fire together, wire together", or more formally: "any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other" (Hebb 1949, p. 70).

The standard framework does not give CA1 a role as consensual as the dentate gyrus and CA3 (Cheng 2013).

4.5.2 Hopfield Network

The Hopfield network is the most well-known recurrent autoassociative memory network (J. J. Hopfield 1982) and its description provides a more complete understanding of the role of CA3 in the standard framework. The network is composed of thresholded units that are updated with the following rule:

$$s_i = \operatorname{sgn}(\sum_{j \neq i} W_{ji} s_j) \tag{4.1}$$

where s_i is the state of neuron i and W_{ji} is the strength of the connection from neuron j to i (typically $W_{ji} = W_{ij}$). The energy function of a Hopfield network associates a scalar value to each state and is defined as:

$$E = -\frac{1}{2} \sum_i \sum_{j \neq i} W_{ji} s_i s_j. \tag{4.2}$$

As s_i is updated, the energy function monotonically decreases:

$$\Delta E = -\Delta s_i \sum_{j \neq i} W_{ji} s_j. \tag{4.3}$$

In this formulation of Hopfield networks (i.e. asynchronous update, $W_{ii} = 0$ and symmetric weights), they are proven to converge to local minima of their energy function (Bruck 1990), so-called stable states (or attractors).

What has been described so far is very similar to the earlier spin glass model (Lenz 1920; Ising 1925). The distinguishing characteristic of Hopfield

network is the use of Hebb (1949)'s learning rule to implement an associative memory, that is the ability to learn the relationship between items, by manipulating attractors of the network. A Hebbian rule is used to modify synaptic weights of the Hopfield network when learning N binary patterns:

$$W_{ji} = \frac{1}{N} \sum_{\mu=1}^{N} \epsilon_i^\mu \epsilon_j^\mu \tag{4.4}$$

where ϵ_i^μ is bit i of pattern μ. The goal with this rule is to make patterns to be stored in a Hopfield network attractor of that network, with the aim of retrieving the patterns when altered versions (e.g. noisy or incomplete) are presented. Intuitively, this can be understood as the formation of attracting and repelling connections: when two bits tend to be positive and negative together ($\epsilon_i^\mu \epsilon_j^\mu = 1$), their neuron will learn to attract each other in state space and conversely, neurons with opposite states will learn to repel each other.

While Hopfield networks have greatly influenced the understanding of memory in both biological and artificial agents, their initial formulation suffers from serious limitations. First and perhaps most obvious is that the threshold activation function does not allow the storage of continuous patterns (but see the continuous generalization of J. J. Hopfield (1984)). Secondly, as there is no distinction between sensory and memory units, the number of patterns that can be stored is thus limited by their dimensionality d which is equal to the number of neurons. The storage capacity for retrieval of patterns free of errors is $C \approx \frac{d}{2\log_2 d}$ and approximately $0.14d$ with a small percentage of errors. The capacity is drastically decreased in the case of correlated patterns. Secondly, it is not clear how Hopfield networks can be adapted to model memory for non-static information. In principle, metastable states could emerge using a nonsymmetric Hebbian rule (allowing for weights to be nonsymmetric). Sequences of patterns could thus be retrieved using the original state update rule, each minimum being replaced in time by another. However, J. J. Hopfield (1982) reported failing to faithfully encode and retrieve sequences of four states and more.

4.5.3 CA3 as a Temporal Associator

The standard framework most often leaves aside the topic of sequence learning and consequently misses a central aspect of episodic memory. A strong body of evidence in fact indicates that the hippocampus is involved in rapid learning of sequential information (Agster, Fortin, et al. 2002; Fortin et al. 2002; Kesner et al. 2002; A. K. Lee and M. A. Wilson 2002; Ergorul and Howard Eichenbaum 2004) and can disambiguate overlapping sequences using splitter cells (Duvelle et al. 2023). Sequence memories are expressed during replays and prospective theta sequences. This cannot be accounted for by computational models that store static patterns of information.

However, a few models which are otherwise compatible with the standard framework have tackled the problem of sequence learning in the hippocampus (e.g. Levy 1982; Levy 1996; Wallenstein and Hasselmo 1997; Wallenstein, Hasselmo, and Howard Eichenbaum 1998). These models propose that CA3 does not only store patterns corresponding to the current representation of the environment, but that it is also capable of learning association between discontiguous events. The models of Levy (1982) and Levy (1996) exploit the sparse connectivity between the dentate gyrus and CA3 to prevent certain CA3 neurons to be clamped to external inputs and instead keep track of past states through recurrent collaterals. As a result, the global state of CA3 at each time point represents a mixture of environment inputs provided by dentate gyrus cells, and temporal context formed through recurrence. Recurrent CA3 connections are also constrained to be sparse in order to form attractors that are not stable through time and recall multiple items successively. Using CA3 plasticity mechanisms similar to those of other standard models but with sparse connections grants these models with sequence learning capabilities. Plasticity indeed stores patterns representing the conjunction of discontiguous items and sequences can be recalled with a process similar to standard models applied iteratively.

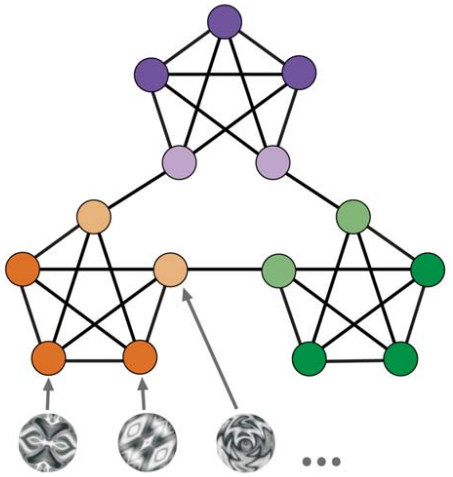
4.5.4 Evidence Against the Standard Framework

While very influential and accounting for some biological phenomena, the standard framework suffers from limitations. In fact, Hebb's learning rule which is a central tenet of the standard framework is computationally sub-optimal, as the small capacity of Hopfield networks suggests. Ketz et al. (2013) proposed a model in which the hippocampus does not learn using a Hebbian rule, but rather an error-driven mechanism scheduled by theta phases, like in the model of Hasselmo, Bodelón, et al. (2002). The authors indeed showed that the hippocampus is a much more powerful learner when this type of learning rule is used instead of the classical Hebbian mechanism. In a more recent version of the model (Y. Zheng et al. 2022), the entorhinal-to-CA3 and recurrent CA3 connections are also trained with an error-driven mechanism, further improving memory capacity and learning speed.

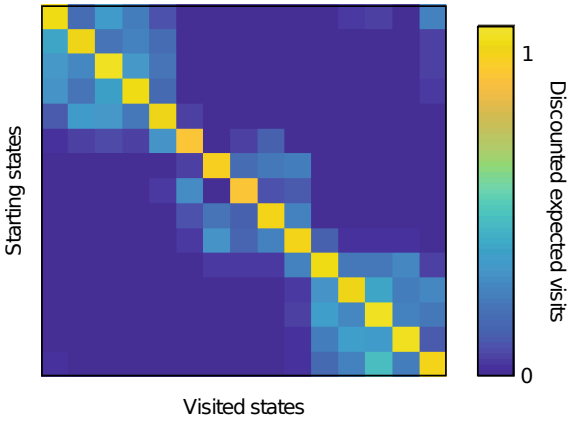
After its original proposal, the model of Ketz et al. (2013) has revealed deeper implications. The standard framework states that the engram (i.e. the substrate of memory) is located at the recurrent connections of CA3. This is at odds with the more consensual view that memory engrams are distributed. Studies suggest the hippocampus not only remembers individual episodes but is also involved in the discovery of statistical regularities among them (Mehta et al. 1997; Strange, Duggins, et al. 2005; Turk-Browne et al. 2009; Schapiro, Kustner, et al. 2012; Schapiro, Turk-Browne, Norman, et al. 2016). In the study of Schapiro, Turk-Browne, Norman, et al. (2016), human participants were shown abstract visual stimuli transitioning according to a community structured graph (Figure 4.7a). The activation of the hippocampus assessed by fMRI not only reflected individual transitions, but the higher order community structure too (Figure 4.7c). Activity evoked by stimuli belonging to a common community was more similar than activity evoked by stimuli of different communities. Schapiro, Turk-Browne, M. M. Botvinick, et al. (2017) investigated how can the hippocampus be involved in both the rapid acquisition of episodes composed of a sequence of transitions and a slower statistical form of learning highlighted by the community structure

task. They discovered that the model of Ketz et al. (2013) can in fact acquire both kinds of information while being initially designed to simulate episodic memory. Like in humans, higher order structure is captured by the model, such that the response to items of the same community is more similar than response patterns of different communities (Figure 4.7d). Most importantly, the structure is better captured by CA1 responses than by the dentate gyrus and CA3. Properties of the trisynaptic and monosynaptic pathways in the model, such as their respective fast and slow learning rate, constrain hippocampal regions to acquire information of different nature. These results suggest that memories are distributed throughout the hippocampus but do not rule out the possibility that true episodic memories (i.e. individual items and transitions) are stored in CA3. Schapiro, Turk-Browne, M. M. Botvinick, et al. (2017) proposed to extend the complementary learning systems with the monosynaptic pathway as an intermediate learning system between the truly episodic learner (trisynaptic pathway) and the truly statistician (cortex).

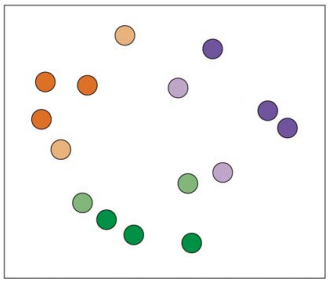
Cheng (2013) argued for the lack of evidence in favor of CA3 recurrent plasticity being necessary for rapid learning. Some studies did find deficits in that ability after experimental suppression of plasticity in CA3 (e.g. Nakazawa et al. 2003; Cravens et al. 2006) but these deficits were not always observed. In a Morris water maze experiment involving new platforms everyday Nakazawa et al. (2003), no deficit was found on days 5-12. In a short exposure contextual fear conditioning study Cravens et al. (2006), mutant mice performed as well as controls when tested after 24 hours. Since the deficit could not be assessed in all conditions, Cheng (2013) argued that CA3 plasticity is not necessary per se. Another study (Kishimoto et al. 2006) used the eyeblink conditioning paradigm and showed the hippocampus is necessary to perform the task, but CA3 plasticity is not necessary for acquisition, only for extinction (gradual decrease in conditioned response when the reinforcement is removed). Some studies in fact suggest that CA3 responses do not necessarily change significantly during learning (Dupret et al. 2010; Miyashita et al. 2009), in comparison with area CA1 whose activity is reorganized. Another prediction of the standard framework is that the



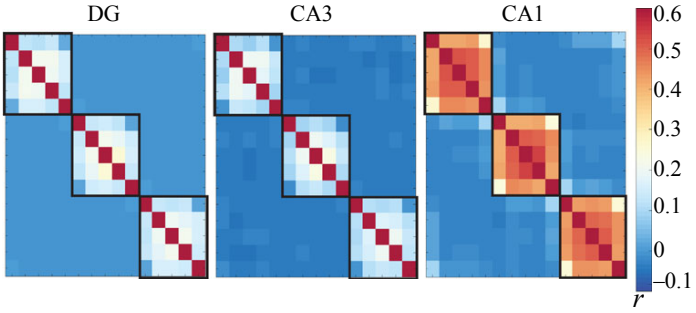
(a) Graph governing the transition between abstract visual stimuli. The graph is structured with three communities. In each community, only two nodes connect with nodes from other communities. Adapted from Schapiro, Turk-Browne, Norman, et al. (2016).



(b) Successor representation reflects the community structure in a simulation of the task. Each state corresponds to a node of the graph represented in Figure 4.7a. Adapted from Stachenfeld et al. (2017).



(c) Dimensionality reduction applied on the fMRI activation of the human hippocampus following the presentation of the visual stimuli depicted in Figure 4.7a. The representation of stimuli belonging to the same community are clustered together. Adapted from Schapiro, Turk-Browne, Norman, et al. (2016).



(d) Settled response of the model proposed by Ketz et al. (2013) in a simulation of the community structure task performed by Schapiro, Turk-Browne, M. M. Botvinick, et al. (2017). The symmetric heatmaps show the average representational similarity for each node pair of the graph depicted in Figure 4.7a. The diagonal represents the correlation between nodes with themselves. Adapted from Schapiro, Turk-Browne, M. M. Botvinick, et al. (2017).

Figure 4.7: Representation of stimuli whose transitions are governed by a community structure, in the human and simulated hippocampus.

storage of novel non-separated patterns in CA3 should interfer with older memories. The dentate gyrus has indeed been found to separate patterns and to be necessary for learning new associations that are similar to previously stored ones (P. E. Gilbert et al. 2001; McHugh et al. 2007). However, Cheng (2013) noted that trying to learn new similar associations without the dentate gyrus did not interfere with previous memories in these studies.

Studies also suggest that the role of the dentate gyrus is overlooked by the standard framework. If the dentate gyrus is a simple pattern separator that orthogonalizes the patterns to be stored in CA3, it should not be necessary for memory recall. In accordance with this proposal, studies that experimentally inactivate the dentate gyrus typically find no impairment in memory recall (e.g. Lassalle et al. 2000). However, inactivating the dentate gyrus did impair retrieval in some other studies, suggesting that memory traces are also located in that region (Denny et al. 2014; S. Park et al. 2016; Bernier et al. 2017). Furthermore, the dentate gyrus is reactivated during retrieval of recent (but not remote) memories (Tayler et al. 2013). In a recent review, Hainmueller and Bartos (2020) proposed that memories are stored in the EC-to-DG and DG-to-CA3 synapses at an early stage and that granule cells support the retrieval of CA3 activity patterns minutes to hours after the experience, while recurrent collaterals of CA3 may still be weak. After successive sharp wave ripple events involving newborn granule cells and consolidating CA3 synapses, memories can be retrieved in absence of the dentate gyrus.

The standard framework is appealing in its simplicity, but this growing body of evidence depicts a much more complex reality calling for alternative theories of learning in the hippocampus.

4.5.5 Alternative Theories

Socratic

A distinct model of the hippocampus was proposed with a focus on the learning of sequences of items (Lisman 1999). This model was called Socratic (i.e. Sequences Of Condensed Representations, Autocorrected, Theta-

Gamma Coded, In Context), and was used to account for early symptoms of schizophrenia (Lisman and Otmakhova 2001). In this model, the DG-CA3 loop stores the sequence of items sent by the entorhinal cortex at a theta-nested gamma rhythm. This model is at the origin of the theta-gamma code described in Section 4.4.2. Here, a later revised version of the model is described in which the respective functions of DG and CA3 were swapped (Lisman, Talamini, et al. 2005). This revised model is illustrated in Figure 4.8. One main idea behind this model is that the temporal separation of events happening in the environment is too big for plasticity to learn their associations and how they unfold through time. To resolve this issue which is central to episodic memory and its temporal nature, the entorhinal cortex is thought to act as a short-term memory buffer sending information to the hippocampus at the right cadence for long-term memory storage (O. Jensen and Lisman 2005; Schon et al. 2016). The stream of events is thus chunked, forming a discrete sequence code in which each item is represented in one gamma cycle (Figure 4.4). According to the revised Socratic model, items are sequentially sent to the hippocampus and CA3 recurrent collaterals learn autoassociations like in the standard framework. However, the CA3-to-DG backprojection is too slow to learn the autoassociation and instead the association between one pattern represented in CA3 and the next pattern represented in DG. This division of labor is thought to support robust sequence learning. While the CA3-to-DG heteroassociation process enables the retrieval of (a corrupted version of) the next item in the sequence, the autoassociation CA3 process is important for cleaning the result such that prediction errors do not accumulate (Figure 4.8, inset). As mentioned above, the initial proposal was that CA3 learns heteroassociations and DG learns to correct them but the exchange of their respective function was later suggested in alignment with the standard framework (Lisman, Talamini, et al. 2005).

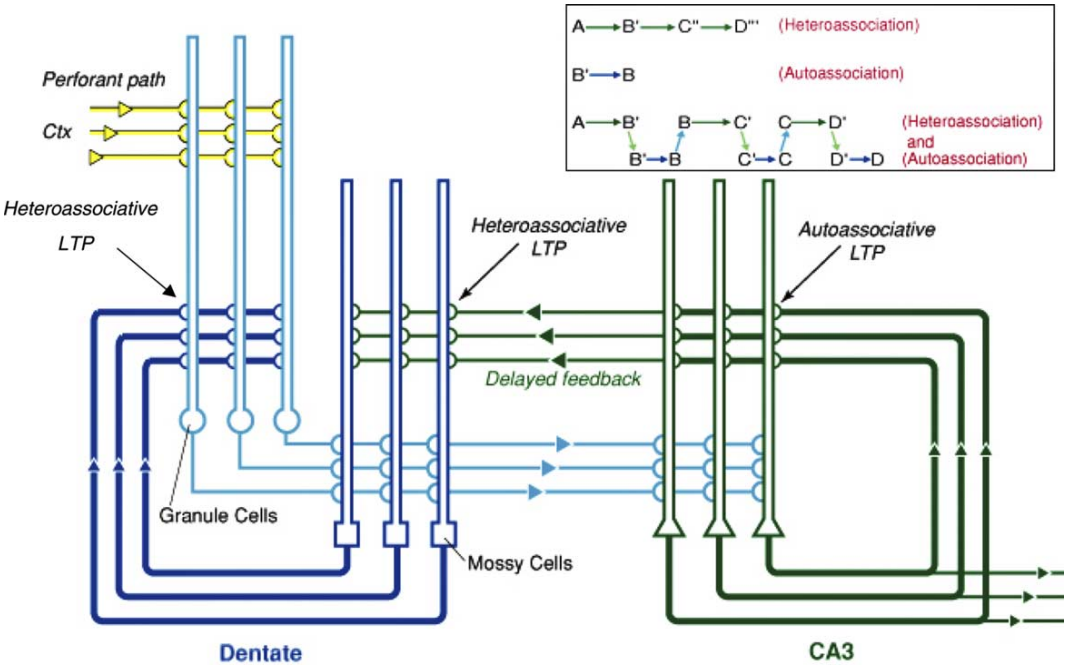


Figure 4.8: Revised Socratic model. Taken from Lisman, Talamini, et al. (2005).

CRISP

CRISP is another hippocampus model that focuses on sequence learning. It was proposed in Cheng (2013) after the review of many arguments against the standard framework. The main idea is that CA3 generates intrinsic sequences that do not necessarily evolve with learning and that sequences of events get mapped to these intrinsic sequences by fast modification of EC-to-CA3 and CA3-to-CA1 synapses. CA1 plays the role of pattern completion such that the monosynaptic pathway is sufficient to learn static patterns while the monosynaptic pathway is necessary to learn sequences. In that model, it is hypothesized that the dentate gyrus can reset the intrinsic CA3 sequence in order to learn new sequences and avoid interference between similar ones, but also to initiate recall.

Modern Hopfield Networks

In the recent years, there has been a resurgence of research related to associative memory. A family of novel algorithms, dubbed dense associative

memories or modern Hopfield networks, have been proposed to overcome the limitations of the original Hopfield network, as reviewed by Krotov (2023). They do not constitute a model of the hippocampus properly speaking, but their success provides valuable information to develop better hippocampal models.

Krotov and J. J. Hopfield (2016) proposed modifications in the energy function and dynamics of the network leading to a super-linear scaling relationship between pattern dimensionality and capacity. Demircigil et al. (2017) extended the idea of Krotov and J. J. Hopfield (2016) and proved storage capacity can even be exponential with an exponential activation function. Ramsauer et al. (2020) proposed another variant of modern Hopfield network based on previous work (Sukhbaatar et al. 2015) that instead makes use of the softmax function. Interestingly, the introduced update rule is equivalent to the self-attention mechanism of transformers (Vaswani et al. 2017):

$$\boldsymbol{z} = \boldsymbol{X}\text{softmax}(\beta \boldsymbol{X}^\top \boldsymbol{q}) \tag{4.5}$$

where $\boldsymbol{q}$ is the query vector, $\boldsymbol{X}$ is the memory matrix and β is an inverse temperature parameter controlling how much memories are separated. Softmax is the normalized exponential function:

$$\text{softmax}(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_j e^{z_j}} \tag{4.6}$$

Universal Hopfield networks were introduced as a general framework for associative memory models (Millidge et al. 2022). The idea is that classical and modern Hopfield networks can be decomposed into a similar series of operations: similarity, separation and projection (Figure 4.9). Retrieval from a universal Hopfield network is thus defined as:

$$\boldsymbol{z} = \boldsymbol{V} \cdot \text{sep}(\text{sim}(\boldsymbol{K}, \boldsymbol{q})), \tag{4.7}$$

where $\boldsymbol{q}$ is the query vector, $\boldsymbol{K}$ is the matrix of stored keys, *sim* is the similarity function, *sep* is the separation function and $\boldsymbol{V}$ is the matrix of stored

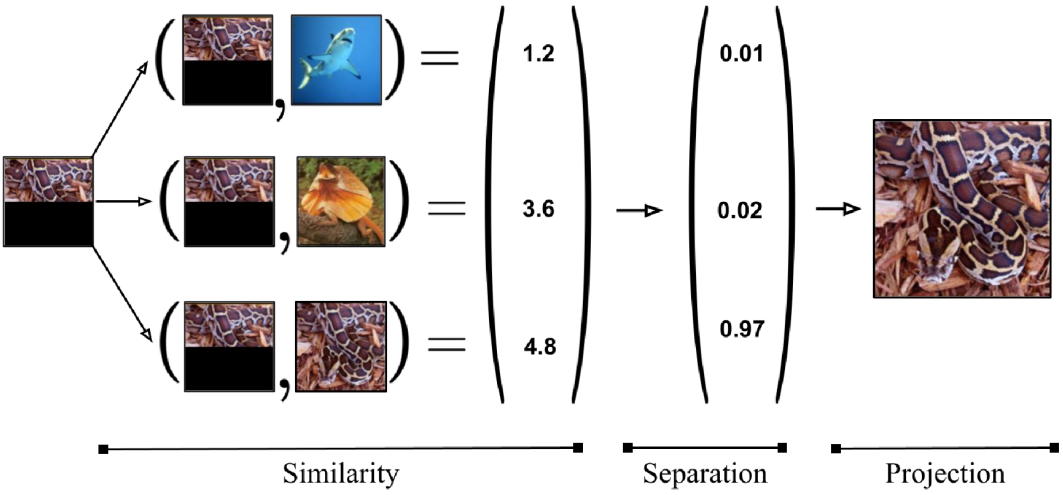


Figure 4.9: Universal Hopfield network operations. The input is compared to stored memories (similarity), the relative comparisons are then exaggerated (separation) and used to construct an output (projection). Taken from Millidge et al. (2022).

values. The concepts of key and value are used to account for heteroassociation where a key element is associated to a distinct value element, but in most applications of Hopfield networks the value is used to reconstruct the same key element (i.e. autoassociation with $\boldsymbol{V} = \boldsymbol{K}^\top$). In the case of the original Hopfield network, the similarity function is the dot product and the separation function is the simple identity function. The lack of a more elaborate separation function, like the softmax of the more modern model of Ramsauer et al. (2020), makes it impossible for the classical Hopfield network to properly separate memories.

The main advantage of modern Hopfield networks over classical ones is that they are not restricted to binary data and are easily integrated into deep learning architectures. Synaptic connections are then trained using the back-propagation algorithm (Linnainmaa 1976; Rumelhart et al. 1986). It is not yet clear to what extent modern Hopfield networks can be trained under more restrictive regimes without impairing their storage capacity. A strength of the classical Hopfield network is its biological plausibility: both the interactions between units and the learning of information are simple enough to be implemented by real neurons. On top of that, information can

be presented sequentially and memorized after a single presentation. On the other hand, Krotov and J. Hopfield (2020) noted how their modern variants involve many-unit interactions that blur the parallel between their connections and biological synapses that typically involve two neurons only. They proposed to replace many-unit interactions by a network of feature neurons and memory neurons that could be implemented as a traditional artificial neural network, which was later demonstrated by Snow and Orchard (2022). Note, however, that the storage capacity becomes linear in the number of memory neurons (Krotov and J. Hopfield 2020).

The use of back-propagation further undermines the plausibility of modern Hopfield networks and their ability to rapidly memorize novel information. Back-propagation is typically employed under a training regime in which information is presented either many times sequentially or as a single batch (as done by Snow and Orchard 2022). Both cases do not align with the requirements of episodic memory which should rapidly acquire new information presented sequentially. In some works, the patterns to memorize are instead added to the weight matrices manually (e.g. Millidge et al. 2022; Spens and Burgess 2023). Recently, Tyulmankov et al. (2021) proposed a learning rule for storing information in a more plausible way. In order to encode a memory, a neuron in the separation layer (i.e. the memory neurons of Krotov and J. Hopfield 2020) is selected for activation, either randomly or through a sequential process. The input pattern is then associated to the activated neuron using a neohebbian three-factor learning rule (Gerstner et al. 2018), such that the memory is stored in the row of the key matrix corresponding to the selected neuron. A hebbian rule is then used to store the pattern in the column of the value matrix corresponding to the neuron. When using this biologically-driven approach with a sequential selection process, or when patterns are manually added to the matrices, the storage capacity is equal to the number of neurons in the separation layer (Tyulmankov et al. 2021). When selecting a separation neuron randomly, the capacity is sublinear and comparable to the capacity of traditional Hopfield networks. Tyulmankov et al. (2021) also demonstrated that modern Hopfield networks can learn sequences by sending the current pattern for key stor-

age and the next pattern for value storage. The network can then be used to recall sequences by recurrently using the output as the next input.

Much like the way non-recurrent transformers (Vaswani et al. 2017) led to the development of performant language models that are now compared to language processing in biological brains (e.g. Caucheteux and King 2022), the modern feedforward variants of Hopfield networks will certainly influence the understanding of memory storage in the hippocampus. In fact, it has already been demonstrated that the popular model of Whittington, T. H. Muller, et al. (2020) is performant with no recurrence and that incorporating elements of transformers improves its performance (Whittington, Warren, et al. 2021). Furthermore, it has been proposed that the hippocampus implements a modern associative memory in CA3 and CA1 neurons, although the precise anatomical mapping is not clear (Krotov and J. Hopfield 2020; Chandra et al. 2023; Spens and Burgess 2023). Their sequence learning ability having been established (Chaudhry et al. 2023; Tyulmankov et al. 2021; Millidge et al. 2022), modern Hopfield networks could be used to model sequence processing in the hippocampus.

Successor Representation

The successor representation was introduced by Dayan (1993) in the context of reinforcement learning, an artificial intelligence paradigm focused on developping agents that can learn to choose actions for maximizing their rewards. Later on, it has influenced theories of learning in the hippocampus. Using the formulation of Stachenfeld et al. (2017), one can define[3] the value V of a state s as the expected sum of rewards R obtained in future states s_t :

$$V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) | s_0 = s\right] \tag{4.8}$$

[3]This is well defined in a Markov process, that is when the transition between states does not depend on past states.

where $\gamma \in [0, 1]$ is an exponentially decaying discount factor. Dayan (1993) factorized this value function into the inner product of the reward function and a matrix M predictive of successor states:

$$V(s) = \sum_{s'} M(s, s')R(s'). \tag{4.9}$$

This successor representation matrix is the expected future occupancy of state s′:

$$M(s, s') = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \delta(s_t, s') | s_0 = s\right], \tag{4.10}$$

where $\delta(s_t, s')$ is the Kronecker delta function whose value is 1 if $s_t = s'$ and 0 otherwise.

It is possible to learn this predictive matrix using an error-driven temporal difference algorithm that incrementally updates the estimated successor representation to minimize the discrepancy between expected and observed state occupancy (Dayan 1993; White 1996). It was later argued that this algorithm can be used to encode episodic memories and reproduce well-known psychological effects (Gershman, C. D. Moore, et al. 2012). The authors made a connection between their model and the temporal context model, the idea of which is to learn the association between individual items and a vector keeping track of the recent context (Howard and Kahana 2002). More recently, Stachenfeld et al. (2017) showed that when states encode space and a column of the successor representation matrix is interpreted as a neuron, the activation of this neuron resembles a place field, giving rise to the theory that hippocampal neurons encode such a prediction of future states. The cognitive map would thus be a predictive map. Since the matrix is influenced by the behavior of the agent and by transitions between states, these place fields can exhibit the experience-dependent asymmetric backward expansion described by Mehta et al. (1997) and the barrier effect of Alvernhe et al. (2011). Successor representation can also account for the activity of the hippocampus in the task of Schapiro, Turk-Browne, Norman, et al. (2016), as shown in Figure b. Finally, Stachenfeld et al. (2017) found that

the eigenvectors of the matrix resembled the activity of grid cells in various situations.

Momennejad and Howard (2018) used multiple γ values for multi-scale representation, drawing inspiration from the gradient of representational scale along the longitudinal axis of the entorhinal cortex and the hippocampus. They showed that this representation can be used for both reconstructing sequences of experienced states and predicting the distance between states.

While being a promising normative theory, it is not clear how the hippocampus learns such a predictive map, but recent modeling work suggests that theta sequences can support the learning of successor representation (George, Cothi, Stachenfeld, et al. 2023).

4.6 Discussion

In this chapter, the anatomy, representations and dynamics of the medial temporal lobe have been presented in terms of episodic memory operations. The computational model developed in Chapter 8 will be designed to follow the same architecture. This new model constitutes a viable alternative to the standard framework for sequence learning, and hippocampal-like representations will be shown to emerge.

The present chapter presents a complex image of the hippocampus, with intertwined processing modes and functions. Although it may function partially autonomously, the next chapter describes how the hippocampus is also controlled by executive regions such as the prefrontal cortex.

5 Episodic Control and Controlled Episodic Memory

An agent remembering an episodic memory and controlling which action to take accordingly

Generative AI

5.1 Introduction

An influential view is that cognition is centered around action, in that it has evolved as a way of selecting the appropriate motor commands[1]. While the level of appropriateness of an action might seem difficult to formalize at first glance, the gene-centered view of evolution (Trivers and Dawkins 1976) provides an objective criterion: an action is deemed appropriate when the level of propagation of one's genes is increased as a consequence of that action. The action-centric view boils down to the proposal that cognition mainly aims to compute that criterion and execute the actions that are most appropriate.[2] The four Fs are often given as primal examples of evolutionary adapted behaviors: *fighting*, *fleeing*, *feeding* and *mating*, but higher-level cognition allows for much richer behaviors with indirect benefits (e.g. language and social interactions). This appeals to the behaviorist notions of primary reinforcers (e.g. food, water, pleasure, sleep), which probably act as an innate proxy for the evolutionary advantage criterion, and secondary reinforcers (e.g. money) which are associated with primary ones, as captured by the model of O'Reilly, M. J. Frank, et al. (2007).

A direct consequence of the action-centered view of cognition is that cognitive functions are instrumental to action selection. This has mostly been studied in the context of perception, and there is a lot of evidence that perception is indeed affected by the ability to act. For example, softball players with better batting performance tend to perceive the ball as bigger than players with poorer performance (Witt and Proffitt 2005). In the same vein, tennis players tend to perceive the ball as slower when they successfully return it (Witt and Sugovic 2010). Crucially, only intended actions appear to have a modulatory effect on perception (Witt, Proffitt, and Epstein 2010). These thrilling results have led to the development of action-specific theories of perception (Witt 2011; Proffitt 2006) which radically depart from

[1] Teleological phrasing is used for convenience throughout the text without implying evolution is goal-directed.

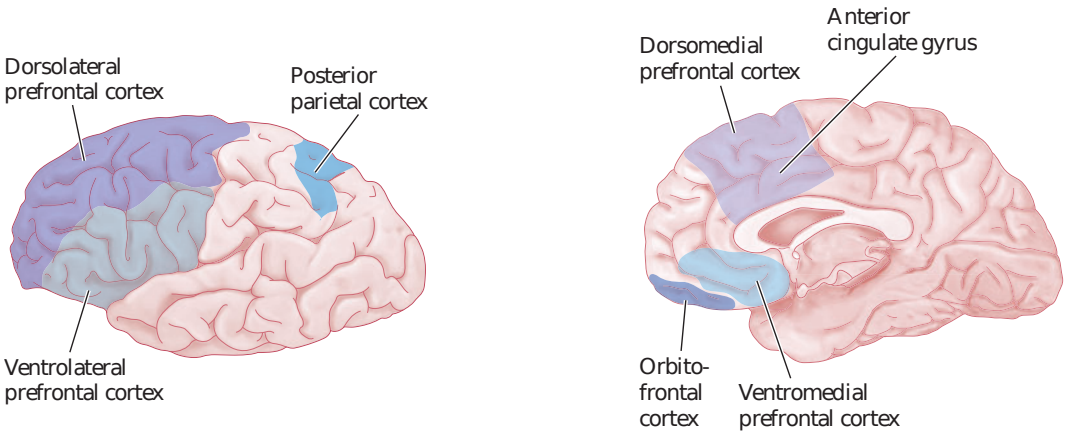
[2] Since the criterion can be very hard to compute, this does not imply that all actions selected by cognitive functions are appropriate.

the early conception that perception is a passive process, the function of which is to depict reality objectively.

Similarly, it is likely that memory has evolved to inform behavioral control. Memory of past experience can indeed provide crucial information about the potential consequences of one's actions and can therefore be used to make better decisions. Memory systems thus have an incentive to encode, store and retrieve informative memories. In this chapter, the goal-directed use of episodic memory is discussed. First, the anatomy of the frontal lobe and connections with the hippocampus will be described in Section 5.2. The involvement of the prefrontal cortex in episodic memory will then be described succinctly (Section 5.3). In Section 5.4, hippocampal representations will be considered in terms of context formation for the recall of relevant memories that can be sent to the frontal cortex to guide decision-making. In Section 5.5, the dynamics of prefrontal-hippocampal interactions will be discussed, and it will be shown that the prefrontal cortex is not only informed by the hippocampus but also modulates memory operations actively. Finally, Section 5.6 presents computational models related to the interactions between memory and control.

5.2 Anatomy

The frontal lobe is a key structure of motor and cognitive control (Goldman-Rakic 1996). It comprises the prefrontal cortex (Figure 5.1), the premotor cortex and the primary motor cortex. It is hierarchically organized along a rostrocaudal axis, such that more rostral regions represent successively more abstract task information (Koechlin et al. 2003; Badre and D'Esposito 2007; Nee and J. W. Brown 2012; J. R. Reynolds et al. 2012). At the lowest level is the premotor cortex which learns the association between sensory stimuli and motor responses with the recruitment of the basal ganglia (i.e. sensorimotor rules; Weinrich et al. 1984; Brasted and Wise 2004). Higher levels (more rostral regions) lie within the prefrontal cortex and learn to modulate lower levels (more caudal regions) according to more abstract rules. A



(a) Lateral structures.

(b) Midline and inferior structures. The orbitofrontal cortex continues to the ventral surface of the frontal lobes (not shown).

Figure 5.1: Major brain regions supporting cognitive control. Taken from Purves et al. (2013).

detailed review of the anatomy of the frontal cortex is outside the scope of this thesis (see Dagar 2023 for an extensive review on the PFC and cognitive control), but a broad description of the medial and lateral divisions of the prefrontal cortex is first provided to better detail its interactions with the hippocampus.

The medial prefrontal cortex (mPFC) comprises the ventromedial PFC (vmPFC), the orbital PFC (OFC) and the dorsomedial PFC (dmPFC). The dmPFC itself comprises the anterior cingulate cortex (ACC). As reviewed by Dagar (2023), the mPFC guides the process of decision making (Kennerley et al. 2011) by evaluating the value of choices (Jocham et al. 2011; Lim et al. 2011) and taking motivation into account (Kouneiher et al. 2009; Holroyd and Yeung 2012). It monitors outcome and error (M. Rushworth et al. 2004; M. F. Rushworth et al. 2007) and the value of the course of action pursued (Hayden and Platt 2010; Noonan et al. 2011). The mPFC is thus thought to monitor behavioral plans by computing prediction errors and initiate rule switching when deemed appropriate.

The lateral prefrontal cortex is divided into dorsal (dlPFC) and ventral (vlPFC) regions. The lateral PFC is thought to support working mem-

ory as suggested by lesion and inactivation studies (M. H. Miller and Orbach 1972; Bauer and Fuster 1976; Funahashi et al. 1993). Neurons of the dlPFC in fact show sustained activity during working memory tasks (Fuster and G. E. Alexander 1971; Kubota and Niki 1971; Niki and Watanabe 1979; Funahashi et al. 1989). The lateral PFC is thus thought to be responsible for maintaining a representation of the current plan and to support the learning of rules that require short-term memory.

In order for the hippocampus to guide cognitive control and for the prefrontal cortex to modulate memory operations, these structures must exchange information through anatomical circuits. Both the hippocampus and the prefrontal cortex are extensively connected to cortical and subcortical regions in such a way that these two regions can indirectly interact in many ways. There are also direct pathways connecting them, most studies focusing on connections with the medial prefrontal cortex (Howard Eichenbaum 2017b). Interactions between the hippocampus and the lateral prefrontal cortex are likely to be mediated by strong bidirectional connections between the lateral and medial prefrontal cortices (Anderson et al. 2016). The most direct hippocampus-to-prefrontal cortex pathway is the monosynaptic projection from CA1 of the ventral to intermediate hippocampus (taking the rodent longitudinal axis as a reference) to the mPFC, including the OFC (Jay, Glowinski, et al. 1989; Jay and Menno P Witter 1991; Hoover and Vertes 2007). More ventral portions of the hippocampus are thought to represent broad information, as exemplified by the big scale of ventral place fields reviewed in the previous chapter. This pathway is thus well suited to provide contextual information to the prefrontal cortex.

A more indirect pathway involves the thalamic nucleus reuniens (Dolleman-Van Der Weel and M. Witter 1996; Vertes 2002; Vertes 2006; Vertes, Hoover, et al. 2007; Cassel et al. 2013; Dolleman-van der Weel et al. 2019). This structure bidirectionally connects CA1 (all portions) and all medial prefrontal areas. It also communicates with the perirhinal cortex and the entorhinal cortex. A lesser known subcortical route involves the small supramammillary nucleus of the hypothalamus, which receives weak projections from the medial prefrontal cortex (F. Wouterlood et al. 1987; Allen and Hopkins 1989; Sesack

et al. 1989; Gonzalo-Ruiz et al. 1992) and projects to the dentate gyrus and area CA3 of the hippocampus (Pasquier and Reinoso-Suarez 1976; Wyss et al. 1979; Vertes and McKenna 2000; M. Li et al. 2023).

The medial prefrontal cortex is also directly and reciprocally connected to the perirhinal and entorhinal cortices (Burwell and Amaral 1998; Menno P Witter, Floris G Wouterlood, et al. 2000; Apergis-Schoute et al. 2006; Agster and Burwell 2009). The mPFC projects to superficial layers of the perirhinal cortex and deep layers of the lateral entorhinal cortex (Burwell and Amaral 1998). It in turn receives projections from these regions that terminate in its layers I, II and VI (Agster and Burwell 2009). Medial prefrontal projections to the entorhinal cortex mainly terminate in the lateral compartment (Burwell and Amaral 1998) which, as reviewed in the previous chapter, connects the hippocampus to cortical regions specialized in semantic processing. This suggests that this pathway modulates semantic information to a larger extent than spatial information.

Finally, it has been reported by Rajasethupathy et al. (2015) and Bian et al. (2019) that ACC neurons project (very sparsely) to CA3 and CA1 neurons. These findings contrast with the previous consensus and have important functional implications since they potentially highlight a straightforward way for the prefrontal cortex to modulate hippocampal activity. Nevertheless, Andrianova et al. (2023) employed multiple complementary methods to validate this discovery and found no evidence of such a direct monosynaptic projection from ACC to the hippocampus.

As reviewed here, the hippocampus and the prefrontal cortex can communicate in many ways. However, the precise pathways underlying this bidirectional communication are difficult to establish due to the involvement of relay structures and inconsistencies between studies. The studies described in the following sections employ a combination of behavioral, functional connectivity and selective lesion analyses that point towards a model of prefrontal-hippocampal interactions.

5.3 Contributions of the Prefrontal Cortex to Memory

Before detailing prefrontal-hippocampal interactions, the prefrontal contribution to memory is introduced. Despite early reports indicating that patient K.M., who had prefrontal damage, had no memory deficits, it is now well established that the prefrontal cortex plays an important role in memory (Janowsky et al. 1989; Moscovitch 1992; Shimamura et al. 1995; Schnider and Ptak 1999; E. K. Miller and J. D. Cohen 2001; Dobbins et al. 2002; Preston and Howard Eichenbaum 2013; Szczepanski and Knight 2014; Howard Eichenbaum 2017b; S. A. Park et al. 2023). The prefrontal cortex is indeed thought to deploy top-down control over memory processes. Just as it uses attentional mechanisms to filter out distracting stimuli in the environment, the prefrontal cortex selects which memories are relevant to reach current goals. Many studies indeed show that prefrontal damage leads to intrusions of irrelevant memories (Schnider and Ptak 1999; Shimamura et al. 1995; Farovik et al. 2008; Chudasama et al. 2012; Peters et al. 2013; Giustino and Maren 2015; Anderson et al. 2016). Evidence from rodent and primate studies suggest that this retrieval control ability is either mediated by connections from the prefrontal cortex to the entorhinal cortex controlling the input and output of the hippocampus, or by the nucleus reuniens which could convey modulatory information directly to the hippocampus (Anderson et al. 2016). The prefrontal cortex is also thought to control memory encoding by inhibiting the hippocampus when new information is consistent with pre-existing knowledge to avoid encoding redundant information (Frankland and Bontempi 2005; Van Kesteren et al. 2012).

5.4 Representation: the Prefrontal-Hippocampal System as a Contextual Memory Map

In the previous chapter, splitter cells have been introduced as neurons not only coding for where the animal is located, but also for where it comes

from (retrospective splitter cells) or where it is about to go (prospective splitter cells). One might wonder how the hippocampus has access to the latter information. The dentate gyrus and CA3 circuitry is known to underlie CA1 retrospective coding (Keinath et al. 2020), but its contribution to prospective coding is unknown (Duvelle et al. 2023). Trajectory-dependent activity can be found in the medial entorhinal cortex but lesion studies suggest that it is not necessary for CA1 activity to be splitted (Sabariego et al. 2019). Splitter cells can also be found in the prefrontal cortex (Baeg et al. 2003; Young and Shapiro 2011; Cowen et al. 2012) and nucleus reuniens (Ito, Zhang, et al. 2015; Ito, E. I. Moser, et al. 2018). Inactivating the medial prefrontal cortex (Guise and Shapiro 2017) or the nucleus reuniens (Ito, Zhang, et al. 2015) both impact splitter coding in CA1, and this impairment seems to be specific to prospective splitter cells (Ito, Zhang, et al. 2015). It was also shown by Ito, E. I. Moser, et al. (2018) that silencing the supramammillary nucleus reduces splitter coding in CA1 and the nucleus reuniens, but not in the prefrontal cortex. Hence, a plausible view is that the retrospective splitting signal is generated internally while the prospective signal is generated in executive structures like the prefrontal cortex and transmitted to the hippocampus through relay structures such as the nucleus reuniens of the thalamus. An alternative theory is that the prospective splitting signal is conveyed by the entorhinal cortex, as this structure is thought to mediate prefrontal control over memory processes, and that the nucleus reuniens orchestrates this communication (Section 5.5; Howard Eichenbaum 2017b).

The hippocampus can thus be seen as representing a temporal and goal-directed context map for retrieving memories that are most appropriate and can best inform executive structures like the prefrontal cortex. It indeed takes into account the recent past through an internally-generated retrospective splitting signal, the current sensory observation provided by the lateral entorhinal cortex, the current structural (e.g. spatial) information provided by the medial counterpart, and the ongoing plan through the prospective splitting signal provided by the prefrontal cortex. This relates to (and perhaps extends with goal-directedness) the interpretation of hip-

pocampal processing as latent state inference (e.g. Fuhs and Touretzky 2007; Whittington, T. H. Muller, et al. 2020) and echoes to the idea of remapping.

5.5 Dynamics: Oscillations as Memory Interactions

The prefrontal cortex and hippocampus interact mainly through ipsilateral (i.e. same side) oscillatory synchrony, as reviewed by Howard Eichenbaum (2017b). Some prefrontal neurons are indeed synchronized to the hippocampal theta rhythm in various tasks (Siapas et al. 2005; Jones and M. A. Wilson 2005). The strength of synchronization is higher in prefrontal neurons coding for spatial information (Hyman et al. 2005), and is increased in successful spatial working memory trials (Jones and M. A. Wilson 2005; Hyman et al. 2010), specifically when the animal is at a choice point (Jones and M. A. Wilson 2005; Benchenane et al. 2010). Increased synchronization has also been found in humans successfully performing a transitive inference task (Backus et al. 2016). Directional flow analyses reveal that information flows from the ventral hippocampus to the prefrontal cortex when a memory is retrieved and maintained (Siapas et al. 2005; O'Neill et al. 2013; Hallock et al. 2016; Place et al. 2016), and from the prefrontal cortex to the hippocampus when memory retrieval is demanded (Hallock et al. 2016; Place et al. 2016). The phase shifts between oscillations in these two regions are on the order of 30ms, exceeding the delay for a monosynaptic projection, which suggests that this synchrony is either mediated by polysynaptic pathways or supports the transfer of information with segments of about 30ms (approximately a gamma cycle), as proposed by Howard Eichenbaum (2017b).

Spellman et al. (2015) inactivated projections from the ventral hippocampus to the medial prefrontal cortex in mice during a spatial working memory task. Their manipulation impaired performance and spatial coding in the prefrontal cortex. A directional flow analysis of theta-gamma oscillations indicated that hippocampus-to-prefrontal functional connectivity was disrupted, but not reciprocal connectivity. This study indicated that the monosynaptic connection arising from CA1 conveys hippocampal rep-

resentations to the prefrontal cortex and is necessary for spatial working memory. Sharp-wave ripples, which support spatial working memory too (Girardeau et al. 2009; Ego-Stengel and M. A. Wilson 2010; Jadhav, Kemere, et al. 2012; Fernández-Ruiz, Oliva, Fermino de Oliveira, et al. 2019; Gridchyn et al. 2020) are also thought to mediate prefrontal-hippocampal communication (Zielinski et al. 2020). Medial prefrontal neurons are indeed modulated during ripple events (Peyrache et al. 2009; Benchenane et al. 2010; Wang and Ikemoto 2016; Jadhav, Rothschild, et al. 2016; Tang et al. 2017; Berners-Lee et al. 2021), especially while awake (Tang et al. 2017). Crucially, the content of hippocampal replay can be decoded from prefrontal activity (Berners-Lee et al. 2021), and inhibiting the medial prefrontal cortex during ripple events disrupts spatial alternation learning (Bakker et al. 2023). While online replays provide hippocampal memory content as context to the prefrontal cortex, offline replays seem to train the prefrontal cortex like any other cortical region involved in the complementary learning systems theory (Peyrache et al. 2009; Benchenane et al. 2010).

Other studies investigated the role of reciprocal connections passing through the nucleus reuniens and found that they are crucial for prefrontal-hippocampal synchrony and memory specificity (Xu and Südhof 2013; Ito, Zhang, et al. 2015; Hallock et al. 2016). Xu and Südhof (2013) found that inactivating the medial prefrontal cortex or the prefrontal-to-reuniens pathway caused an overgeneralization in a contextual fear acquisition task, suggesting that the hippocampus then lacked top-down information. Optogenetic stimulation of the nucleus reuniens caused increased freezing in novel contexts. However, as noted by Howard Eichenbaum (2017b), these broad artificial stimulations probably disrupted the modulatory role of the nucleus reuniens rather than disrupting the direct exchange of information between the thalamus and the hippocampus. In fact, unlike neurons of the medial prefrontal cortex and the hippocampus, nucleus reuniens neurons do not encode detailed information about spatial trajectories (Ito, Zhang, et al. 2015).

If the nucleus reuniens does not directly convey information from the prefrontal cortex to the hippocampus, the question remains as to how top-down signals enabling memory control and prospective coding are con-

veyed. Despite some reports (Rajasethupathy et al. 2015; Bian et al. 2019), it is very unlikely that a monosynaptic projection is involved (Andrianova et al. 2023). Studies suggest that ipsilateral prefrontal-to-entorhinal connections play a direct role (Navawongse and Howard Eichenbaum 2013; Jayachandran et al. 2019). Disconnecting this pathway indeed causes dorsal CA1 neurons to lose their object specificity without affecting their spatial specificity (Navawongse and Howard Eichenbaum 2013). After reviewing many of the results presented here, Howard Eichenbaum (2017b) proposed a model of prefrontal-hippocampal interactions for episodic memory (Figure 5.2), in which he proposed that these structures interact bidirectionally, and that the nucleus reuniens synchronizes them and controls in which direction information flows. When contextual cues are presented, ventral hippocampal information is sent to the medial prefrontal cortex to provide context to executive functions. Subsequently, when context-guided memory retrieval is required, the prefrontal cortex biases retrieval of specific memories in the dorsal hippocampus through the lateral entorhinal and perirhinal cortices. This simple model accounts for many empirical findings and for why prefrontal projections do not terminate on the same end of the longitudinal axis as where hippocampal projections originate: dorsal hippocampal neurons encode memory specifics while more ventral neurons represent broader contextual information. However, this model might underestimate the role of the nucleus reuniens. In a recent study, (Jayachandran et al. 2019) analyzed the selective contribution of the perirhinal and nucleus reuniens pathways in a contextual sequence memory task. Both pathways were necessary for rats to differentiate correct in sequence odors from incorrect out of sequence ones. In some conditions, silencing the reuniens pathway impaired performance more than silencing the perirhinal pathway. More strikingly, the authors found that the two pathways supported different top-down control strategies. These results are at odds with the proposal of Howard Eichenbaum (2017b) that nucleus reuniens is just an enabling structure of the perirhinal and entorhinal transfer of prefrontal information to the hippocampus. However, it is possible that the entorhinal pathway, which was not specifically silenced in the study, is indeed con-

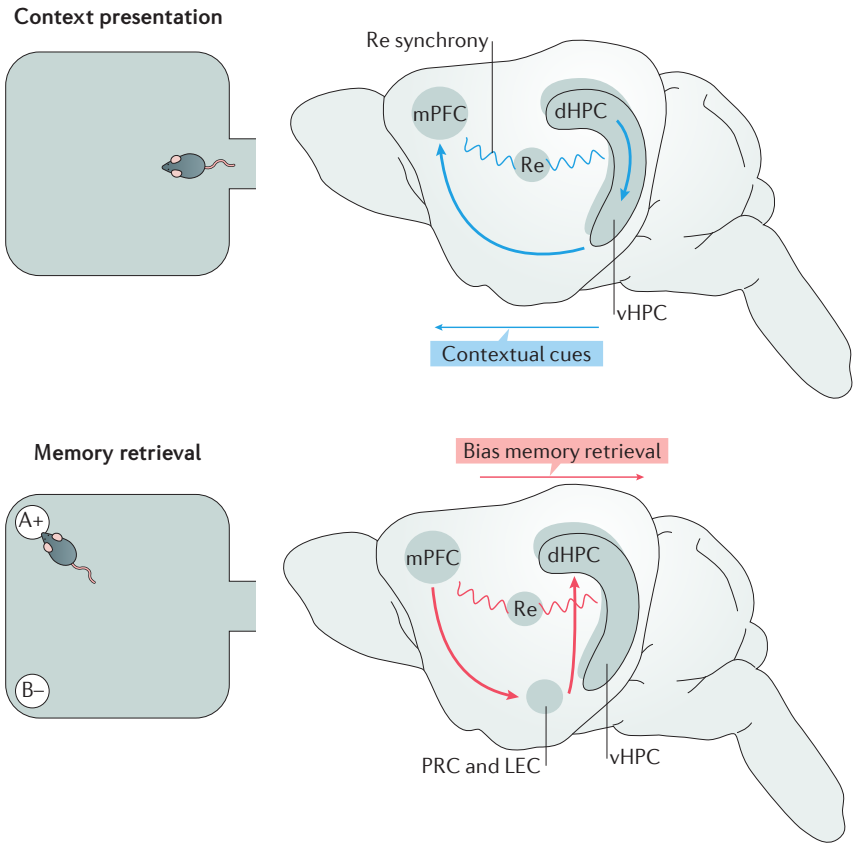


Figure 5.2: Model of prefrontal-hippocampal interactions in episodic memory. Adapted from Howard Eichenbaum (2017b).

trolled by the nucleus reuniens, and that the perirhinal pathway operates more autonomously. The role of the supramammillary nucleus, which also connects to the prefrontal cortex and the hippocampus, also remains to be determined in this model.

5.6 Computational Mechanisms and Models

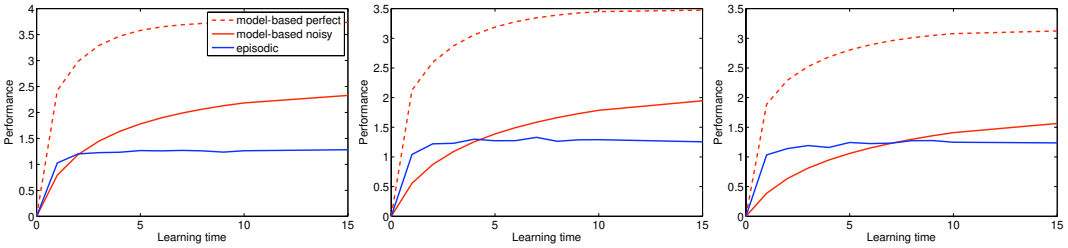
5.6.1 Experience Replay

Experience replay can be thought of as the reinforcement learning equivalent of the semantic replay postulated by the complementary learning systems theory (Section 3.3.1). Beyond specific implementations, what is

referred to as experience replay here is the general machine learning approach of storing and replaying past experiences to speed up the acquisition process of reinforcement learning agents (Lin 1992; A. W. Moore and Atkeson 1993; Sutton and Barto 2014; Mnih et al. 2015; Schaul et al. 2015; Mattar and Daw 2018; Cazé et al. 2018). Experience replay is at least partially inspired by hippocampal replays, which are known to play a role in reinforcement learning (De Lavilléon et al. 2015; Girardeau et al. 2009). Trajectories generated with some experience replay methods are in fact comparable to hippocampal reactivations (Mattar and Daw 2018; Cazé et al. 2018).

5.6.2 Episodic Memory as a Third Way of Control

Reinforcement learning algorithms are classically divided into two categories. Model-free reinforcement learning algorithms extract the value of individual states or actions from many experiences, by bootstrapping (compiling) information about which states/actions follow individual state/action (i.e. transition dynamics) and how rewarded these are. On the other hand, in model-based algorithms, transition dynamics and rewards are learned separately and only combined during inference, making these algorithms more flexible but more computationally expensive. The hippocampus and episodic memory are thought to contribute to control in a third way (Hasselmo and Howard Eichenbaum 2005; Lengyel and Dayan 2007; Gershman and Daw 2017; Cazin 2018; M. Botvinick et al. 2019). Instead of learning by bootstrapping (model-free) or performing expensive inference (model-based), the episodic controller of Lengyel and Dayan (2007) explicitly stores and retrieves experiences leading to unexpectedly high rewards. The authors showed that while such a controller is eventually suboptimal, it can learn tasks more rapidly than a model-based controller, especially when the task is difficult and the model-based controller is noisy (Figure 5.3). Unlike offline experience replay, episodic control thus combines episodes at inference time to compensate the rigidity of model-free bootstrap and the cost of model-based inference. Contrary to classical reinforcement learning algorithms, episodic control is nonparametric in that experiences are stored

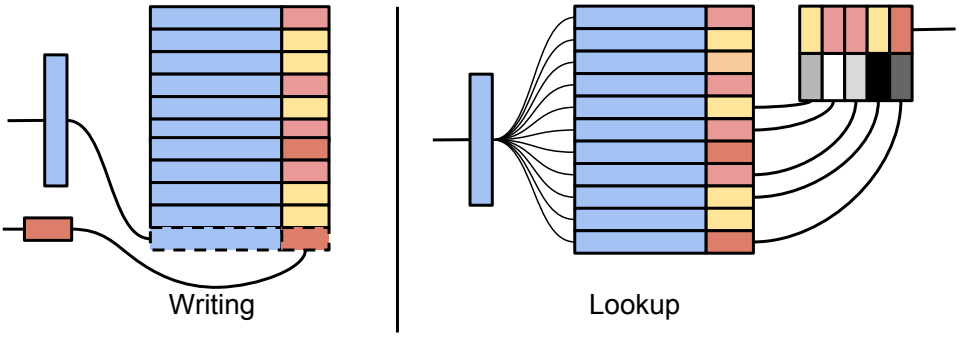


(a) $B = 2$ (b) $B = 3$ (c) $B = 4$

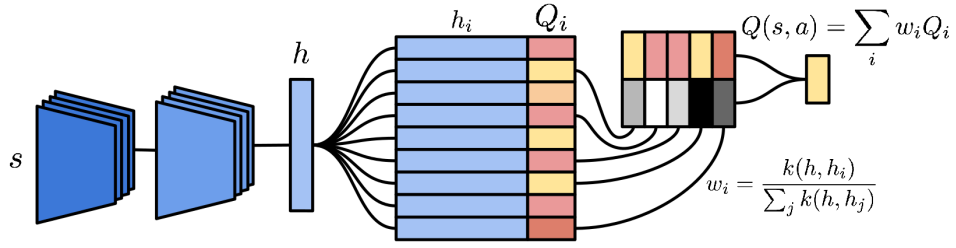
Figure 5.3: Performance of model-based and episodic controllers in a tree-structured Markov decision problem with varying complexity captured by the branching factor B. The noisy model-based controller only has access to noisy versions of the action values it learns. This was done to model the effects of imperfect working memory. Taken from Lengyel and Dayan (2007).

explicitly instead of being transformed and stored as parameters of a transition model (model-based) or value estimates (model-free). In the complementary learning systems framework, this corresponds to experiences being stored separately from each other in episodic memory, as compared to the way semantic memory blends experiences into coherent schemas.

This work has led to the development of differentiable memory-augmented controllers that can be integrated into deep learning architectures (Graves, Wayne, and Danihelka 2014; Graves, Wayne, M. Reynolds, et al. 2016; Blundell et al. 2016; Pritzel et al. 2017; Ritter et al. 2018; Wayne et al. 2018; Fortunato et al. 2019; Lu et al. 2022). Pritzel et al. (2017) for example introduced Neural Episodic Control (NEC), a deep reinforcement learning agent that employs a differentiable neural dictionary (DND) to store state-action-value tuples and act according to retrieval from this DND. More precisely, DND has a memory module $M_a = (K_a, V_a)$ for each action a, where K matrices store experienced states as keys and V matrices store the estimated values of these states. Operations of the DND are illustrated in Figure 5.4 and described below.



(a) A new experience is written in the dictionary with the hidden state of the convolutional neural network as key and Q-value estimate as value. The value of a state is evaluated by looking at the value of similar states in the dictionary.



(b) Architecture of episodic memory module for a single action a. Pixels representing the current state enter through a convolutional neural network on the bottom left and an estimate of $Q(s,a)$ exits top right. Gradients flow through the entire architecture.

Figure 5.4: Illustration of operations of a differentiable neural dictionary. Taken from Pritzel et al. (2017).

Writing to a Differentiable Neural Dictionary

The NEC agent perceives the world through a convolutional neural network (the same as described by Mnih et al. 2015) which processes observation s_t from the environment and outputs an embedding vector h. This vector is used as a key for storing a memory in the DND, this means that it is added to the key matrix of the selected action a (i.e. K_a). The corresponding value to be added to V_a is computed as a Q-value estimate combining N-step on-policy rewards and off-policy bootstrap:

$$Q^{(N)}(s_t, a) = \sum_{j=0}^{N-1} \gamma^j r_{t+j} + \gamma^N \max_{a'} Q(s_{t+N}, a'). \tag{5.1}$$

The bootstrap term $\max_{a'} Q(s_{t+N}, a')$ is found by querying all memories M_a for each action a and taking the highest estimated Q-value returned. Because a memory cannot be added before waiting N steps to compute the Q-value estimate, memories for time t must be written in the DND at $t' \geq t + N$.

If a state is already present in K_a, the corresponding value of V_a, Q_i, is updated with rate α:

$$Q_i \leftarrow Q_i + \alpha(Q^N(s, a) - Q_i). \tag{5.2}$$

When maximum capacity is reached, the item that has least recently shown up as a neighbour in the lookup operation described below is overwritten.

Lookup from a Differentiable Neural Dictionary

The lookup operation of a DND is performed as follows:

$$o = \sum_i w_i v_i \tag{5.3}$$

where v_i is the ith value and

$$w_i = k(h, h_i) / \sum_j k(h, h_j) \tag{5.4}$$

with $k(x, y)$ a kernel between vectors x and y defined as:

$$k(h, h_i) = \frac{1}{||h - h_i||_2^2 + \delta} \tag{5.5}$$

$$\delta = 10^{-3}. \tag{5.6}$$

To ensure scalability, the lookup operation is limited to the 50 nearest neighbors with respect to their key vector. This is done using a kd-tree that approximates nearest-neighbor lookup (Bentley 1975). The agent learns to minimize the discrepancy between the output o and a target Q-value estimate and selects the best action using the traditional Deep Q-Network method (Mnih et al. 2015). Like other episodic controllers, NEC leverages fast episodic storage to maximize data efficiency. Pritzel et al. (2017) showed that it indeed learns very rapidly and outperforms previous reinforcement learning techniques on Atari games.

5.6.3 Controlled Episodic Memory

While experience replay and episodic control have greatly influenced the field of reinforcement learning by drawing inspiration from hippocampal projections to the prefrontal cortex, the reciprocal connections are still underexplored in this context. Some notable exceptions do not explicitly refer to these anatomical connections but implement elements of top-down control over memory (e.g. Fortunato et al. 2019; Lu et al. 2022; K. T. Jensen et al. 2023). For example, Fortunato et al. (2019) augmented NEC with a long short-term memory (LSTM; Hochreiter and Schmidhuber 1997) which serves as a working memory from which the action is selected. Instead of learning the heteroassociation between sensory states and Q-value estimates like in NEC, the DND of Fortunato et al. (2019) stores sensory and working mem-

ory states as keys, and the retrieved value is a combination of past working memory states that is fed back to the working memory. Crucially, the reading operation not only takes into account the sensory state but also the state of the working memory, allowing for top-down interactions. This echoes the prefrontal control over hippocampal operations, and Fortunato et al. (2019) showed that their architecture outperforms NEC in a variety of tasks.

Other related works include model-based episodic learning (Le et al. 2021), episodic meta-learning (L. C. Melo 2022; Ritter et al. 2018), separate retrieval mechanisms (Goyal et al. 2022), and so on. Yet, this line of machine learning research is largely disconnected from the biology too.

In computational neuroscience, prefrontal-to-hippocampus interactions have been modeled more explicitly by Pilly et al. (2018). They extended the model of Ketz et al. (2013) described in the previous chapter to account for the contextual modulation of hippocampal memory processing by the prefrontal cortex. In order to do that, they used context information to bias the excitability of granule cells in the dentate gyrus. They thus divided the population of dentate gyrus cells into four subsets corresponding to four contexts. When an association was presented to the hippocampus, the excitability of context-inappropriate dentate gyrus cells was suppressed. The model performed the task of Peters et al. (2013) in which rats were presented with eight odor discrimination problems. When presented in a different context, the reward contingency was reversed, such that the correct odor of the first context was unrewarded in the second one, and conversely. Thanks to the modulation of granule cells, the model successfully acquired this contextual memory task and reproduced biological findings. Unlike the proposal of Howard Eichenbaum (2017b), the authors argued that prefrontal information is potentially conveyed to the dentate gyrus by the supramammillary nucleus, yet it is not clear whether this small region is able to support this function.

5.7 Discussion

In this chapter, interactions between episodic memory and cognitive control in the brain have been described. Hippocampus-to-prefrontal connections have been introduced as a way to provide episodic memory information to decision-making regions. Conversely, prefrontal-to-hippocampus connections have been shown to exert cognitive control on cognitive actions instead of motor actions, more precisely they contextually control memory operations. Finally, related computational models have been presented.

We are still a long way from understanding how the simple interactions revealed by experiments enable complex functions of the construction system of the human brain (Hassabis and Maguire 2009), such as imagination. This is partly due to a lack of computational models related to this topic, and this thesis aims to help fill this gap. First, the link between episodic control algorithms and the biological implementation of episodic memory is unclear. In the next chapter, the differentiable neural dictionary of neural episodic control will be related to Hopfield networks (Chapter 6). Furthermore, computational models of the prefrontal influence on memory recall in the hippocampus are rare and involve a direct anatomical connection. The contextual modulation of modern Hopfield networks through the intermediate entorhinal cortex will be investigated in Chapter 7. Finally, a robust model of one-shot sequence learning in the hippocampus will be developed in Chapter 8, building upon the results of the previous chapters. Among other things, it will be shown that this model is capable of disambiguating overlapping sequences by forming contextual representations.

Part II

Computational Modeling

In this part, computational tools are employed to investigate episodic memory and its relationship with cognitive control. Chapters 6 and 7 are introductory studies on elements of prefrontal-hippocampal interactions. The most important contribution is the CONSEQUENCE model of one-shot sequence learning in the hippocampus, introduced in Chapter 8, that builds upon the modern Hopfield network studied in Chapters 6 and 7.

6 Hopfield Episodic Control: Relating Associative Memory to Reinforcement Learning Algorithms

6.1 Introduction

In this chapter, a novel connection is established between the fields of associative memory and reinforcement learning. It is shown in Section 6.2 that the differentiable neural dictionary (DND ; Section 5.6.2) introduced in the context of episodic control (Pritzel et al. 2017) is a Hopfield network. Retrieval from a DND can indeed be decomposed into similarity, separation and projection operations, just as any instance of the recently proposed universal Hopfield network framework (UHN ; Millidge et al. 2022). In a DND, similarity scores are computed using the Euclidean distance and separated with a k-nearest neighbor algorithm. The projection operation of DND is the same as for all existing UHN instances. DND can thus be thought of as a single-shot associative memory model, just like Hopfield networks and their modern continuous variants. The implications are twofold. The first one is that, while DND was introduced in the context of reinforcement learning, it can be used to store and retrieve information in a wide variety of forms. This ability is illustrated in Section 6.3 using three image datasets. The second implication is that any UHN model can be used to

store experiences in place of the DND for episodic control, as shown in Section 6.4.

6.2 Differentiable Neural Dictionary as a Hopfield Network

In the UHN framework, retrieval is performed by applying similarity and separation functions sep and sim respectively, then projecting the results to the output space with some value matrix $\boldsymbol{V}$ (Equation 4.7):

$$\boldsymbol{z} = \boldsymbol{V} \cdot \text{sep}(\text{sim}(\boldsymbol{K}, \boldsymbol{q})). \tag{6.1}$$

Rewriting the reading operation of a DND (Equation 5.3) with different variable names yields:

$$z = \sum_i v_i w_i, \tag{6.2}$$

where $\boldsymbol{w}$ contains the normalized distances between the query and the keys of the k-nearest neighbors. Using vector notation, the parallel with Hopfield networks becomes more explicit:

$$z = \boldsymbol{v} \cdot \boldsymbol{w} \tag{6.3}$$

$$z = \boldsymbol{v} \cdot \text{sep}(\text{sim}(\boldsymbol{K}, \boldsymbol{q}), k), \tag{6.4}$$

where sim is the kernel function:

$$\text{sim}_i(\boldsymbol{K}, \boldsymbol{q}) = \frac{1}{||\boldsymbol{K}_i - \boldsymbol{q}||_2^2 + \delta} \tag{6.5}$$

which is equal to the Euclidean similarity function defined in Millidge et al. (2022). The separation function finds the top k similarity values (thus implementing the k-nearest neighbor algorithm) and normalizes them:

$$\text{sep}(\boldsymbol{x}, k) = \kappa(\boldsymbol{x}, k) / \sum_i \kappa_i(\boldsymbol{x}, k) \tag{6.6}$$

$$\kappa_i(\boldsymbol{x}, k) = \begin{cases} x_i & \text{if } x_i \text{ is among the top } k \text{ values of } \boldsymbol{x}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

In sum, the kernel function of DND is a similarity function, as it is equivalent to the Euclidean similarity function of Millidge et al. (2022). Furthermore, the k-nearest neighbor algorithm sparsifies the result of the similarity function by cancelling the contribution of the most distant experiences. The output of the algorithm is then normalized before being projected to the value space. This constitutes a novel separation function for the UHN framework. It is worth noting that this separation function is similar to applying a threshold on the similarity function, like what is done in the sparse distributed memory model (Kanerva 1988), which has also been cast as a UHN (Millidge et al. 2022). The only difference is that the threshold for sparse distributed memory is fixed, while it must be dynamic for selecting a constant number k of neighbors. Furthermore, with $k = 1$, the separation function is equivalent to the max function benchmarked by Millidge et al. (2022). The remaining difference between DND and other UHN instances is that the output of the DND is a scalar value, while UHN models can store vectors. As mentioned in Chapter 5, a DND has been used by Fortunato et al. (2019) to store multidimensional values. In fact, the DND can simply be extended with a matrix of value vectors. Equation 6.3 thus becomes:

$$\boldsymbol{z} = \boldsymbol{V} \cdot \boldsymbol{w}. \tag{6.8}$$

6.3 Differentiable Neural Dictionary as an Associative Memory

In the previous section, the differentiable neural dictionary of neural episodic control has been shown to be a universal Hopfield network. In principle, DND can thus be used as an associative memory. In this section, the MNIST, CIFAR10 and Tiny ImageNet datasets are used to test the robustness and capacity of DND as an associative memory model, using the same methods as in Millidge et al. (2022), unless otherwise specified.

Example reconstructions of MNIST digits by DND are shown in Figure 6.1. Memories are separated by keeping the k-nearest neighbors only (or k-max). For simplicity, the kd-tree of the original neural episodic control implementation is not used. Instead, all similarity scores are computed and those not selected are zeroed out. Using $k = 50$ like in Pritzel et al. (2017) gives an output from which the original image can be recognized, but the model is unable to properly separate memories and the resulting output is blurry. The max separation function is equivalent to selecting the nearest neighbor ($k = 1$) and provides a much clearer output. In fact, max was the best separation function benchmarked by Millidge et al. (2022). Even when the output is blurry, the $k > 1$ models seem to properly capture the statistics of the dataset, such as the fact that central pixels are globally more active than those on the borders. Recall accuracy is typically assessed in absolute terms, by checking that the difference between output and response is below a threshold (Millidge et al. 2022). It is worth exploring whether the statistical modeling capacities of $k > 1$ can consistently improve performance with this criterion, despite the max function having theoretically unbounded capacity with respect to the dimensionality of the query (Millidge et al. 2022). On the other hand, a new criterion will have to be introduced to evaluate recall in relation to other stored patterns, given that the main function of episodic memory is to reconstruct information corresponding to a particular situation rather than to generalize.

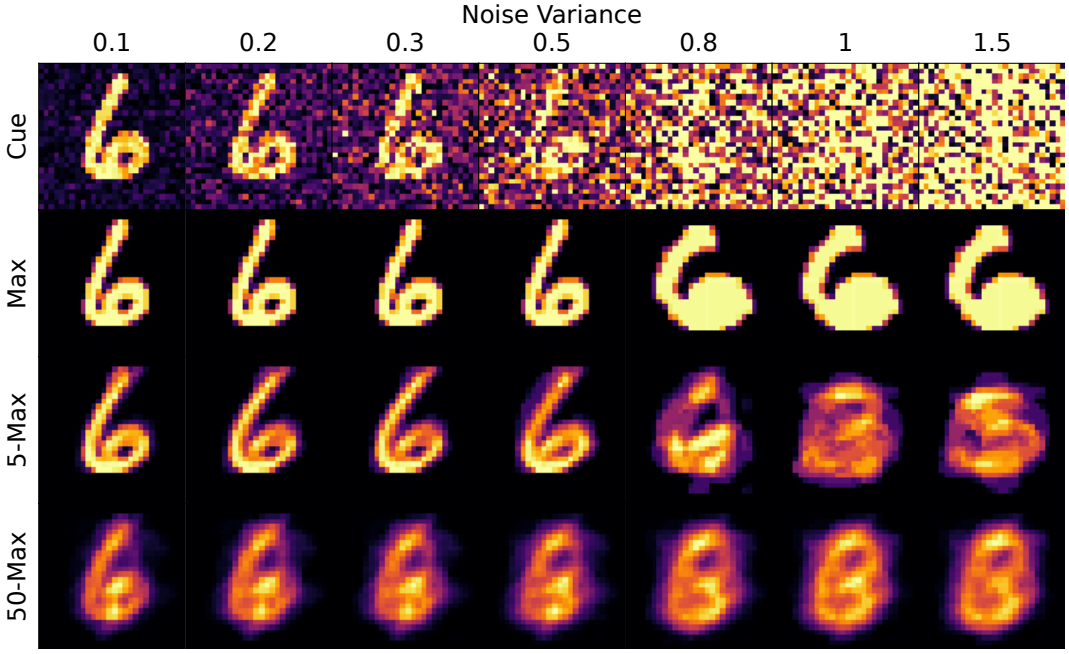


Figure 6.1: Example reconstructions of noisy MNIST digits by DND. The top row shows the input cue with increasing amount of noise. The following rows show the reconstruction of the stored memory using $k = 1, 5, 50$.

6.3.1 Capacity with Different Functions

In DND, the similarity between memories and query is computed using a Euclidean function (Equation 6.5). While rarely used in associative memory models, this function was introduced by Millidge et al. (2022) and found to outperform the more common dot product. An even better performing similarity function was the inverse of the Manhattan distance:

$$\text{sim}_i(\boldsymbol{K}, \boldsymbol{q}) = \frac{1}{\delta + \sum_j \text{abs}(\boldsymbol{K}_{ij} - \boldsymbol{q}_j)} \tag{6.9}$$

The capacity of the model under different similarity and separation functions is assessed by quantifying correctly retrieved data when increasing number of MNIST, CIFAR10 and Tiny ImageNet images is stored (Figure 6.2 ; Table 6.1). Half-masked images are given as input, and a trial is correct if the sum of squared pixel differences between the output and the actual image is less than a threshold of 50. The Manhattan similarity function out-

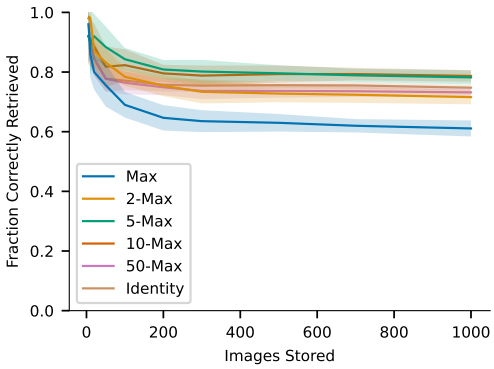
Similarity	Separation	MNIST	CIFAR10	Tiny ImageNet
Euclidean	Max	0.739 ± 0.14	0.220 ± 0.18	$\mathbf{0.223 \pm 0.21}$
	2-Max	0.826 ± 0.11	$\mathbf{0.236 \pm 0.18}$	0.015 ± 0.02
	5-Max	$\mathbf{0.851 \pm 0.08}$	0.117 ± 0.09	0.010 ± 0.02
	10-Max	0.838 ± 0.08	0.095 ± 0.08	0.010 ± 0.02
	50-Max	0.801 ± 0.09	0.087 ± 0.09	0.010 ± 0.02
	Identity	0.809 ± 0.08	0.088 ± 0.09	0.010 ± 0.02
Manhattan	Max	0.835 ± 0.10	$\mathbf{0.451 \pm 0.21}$	$\mathbf{0.669 \pm 0.24}$
	2-Max	0.886 ± 0.08	0.369 ± 0.20	0.011 ± 0.02
	5-Max	$\mathbf{0.887 \pm 0.07}$	0.106 ± 0.08	0.010 ± 0.02
	10-Max	0.826 ± 0.08	0.075 ± 0.07	0.010 ± 0.02
	50-Max	0.775 ± 0.11	0.067 ± 0.06	0.010 ± 0.02
	Identity	0.804 ± 0.09	0.063 ± 0.06	0.010 ± 0.02

Table 6.1: Capacity of associative memory with different similarity and separation functions assessed with MNIST, CIFAR10 and Tiny ImageNet datasets. Reported are means and standard deviations of the 10 simulations of Figure 6.2. For each dataset and similarity function, the best performance is highlighted in bold.

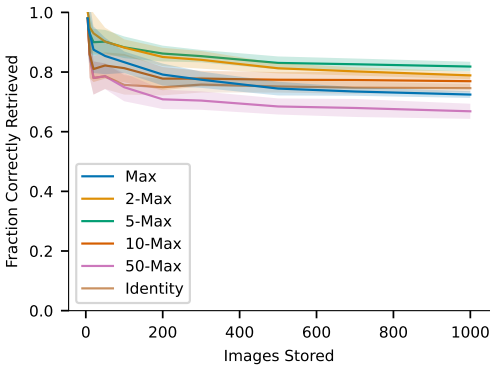
performs the Euclidean function, especially when k is low. Furthermore, the best k value is highly dependent on the dataset. In MNIST, the best k is 5 for both Euclidean and Manhattan functions. In CIFAR10, the best k is 2 with Euclidean similarity and 1 with the Manhattan function. In Tiny ImageNet, max ($k = 1$) outperforms other functions.

6.3.2 Retrieval with Different Functions

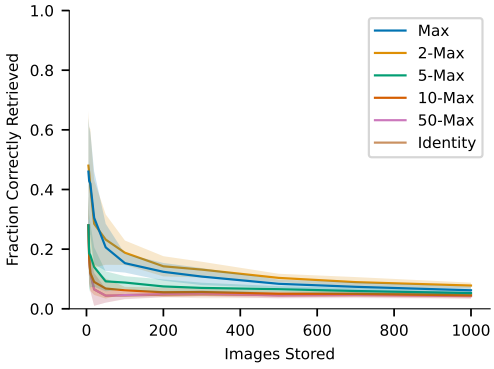
In order to test robustness of the memory, the ability to recall memories from noisy cues is analyzed. Independent zero-mean Gaussian noise with variance σ is thus added to the query images pixelwise. Like capacity, the best k for retrieval depends on the dataset (Figure 6.3 ; Table 6.2). Here again, the performance is better with the Euclidean similarity than with the Manhattan similarity for low k, and worse for high k. In MNIST, the best k is 50 with both Euclidean and Manhattan similarities. In CIFAR10, 2



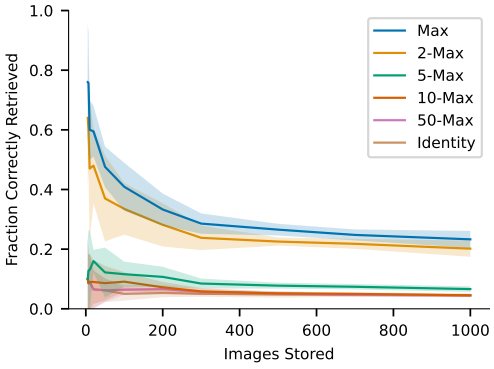
(a) MNIST with Euclidean similarity



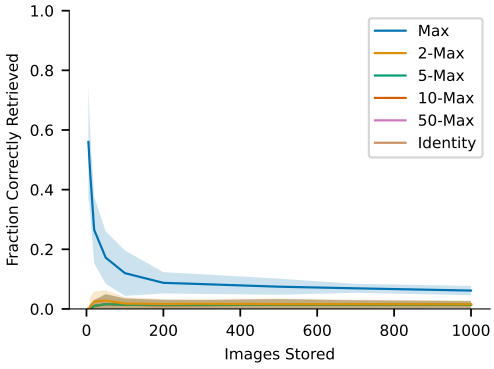
(b) MNIST with Manhattan similarity



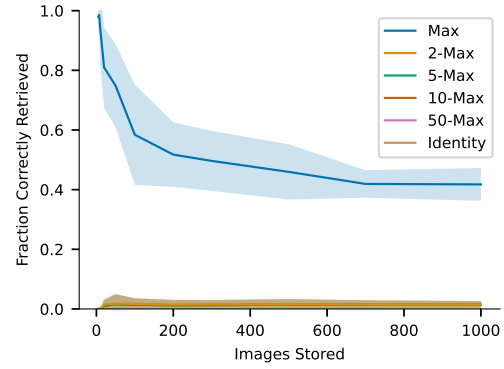
(c) CIFAR10 with Euclidean similarity



(d) CIFAR10 with Manhattan similarity



(e) Tiny ImageNet with Euclidean similarity



(f) Tiny ImageNet with Manhattan similarity

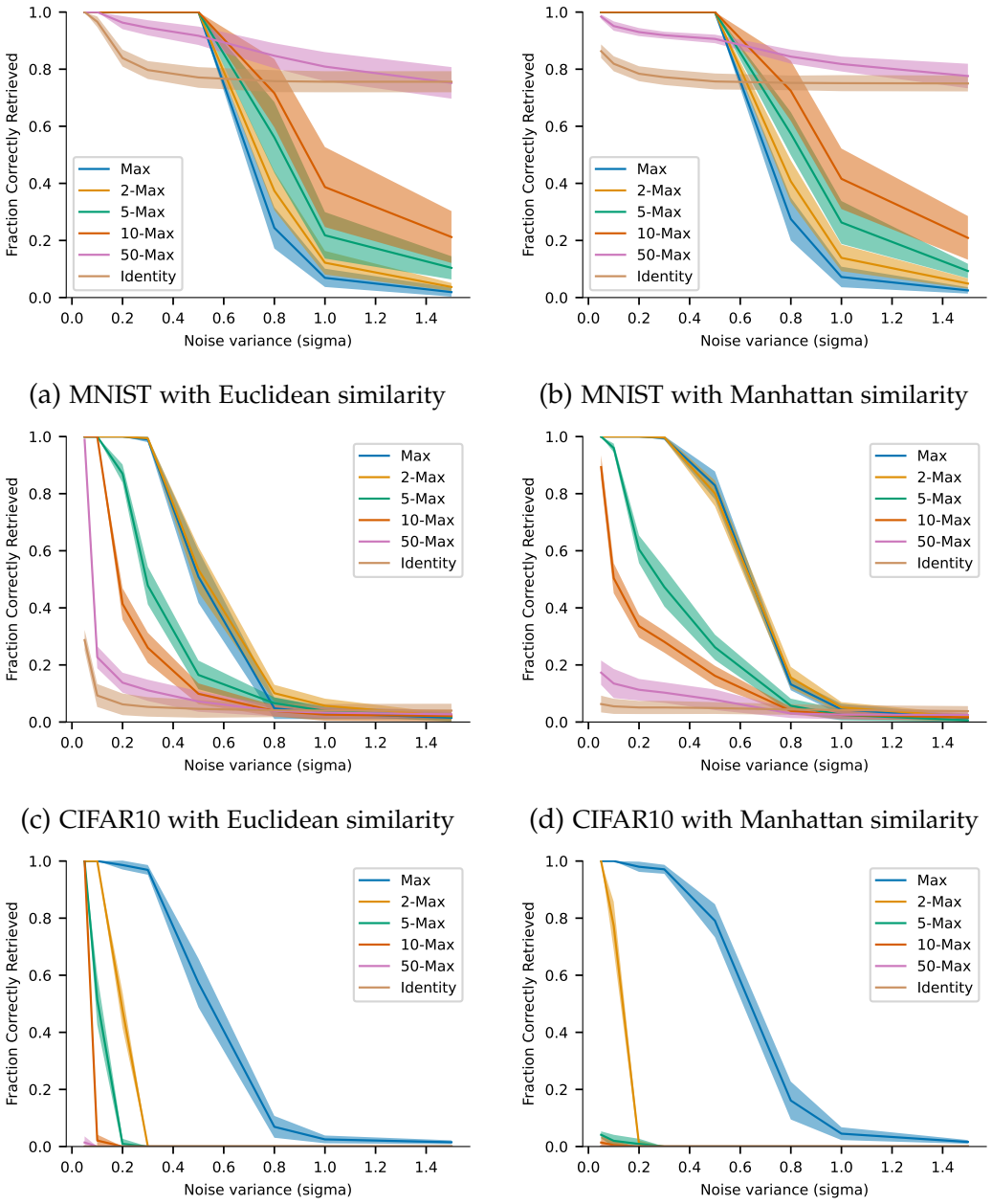
Figure 6.2: Capacity of associative memory with different similarity and separation functions assessed with MNIST, CIFAR10 and Tiny ImageNet datasets. Plots represent the means and standard deviations of 10 simulations. A trial is correct when the difference between the output and the actual memory is under a threshold.

Similarity	Separation	MNIST	CIFAR10	Tiny ImageNet
Euclidean	Max	0.667 ± 0.44	0.574 ± 0.45	$\mathbf{0.580 \pm 0.44}$
	2-Max	0.692 ± 0.41	$\mathbf{0.588 \pm 0.44}$	0.310 ± 0.43
	5-Max	0.735 ± 0.37	0.455 ± 0.41	0.190 ± 0.35
	10-Max	0.789 ± 0.31	0.357 ± 0.39	0.128 ± 0.33
	50-Max	$\mathbf{0.904 \pm 0.09}$	0.205 ± 0.30	0.002 ± 0.01
	Identity	0.830 ± 0.10	0.083 ± 0.09	0.000 ± 0.00
Manhattan	Max	0.672 ± 0.43	0.627 ± 0.44	$\mathbf{0.620 \pm 0.43}$
	2-Max	0.699 ± 0.40	$\mathbf{0.628 \pm 0.43}$	0.223 ± 0.39
	5-Max	0.741 ± 0.36	0.424 ± 0.38	0.009 ± 0.02
	10-Max	0.794 ± 0.30	0.282 ± 0.28	0.002 ± 0.01
	50-Max	$\mathbf{0.891 \pm 0.07}$	0.085 ± 0.06	0.000 ± 0.00
	Identity	0.781 ± 0.05	0.049 ± 0.02	0.000 ± 0.00

Table 6.2: Retrieval capability against noise. Reported are means and standard deviations of the 10 simulations of Figure 6.3. For each dataset and similarity function, the best performance is highlighted in bold.

is the best k. Like for capacity, max again outperforms other functions for retrieval of noisy Tiny ImageNet images.

As hypothesized with Figure 6.1, max is not always the best performing function with the absolute accuracy criterion, both for capacity and retrieval. It can indeed be outperformed by higher k values, meaning that taking into account more memories than the single most similar one can lead to more precise recall in absolute terms. However, the performance of the identity and 50-Max functions in the MNIST dataset are surprisingly good, even under very high levels of noise. In fact, as pixel values are restricted to lie in the range $[0, 1]$, it is very unlikely that enough information remains in the image to correctly identify it when $\sigma > 1$. Hence, a plausible alternative is that high k functions model the dataset such that they output a mixture of many images that is sometimes classified as correct retrieval, although it is not necessarily closer to the query image than any other of the dataset. This is especially true for MNIST which contains simple pictures that are more similar to each other than CIFAR10 and Tiny ImageNet.



(a) MNIST with Euclidean similarity

(b) MNIST with Manhattan similarity

(c) CIFAR10 with Euclidean similarity

(d) CIFAR10 with Manhattan similarity

(e) Tiny ImageNet with Euclidean similarity

(f) Tiny ImageNet with Manhattan similarity

Figure 6.3: Retrieval capability against increasing levels of noise. Plots represent the means and standard deviations of 10 simulations with different sets of images. A trial is correct when the difference between the output and the actual memory is under a threshold.

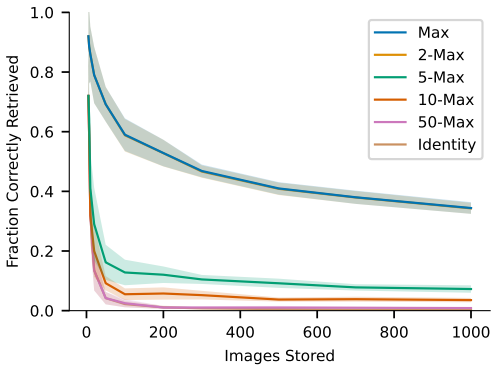
Similarity	Separation	MNIST	CIFAR10	Tiny ImageNet
Euclidean	Max	$\mathbf{0.624 \pm 0.22}$	$\mathbf{0.223 \pm 0.19}$	$\mathbf{0.223 \pm 0.21}$
	2-Max	$\mathbf{0.624 \pm 0.22}$	0.222 ± 0.19	0.222 ± 0.21
	5-Max	0.256 ± 0.24	0.132 ± 0.13	0.112 ± 0.14
	10-Max	0.199 ± 0.24	0.104 ± 0.14	0.098 ± 0.13
	50-Max	0.171 ± 0.25	0.086 ± 0.14	0.095 ± 0.14
	Identity	0.169 ± 0.25	0.085 ± 0.14	0.095 ± 0.14
Manhattan	Max	$\mathbf{0.793 \pm 0.14}$	$\mathbf{0.502 \pm 0.24}$	$\mathbf{0.669 \pm 0.24}$
	2-Max	0.790 ± 0.14	0.486 ± 0.25	0.505 ± 0.36
	5-Max	0.255 ± 0.22	0.176 ± 0.19	0.155 ± 0.20
	10-Max	0.187 ± 0.22	0.103 ± 0.15	0.119 ± 0.18
	50-Max	0.163 ± 0.23	0.090 ± 0.15	0.113 ± 0.18
	Identity	0.162 ± 0.23	0.088 ± 0.15	0.113 ± 0.18

Table 6.3: Capacity of associative memory with different similarity and separation functions assessed with MNIST, CIFAR10 and Tiny ImageNet datasets. Reported are means and standard deviations of the 10 simulations of Figure 6.4. For each dataset and similarity function, the best performance is highlighted in bold.

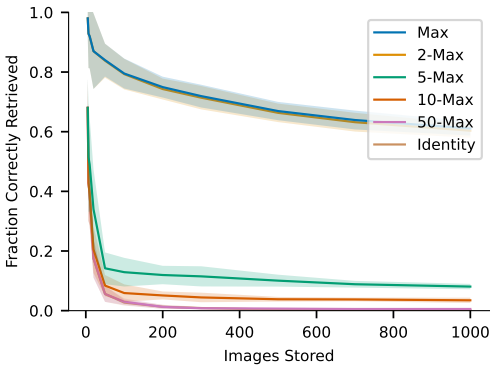
6.3.3 Performance with Relative Criterion

In order to prevent associative memory models to model the statistics of the dataset rather than focusing on the query image to output the actual memory, a novel criterion is introduced. Instead of the absolute threshold, retrieval must be good relatively to other images. More precisely, the novel criterion is such that a trial is correct if and only if the squared pixel difference between the truth and the output is lower than it is between the output and any other memory. Capacity is again assessed using this new criterion (Figure 6.4 ; Table 6.3). The Manhattan function still outperforms the Euclidean similarity. Most crucially, the best performance is always obtained with the max function.

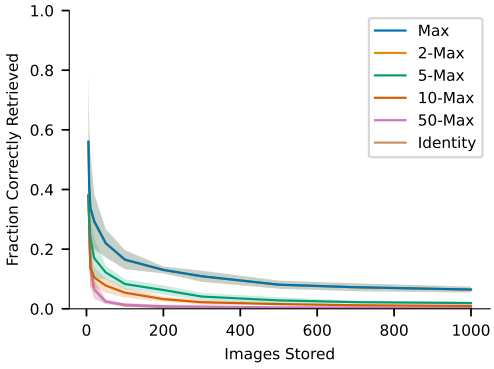
Retrieval is then tested with the new criterion (Figure 6.5 ; Table 6.4). Once again, the best performance is obtained with the Manhattan similarity. Furthermore, $k = 1$ almost always outperforms other values. Note that performance with $k = 2$ is very similar.



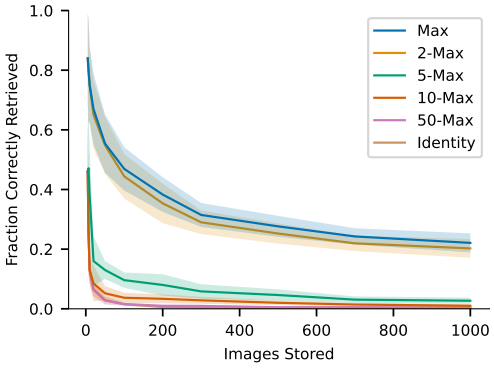
(a) MNIST with Euclidean similarity



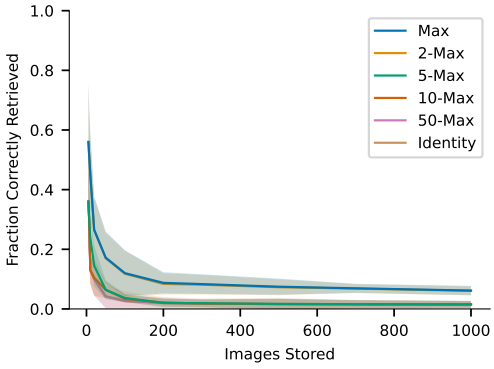
(b) MNIST with Manhattan similarity



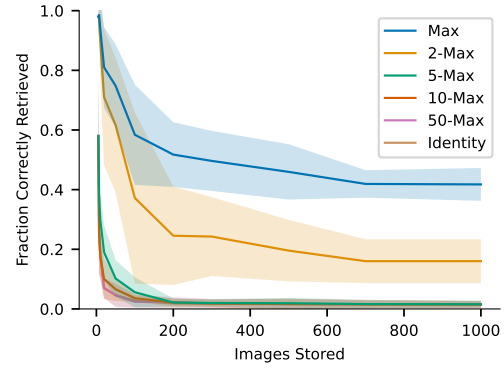
(c) CIFAR10 with Euclidean similarity



(d) CIFAR10 with Manhattan similarity

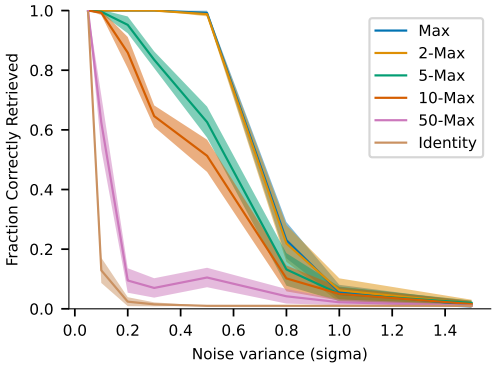


(e) Tiny ImageNet with Euclidean similarity

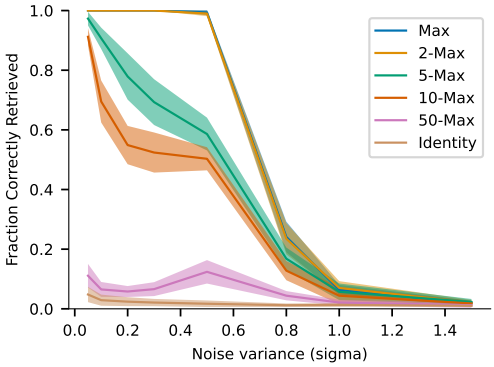


(f) Tiny ImageNet with Manhattan similarity

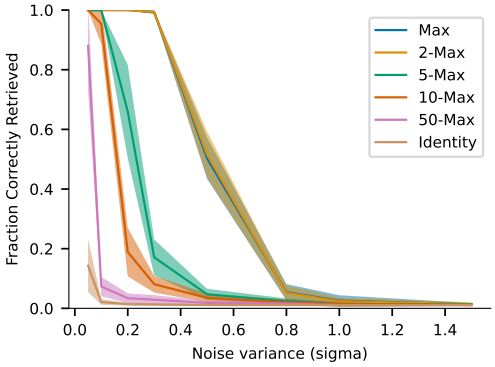
Figure 6.4: Capacity of associative memory with different similarity and separation functions assessed with MNIST, CIFAR10 and Tiny ImageNet datasets. Plots represent the means and standard deviations of 10 simulations. Here, a trial is correct when the difference between the output and the actual memory is lower than the difference between the output and any other stored memory.



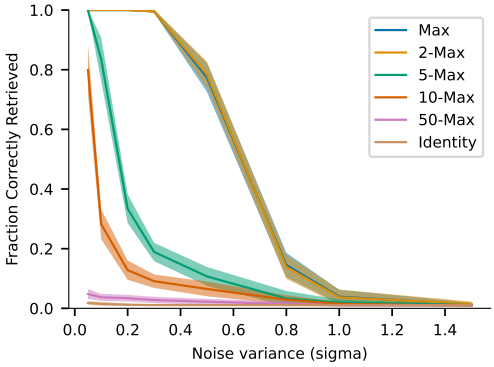
(a) MNIST with Euclidean similarity



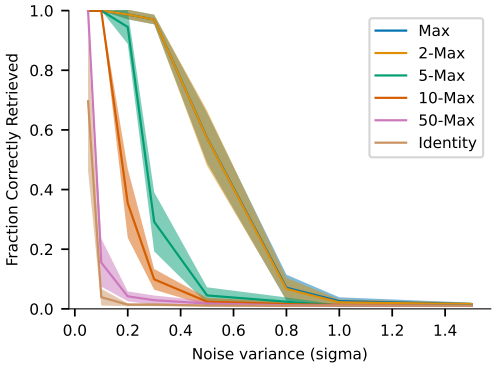
(b) MNIST with Manhattan similarity



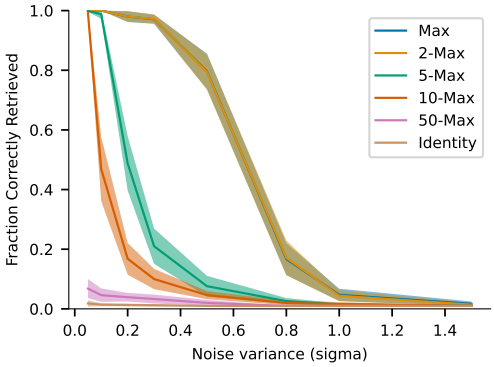
(c) CIFAR10 with Euclidean similarity



(d) CIFAR10 with Manhattan similarity



(e) Tiny ImageNet with Euclidean similarity



(f) Tiny ImageNet with Manhattan similarity

Figure 6.5: Retrieval capability against increasing levels of noise. Plots represent the means and standard deviations of 10 simulations with different sets of images. Here, a trial is correct when the difference between the output and the actual memory is lower than the difference between the output and any other stored memory.

Similarity	Separation	MNIST	CIFAR10	Tiny ImageNet
Euclidean	Max	$\mathbf{0.661 \pm 0.44}$	0.574 ± 0.45	$\mathbf{0.580 \pm 0.44}$
	2-Max	$\mathbf{0.661 \pm 0.44}$	$\mathbf{0.575 \pm 0.44}$	0.579 ± 0.44
	5-Max	0.576 ± 0.41	0.366 ± 0.42	0.416 ± 0.45
	10-Max	0.522 ± 0.39	0.288 ± 0.40	0.314 ± 0.41
	50-Max	0.246 ± 0.34	0.134 ± 0.29	0.160 ± 0.32
	Identity	0.150 ± 0.32	0.029 ± 0.05	0.101 ± 0.24
Manhattan	Max	$\mathbf{0.665 \pm 0.44}$	$\mathbf{0.621 \pm 0.44}$	$\mathbf{0.622 \pm 0.43}$
	2-Max	0.664 ± 0.43	$\mathbf{0.621 \pm 0.44}$	$\mathbf{0.622 \pm 0.43}$
	5-Max	0.523 ± 0.36	0.316 ± 0.37	0.352 ± 0.40
	10-Max	0.422 ± 0.31	0.177 ± 0.25	0.228 ± 0.33
	50-Max	0.063 ± 0.04	0.026 ± 0.02	0.030 ± 0.03
	Identity	0.022 ± 0.02	0.012 ± 0.00	0.012 ± 0.01

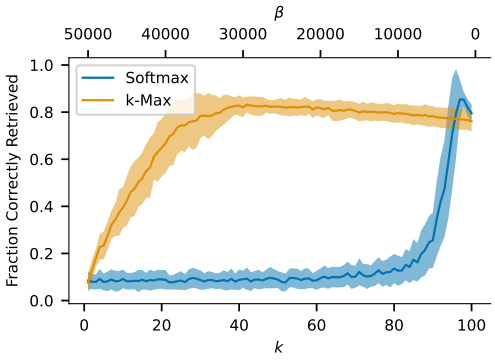
Table 6.4: Retrieval capability against noise. Reported are means and standard deviations of the 10 simulations of Figure 6.5. For each dataset and similarity function, the best performance is highlighted in bold.

6.3.4 Relationship Between k-Max and Softmax

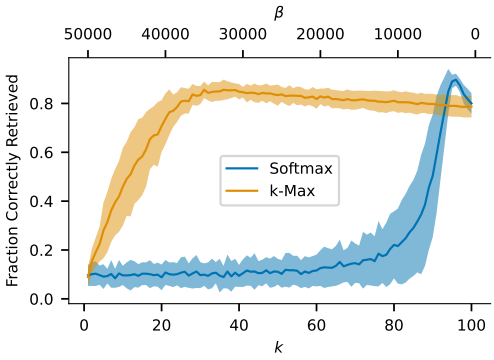
Like k-Max, the softmax function virtually cancels out the contribution of distant memories, especially when β, the scaling parameter of its input, is high. It does it by normalizing exponentiated similarity scores (Equation 4.6). While k is a discrete parameter, β is continuous, which potentially makes the softmax function more flexible to optimize. Here, the two separation functions are compared. For each dataset, 100 images are encoded. The noise is set to 1 for MNIST and 0.75 for CIFAR10 and Tiny ImageNet. The results are shown in Figure 6.6 with the absolute criterion and in Figure 6.7 with the relative criterion. Except for the condition with CIFAR10, Euclidean similarity and absolute criterion (Figure 6.6c), the softmax can always outperform k-Max.

6.4 Hopfield Network for Episodic Control

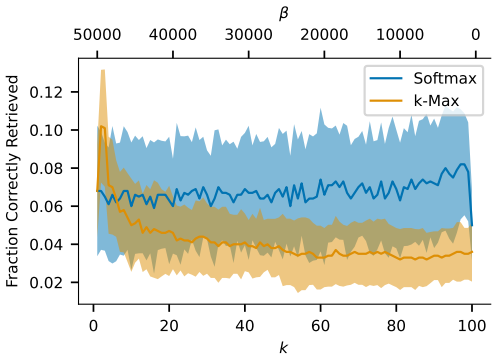
Neural episodic control greatly improves the sample efficiency of reinforcement learning, as it rapidly assimilates new experience in the dictionary



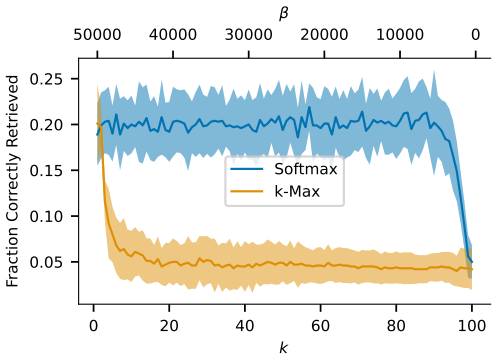
(a) MNIST with Euclidean similarity



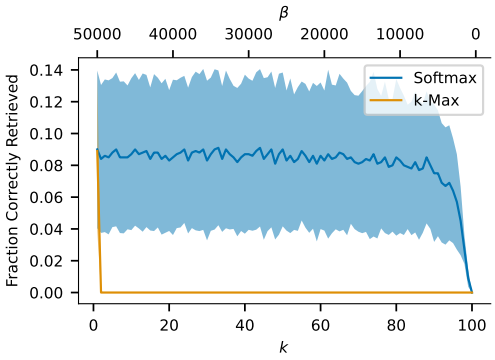
(b) MNIST with Manhattan similarity



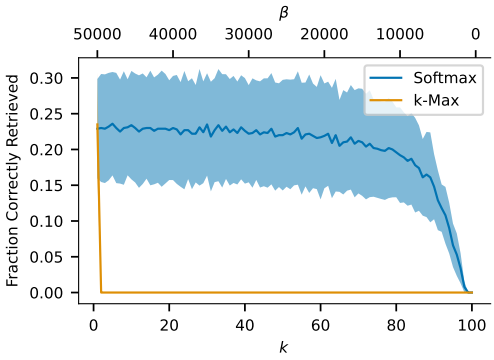
(c) CIFAR10 with Euclidean similarity



(d) CIFAR10 with Manhattan similarity

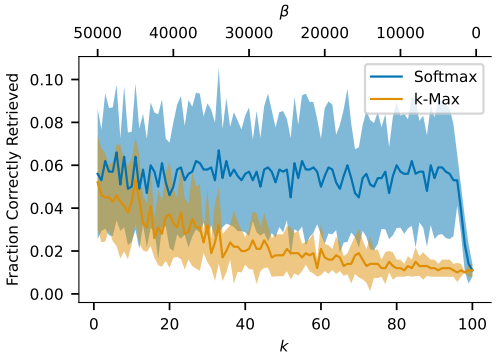


(e) Tiny ImageNet with Euclidean similarity

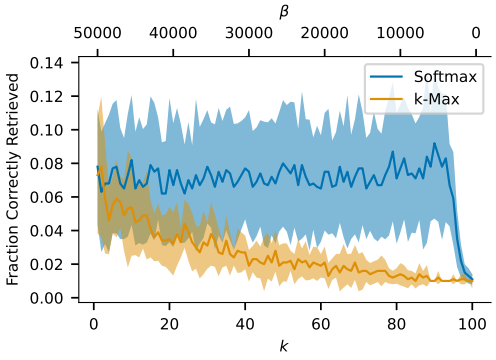


(f) Tiny ImageNet with Manhattan similarity

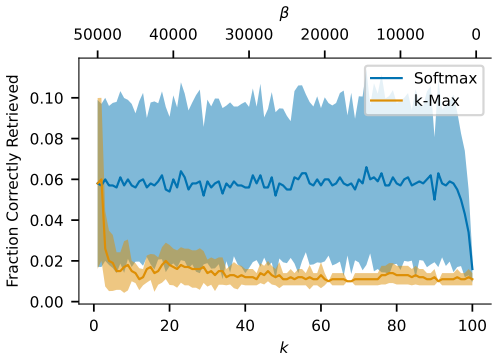
Figure 6.6: Retrieval capability as a function of k and β parameters of the k-Max and softmax separation functions respectively. Plots represent the means and standard deviations of 10 simulations with different sets of images. A trial is correct when the difference between the output and the actual memory is under a threshold.



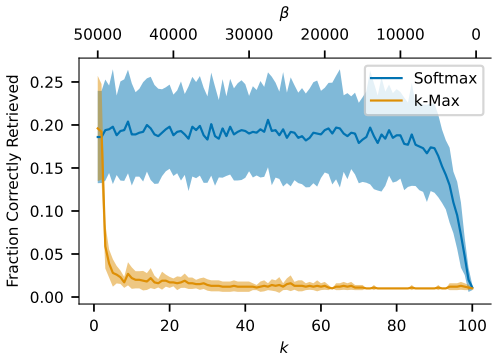
(a) MNIST with Euclidean similarity



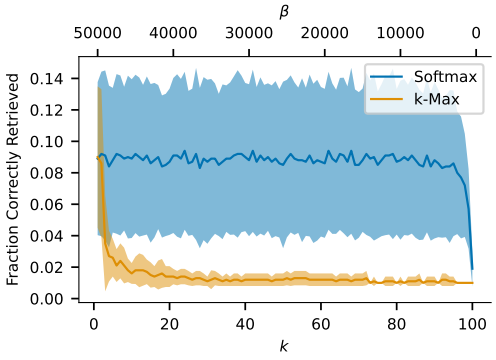
(b) MNIST with Manhattan similarity



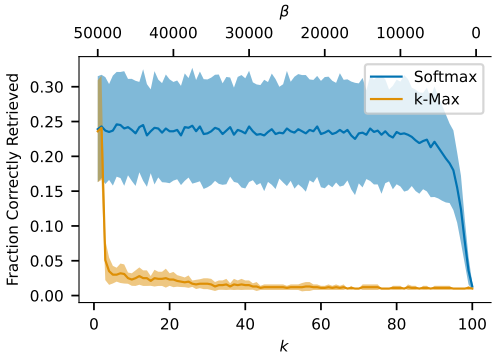
(c) CIFAR10 with Euclidean similarity



(d) CIFAR10 with Manhattan similarity



(e) Tiny ImageNet with Euclidean similarity



(f) Tiny ImageNet with Manhattan similarity

Figure 6.7: Retrieval capability as a function of k and β parameters of the k-Max and softmax separation functions respectively. Plots represent the means and standard deviations of 10 simulations with different sets of images. Here, a trial is correct when the difference between the output and the actual memory is lower than the difference between the output and any other stored memory.

(Pritzel et al. 2017). Since the differentiable neural dictionary is a Hopfield network, a question worth exploring is whether other instances of the universal Hopfield network framework can be used for reinforcement learning, and perhaps outperform the DND. The strength of this framework is that it unifies disparate models that sequentially apply a set of operations. In fact, both the Euclidean kernel and k-nearest neighbor of NEC can be replaced by any differentiable similarity and separation functions, respectively.

The code of Sarrico et al. (2019) is used. It is made more flexible to incorporate Manhattan and dot product kernels, as well as the softmax separation function. The models are tested in two classic problems, CartPole and Acrobot, implemented in OpenAI Gym (Brockman et al. 2016). When possible, the original hyperparameters of Pritzel et al. (2017) are used, although some values are taken from Sarrico et al. (2019) who trained their algorithms in the same reinforcement learning environments instead of Atari games like in Pritzel et al. (2017). The number of nearest neighbors is set to $k = 11$ and the memory can store a total of 10,000 entries. A more exhaustive report of hyperparameters can be found in Table 6.5. The agents are trained during 500,000 steps and evaluated every 5000 steps.

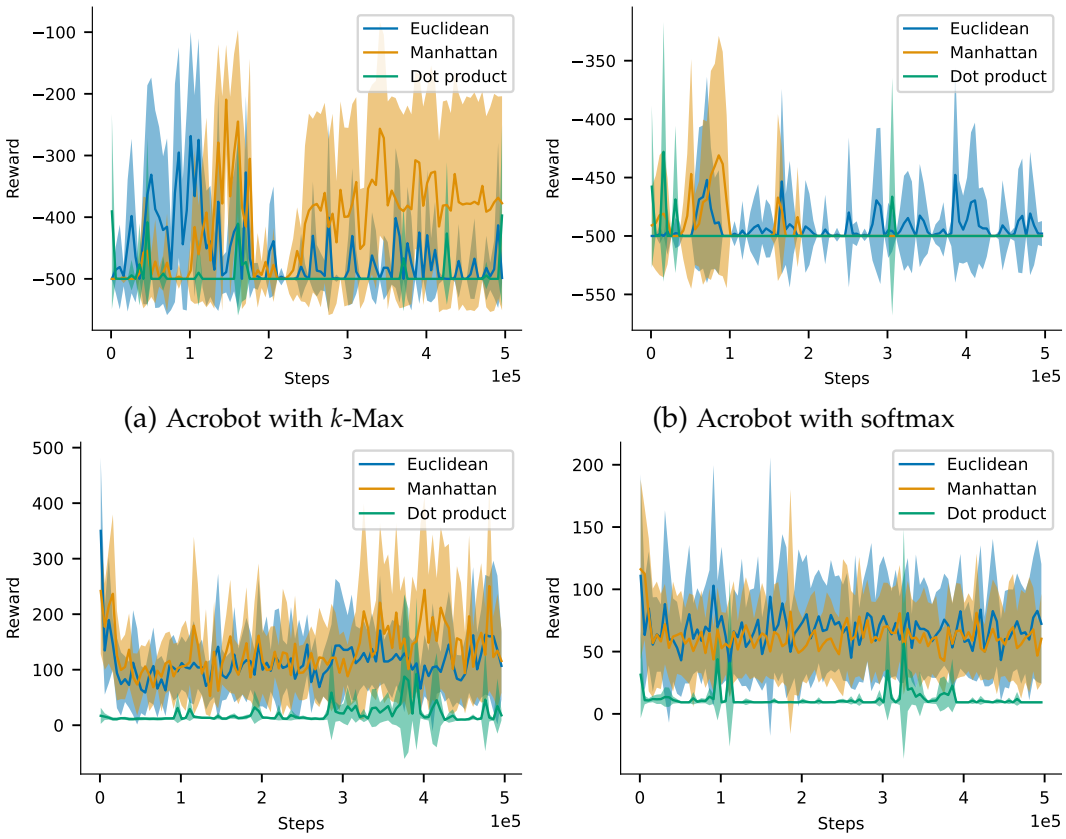
The results of 3 runs are shown in Figure 6.8. In this (very preliminary) experiment, the dot product did not perform as good as the other similarity functions. Furthermore, while the performance of the Manhattan and Euclidean functions is very similar, Manhattan similarity performed better in the Acrobot environment with k-Max separation. Finally, the softmax function with randomly chosen $\beta = 100$ did not perform as great as k-Max. Overall, the agents did not successfully learn the tasks.

6.5 Discussion

In this chapter, DND, which has initially been introduced in the context of reinforcement learning (Pritzel et al. 2017), has been shown to be mathematically related to associative memory models like Hopfield networks. The

Parameter	Value
Initial policy ϵ	1
Final policy ϵ	5×10^{-3}
Starting step of policy ϵ annealing	5×10^{3}
Ending step of policy ϵ annealing	25×10^{3}
Discount factor γ	0.99
Number of neighbors k	11
Kernel δ	10^{-3}
Softmax β	100
Experience replay size	10^{5}
RMSprop learning rate	7.92×10^{-6}
RMSprop momentum	0.95
RMSprop ϵ	10^{-2}
Starting step of training	10^{3}
Batch size	32
Memory learning rate	0.1
n-step return	100
Key size	64

Table 6.5: Hyperparameters used for neural episodic control simulations.



(a) Acrobot with k-Max

(b) Acrobot with softmax

(c) Cartpole with k-Max

(d) Cartpole with softmax

Figure 6.8: Mean and standard deviation learning curves of 3 different runs in Acrobot and Cartpole using different similarity and separation functions.

universal Hopfield network framework has recently been introduced to encompass the traditional Hopfield network, modern variants and related models (Millidge et al. 2022). These models recall memories with a common sequence of operations: similarity, separation and projection. It has been shown that retrieval from a DND is also done with such operations. Hence, a DND is an instance of the universal Hopfield network framework.

This novel link contributes to the field of associative memory. The similarity function of DND is Euclidean and has already been shown to yield high capacity (Millidge et al. 2022). On the other hand, the k-nearest neighbor is not commonly used as a separation function for associative memory. Thus, one of the present objectives was to assess its performance. Interestingly, k controls the separation function, and setting $k = 1$ is equivalent to using the max function studied by Millidge et al. (2022). While having theoretically unbounded capacity, the max function can transition sharply from one memory attractor to another when noise of increasing amplitude is added to the query. With Figure 6.1, it was hypothesized that higher k values could be better at modelling datasets and improve the performance assessed in absolute terms. Simulations indeed revealed higher capacity and better retrieval from noisy queries with $k > 1$, especially with simple datasets like MNIST. However, these results depend on the way performance is evaluated. In Millidge et al. (2022), the evaluation of retrieval is based on some distance evaluation of the memory output and the actual image, which must not exceed some threshold fixed by the experimenter. This is a widespread method for evaluating associative memory models, but one must choose the threshold wisely, as setting it too high can result in false positives with the model reproducing statistics of the dataset grossly. This for example seems to be the case when retrieval is assessed using the MNIST dataset. The performance of 50-Max (and even the identity function) remains high despite very strong noise. Therefore, another way of evaluating retrieval was used which does not consider the output in absolute terms, but rather compares it to the whole memory set. Retrieval is deemed correct if and only if the output resembles the actual image more than any other stored memory. Memory models thus cannot benefit from

modeling statistics of the dataset, and must rather focus on recalling the distinguishing characteristics of the query. Using this method, the max function outperforms the others. Ideally, performance should be evaluated in both absolute and relative terms to ensure that recall is accurate and stands out from other memories.

This raises the question of what is the function of associative memory. Modeling statistics of a dataset is related to generalization, which is typically the main goal of machine learning. The objective of associative memory is somewhat different. Instead of generalizing, an associative memory aims to recall the exact information corresponding to the individual memory. This is reminiscent of the division of labor between episodic and semantic memory. When it comes to episodic control however, that is the use of episodic memories for action control, some generalization is desirable. This is especially the case in neural episodic control in which the selection of actions only relies on episodes, the hippocampal module constituting a bottleneck. Initially, episodic control (not to be confused with its implementation in neural episodic control) has been introduced as a way of speeding up the learning of reinforcement agent and, after the initial episodic control phase, it is desirable that more robust controllers can take over (Lengyel and Dayan 2007). A biologically inspired alternative to neural episodic control would be to supplement episodic memory with other controllers whose function is to generalize. The hippocampus would then no longer be a bottleneck, and could instead be devoted to memorizing the specifics of situations.

Another thing that was briefly explored is the use of different functions for neural episodic control. The simulations were done in simple environments, as a proof of concept. It is possible that the Manhattan function could consistently outperform the Euclidean kernel, as it does on the associative memory tasks. As their biological plausibility has yet to be proven, the Manhattan and Euclidean functions will not be used in the remainder of this thesis. The softmax function, which has been proven powerful in transformers (Vaswani et al. 2017), could also improve episodic control agents with a more informed choice of β. More conclusive evidence could

come from training agents to play Atari games, as in the original NEC study, and optimizing hyperparameters such as k and β. Unfortunately, a proper benchmark of the new functions on Atari games would require far greater computing resources than are presently available in the laboratory.

The next chapter focuses on how modern Hopfield networks can be contextually modulated. Then, Chapter 8 will present a model of how they can be implemented in the hippocampus.

7 Contextual Control Over Memory Recall

Before presenting the new sequence learning model, the connections from the prefrontal cortex to the hippocampus are studied. In order to guide behavior efficiently, recalled memories should be relevant to the current situation. More precisely, the memory selection process should be influenced by both contextual and sensory information. In contrast to the information that is directly accessible to the agent through its perception of the environment, contextual information encompasses any internal representation that is not directly perceptible, but is susceptible to help the agent predict future rewards and states. This includes information about the temporal unfolding of events, spatial integration, but also the goals pursued by the agent. While stimulus-driven retrieval has been studied both empirically and theoretically, much less is known about the contextual modulation of episodic recall. A better comprehension of the mechanisms involved is necessary to develop a complete theory of episodic memory, but also to develop machine learning algorithms that can comprehend a wider variety of spatiotemporal contexts and direct their behavior towards goals. As reviewed in Chapter 5, the prefrontal cortex controls the hippocampus to recall contextually relevant memories. In the standard framework (Chapter 4.5), episodic memories are thought to be stored in recurrent collaterals of CA3. However, there is no direct anatomical connection from the prefrontal cortex to CA3 (Andrianova et al. 2023), and it is thus not clear how retrieval in the latter can be modulated by the former.

In a modeling study (Pilly et al. 2018), contextual information was fed to the dentate gyrus to modulate recall. This is at odds with the current un-

derstanding of prefrontal-to-hippocampus projections (Chapter 5; Howard Eichenbaum 2017b). The aim is not to compete with their model, but rather to study how prefrontal signals can modulate memory through the entorhinal cortex. An artificial neural network with Hopfield associative memory is used to study performance and contextual representations based on the memory locus and the target of contextual information.

7.1 Methods

7.1.1 Task

Like the model of Pilly et al. (2018), our model performs the task described in Peters et al. (2013). Object-discrimination problems are presented in two different contexts and used to evaluate context-guided memory recall. In a given task t_i and context c, the network learns to choose a rewarded object $R_{i,c}$ over a non-rewarded object $\bar{R}_{i,c}$. In the alternative context c', it learns to choose between the same objects as in context c, except that the non-rewarded object in context c is rewarded in c', and conversely (i.e. $\bar{R}_i^c = R_i^{c'}$ and $R_i^c = \bar{R}_i^{c'}$). Objects are represented in the network as 6×4 matrices with 6 active entries of value 1 placed at random locations, and other entries with value 0 (Figure 7.2).

7.1.2 Model

The hippocampus model is a fully-connected neural network which follows the anatomy of the biological medial temporal lobe (Figure 7.1). Information flows through a trisynaptic pathway (Input $\rightarrow$ EC_{II} $\rightarrow$ DG $\rightarrow$ CA3 $\rightarrow$ CA1 $\rightarrow$ EC_V $\rightarrow$ Output) and a monosynaptic pathway (Input $\rightarrow$ EC_{III} $\rightarrow$ CA1 $\rightarrow$ EC_V $\rightarrow$ Output). The number of neurons in each region is set as the number of neurons in the analog rat region (as retrieved from Cutsuridis et al. 2019), scaled down by a factor $N = 0.003$ (Table 7.1). The network input consists of three pattern slots: two for the choice objects and one for the rewarded object (Table 7.2). The assignment of R_i^c and $\bar{R}_i^c$ to the first

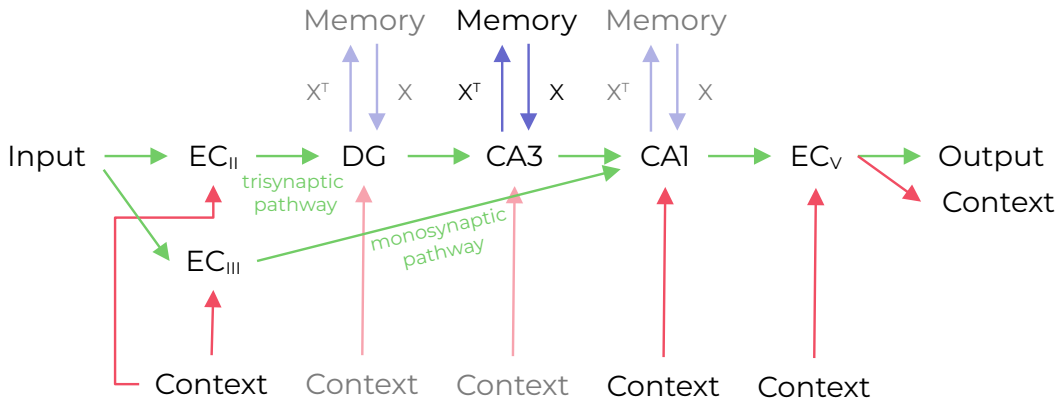


Figure 7.1: Architecture of the model. The deep learning network is trained to autoencode its input (green), and the context when indicated (red). The associative memory is modified during the encoding phase, and used to recall memories during the test phase. In the standard framework, it is in CA3 (blue), but other locations are also tested here (transparent blue). Context information can target various layers (transparent red), but the current understanding (Howard Eichenbaum 2017b) is that it is provided to the entorhinal cortex (red).

or second slot is random to prevent the network from learning to always output the same slot. ReLU is used as the activation function of hidden layers and sigmoid is used in the output layer to map latent activity back to the pattern space. An additional input is used to modulate layers of the hippocampal network contextually. It represents the two possible contexts as a two-dimensional one-hot vector. When a layer is targeted by this contextual layer, this simply means that it receives it as an additional input, in a feedforward way. Note that the output layer is never modulated, since it does not model a medial temporal lobe region.

One of the hippocampal layers (CA3, unless otherwise mentioned) is augmented with an autoassociative memory similar to the modern Hopfield network of Ramsauer et al. (2020). During the memory encoding phase, activity vectors of the hippocampal layer are manually added to the au-

Table 7.1: Network summary.

Layer	Output size	Function
EC_{II}	$110,000N$	ReLU
EC_{III}	$250,000N$	ReLU
DG	$1,200,000N$	ReLU
CA3	$250,000N$	ReLU
CA1	$390,000N$	ReLU
EC_{V}	$330,000N$	ReLU
Output	$3 \times 6 \times 4$	sigmoid

Table 7.2: Input summary. The “Objects" input size corresponds to the product of the number, width and height of input patterns.

Input	Size
Context	2
Objects	$3 \times 6 \times 4$

toassociative weight matrix $\boldsymbol{X}$. During the test phase, activity of the layer $\boldsymbol{a'}$ is completed with memory recall before being sent to the next layer:

$$\boldsymbol{a} = \boldsymbol{X}\text{softmax}(\beta \boldsymbol{X}^\top \boldsymbol{a'}), \tag{7.1}$$

where β is set to 1,000 to strongly separate memories.

7.1.3 Simulation Details

All task-context combinations are presented once during each epoch, in random order. Similarly to the recently proposed autoencoder model of the hippocampus (Santos-Pata et al. 2021), the network learns to output its input, namely the two input object patterns and the rewarded one, as well as the context vector (unless otherwise mentioned). This is done by updating feedforward weights using backpropagation, in a pretraining phase. The loss function to be minimized is the mean squared error between the input and the output. Learning is online, as training examples are presented one by one. Subsequently, in the acquisition phase, weights are fixed and memories are stored in the associative memory. Two training regimes are tested. In the many-shot regime, pretraining lasts 20 epochs with 50 tasks. The same 50 tasks are then acquired in the memory and used to test performance. Contrastingly, in the one-shot regime, pretraining lasts a single epoch with 1000 tasks, such that networks are updated as many times as

in the many-shot regime. In the acquisition phase of the one-shot regime, 50 novel tasks are presented to the network and used to test performance.

While the rewarded pattern is provided as input to the network during the pretraining and training phases, it is masked during test (i.e. values of the third input slot are set to 0). In the test phases, a trial is considered correct when the pattern stored in the third slot of the output is closer to the rewarded pattern than the non-rewarded one, as assessed by the mean squared error. Thus, the expected performance of a random network, or a network with no contextual target, is 50% and serves as baseline. The learning rate is a sensible parameter, as its optimal value varies among the conditions tested in this work. Therefore, for each tested condition, the learning rate is set as the average of the five learning rates that best performed during a random search. This purely stochastic optimization process evaluated the performance of the model in each condition, with 300 random learning rate values. After this process, 50 different simulations were performed with the optimal learning rate of each condition.

7.2 Results

7.2.1 Behavioral Performance

The need for associative memory is first assessed to ensure that the task requires episodic processing. In the many-shot task, the results with CA3 memory (M = 65.06, SD = 7.85) and without modern Hopfield network (M = 70.64, SD = 4.05) indicate that memory impairs performance with this training regime ($t(49) = -4.97$, $p < .001$), as shown in Figure 7.3b. On the other hand, in the one-shot task, results with CA3 memory (M = 75.06, SD = 4.64) and without modern Hopfield network (M = 49.54, SD = 3.19) indicate that memory improves performance when test tasks are presented once during encoding ($t(49) = 34.89$, $p < .001$). In these conditions, CA3 is directly targeted by context information. To test whether the model can perform the task in accordance with the model of Howard Eichenbaum (2017b), context information is then provided to entorhinal layers. The per-

formance with entorhinal target (M = 51.06, SD = 3.05) and CA3 target (M = 75.06, SD = 4.64) indicate that the model is much less performant when context information is given to the entorhinal cortex ($t(49) = 28.2$, $p < .001$), almost no better than chance (Figure 7.3c). It thus appears that context information does not reach the memory layer in absence of a direct connection. This constraint is however alleviated by placing the memory in other subfields, such as the dentate gyrus and CA1. When the dentate gyrus is the memory-augmented layer, the model almost performs equally well whether the dentate gyrus is directly targeted or contextual information transits through superficial (input) layers of the entorhinal cortex (Figure 7.3a). Similarly, when the associative memory is in CA1, the model performs well when context is provided to upstream layers (Figure 7.3e). When superficial layers of the entorhinal cortex are targeted and memory is in CA1, the performance of the model with the monosynaptic pathway (M = 70.14, SD = 4.71) and without it (M = 56.28, SD = 4.05) indicate that this pathway is necessary ($t(49) = 17.55$, $p < .001$), as shown in Figure 7.3f. By combining results of when the dentate gyrus, CA3 or CA1 (no monosynaptic pathway for simplicity) is memory-augmented, it is possible to evaluate performance according to the position of the targeted layer relative to the memory layer. This analysis reveals that targeting regions downstream memory has no effect (Figure 7.3d). Spearman's rank correlation was computed to assess the relationship between the relative position of the target (with positions less than or equal to 0) and performance. There was a positive correlation between the two variables, $r(898) = 0.78, p < .001$. Furthermore, a two-way ANOVA was performed to evaluate the interaction effects of relative position (less than or equal to 0) and context reconstruction on performance. The means and standard deviations for performance are presented in Table 7.3. The results indicated a significant main effect for relative position ($\mathrm{F}(3, 899) = 545.59, p < .001$), main effect for context reconstruction ($\mathrm{F}(1, 899) = 24.29, p < .001$), and interaction between relative position and context reconstruction ($\mathrm{F}(3, 899) = 9.06, p < .001$). Figure 7.3d indeed shows that the steepness of performance decline, as target is located more and more upstream memory, is greater when the model only

Table 7.3: Descriptive statistics for performance.

Objective	Target position	M	SD
Sensory	-3	50.54	3.33
	-2	52	3.43
	-1	60.69	6.93
	0	71.91	5.27
Sensory + Context	-3	56.28	4.05
	-2	53.91	4.53
	-1	63.53	6.96
	0	71.43	5.86

tries to reconstruct sensory input than when it also tries to reconstruct context information in the pretraining phase.

7.2.2 Neural Coding of Context

In order to understand how the model either succeeds or fails at performing the task, the representations employed in the different layers can be analyzed. Principal component projections reveal that contexts are well separated at the targeted EC_{II} population (Figure 7.4a), but this separation diminishes in subsequent layers DG (Figure 7.4b) and CA3 (Figure 7.4c), to such an extent that the associative memory in CA3 then struggles to separate memories contextually (Figure 7.4d). Downstream layers thus cannot know the context to answer correctly.

One possibility is that some neurons are specifically modulated by context identity and underlie the separation in the principal component space. In order to find such context cells, paired t-tests are used to assess whether the activity of each cell is dependent on the context (either stronger or weaker), treating the 50 test tasks as samples. Neurons with $p < .001$ are considered context cells. This analysis reveals that 68% of EC_{II} neurons are context cells, and that this number decreases in the next populations (Figure 7.4).

The presence of context-modulated cells is very likely to help the model performing the tasks. In order to ensure that context cells indeed have

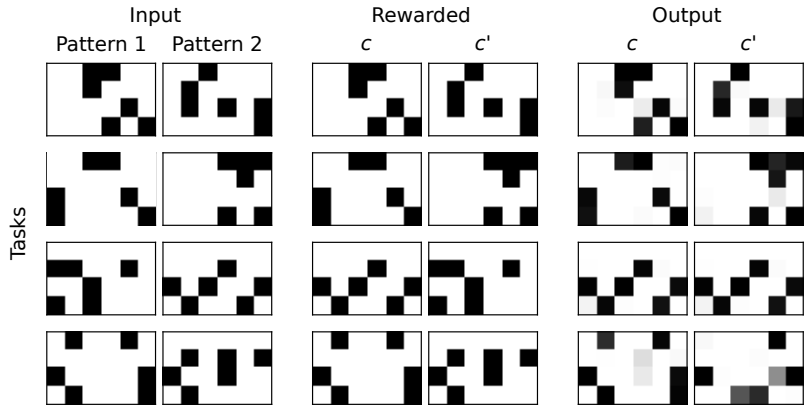
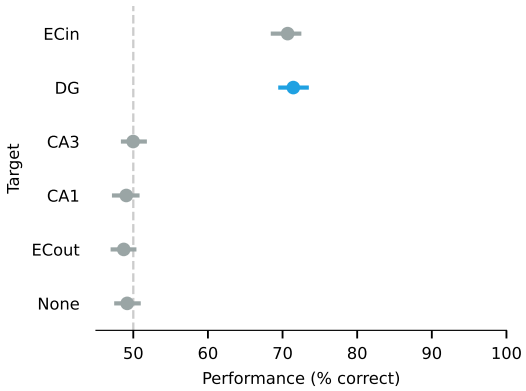
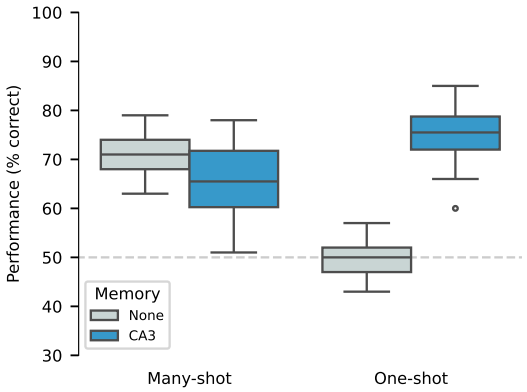


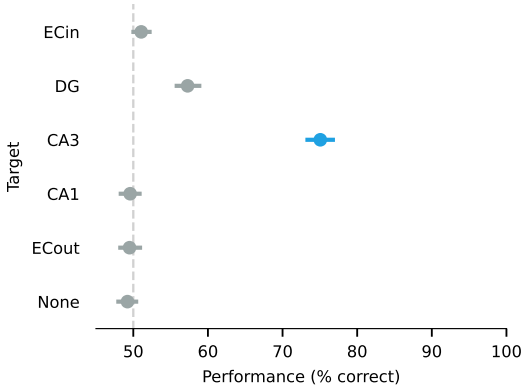
Figure 7.2: Example tasks in which the network succeeds (first two rows) and fails (last two rows) to recall the correct memory during the test phase. The two choice patterns that are placed in the first slots of the input are shown on the left. The rewarded patterns of contexts c and c' are shown in the middle. On the right, the choice of the network (i.e. third slot of the output) is shown for both contexts. In the first two tasks, the output of the network successfully corresponds to the rewarded pattern in both contexts. In the third task, however, the network chooses the same pattern in both contexts and thus fails to complete the task. In the fourth task, the network outputs the correct pattern in context c but outputs a pattern that is not in the choice set for that task in c'. However, the trial is still considered a success, as the output is closer to $R_i^{c'}$ than $\bar{R}_i^{c'}$. For these examples, the network was pretrained on 10000 tasks with a learning rate of $10^{-4.5}$.



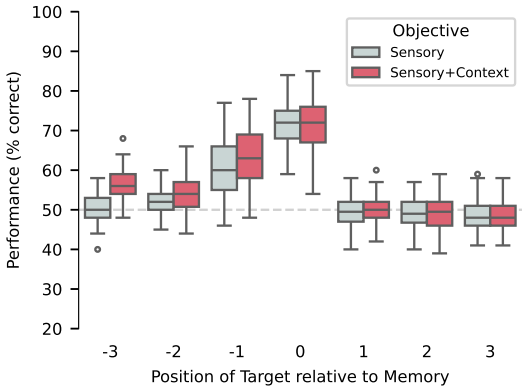
(a) Performance of the model when the associative memory is in DG (blue).



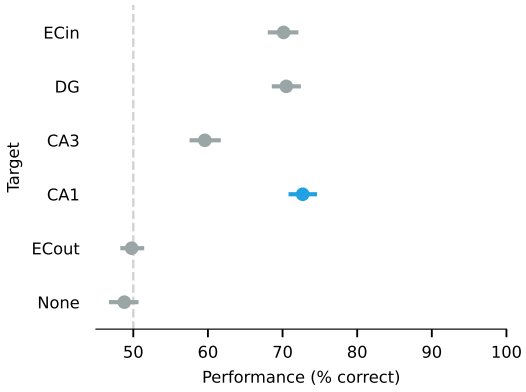
(b) Performance of training regimes with and without associative memory (target is CA3).



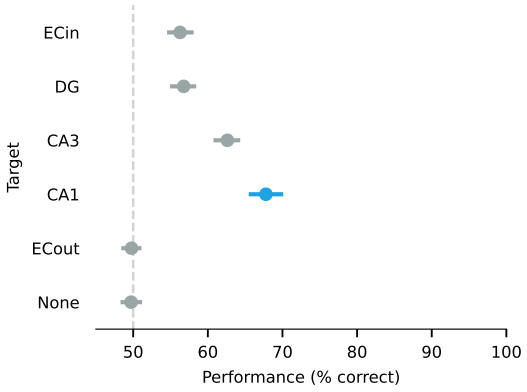
(c) Performance of the model when the associative memory is in CA3 (blue).



(d) Performance as a function of the location of the target relative to the associative memory, using sensory only and full reconstruction losses. No monosynaptic pathway was used with CA1 memory. Relative positions P and -P indicate that the target and the memory are P layers apart. The target is located before the memory with -P, and after with P.

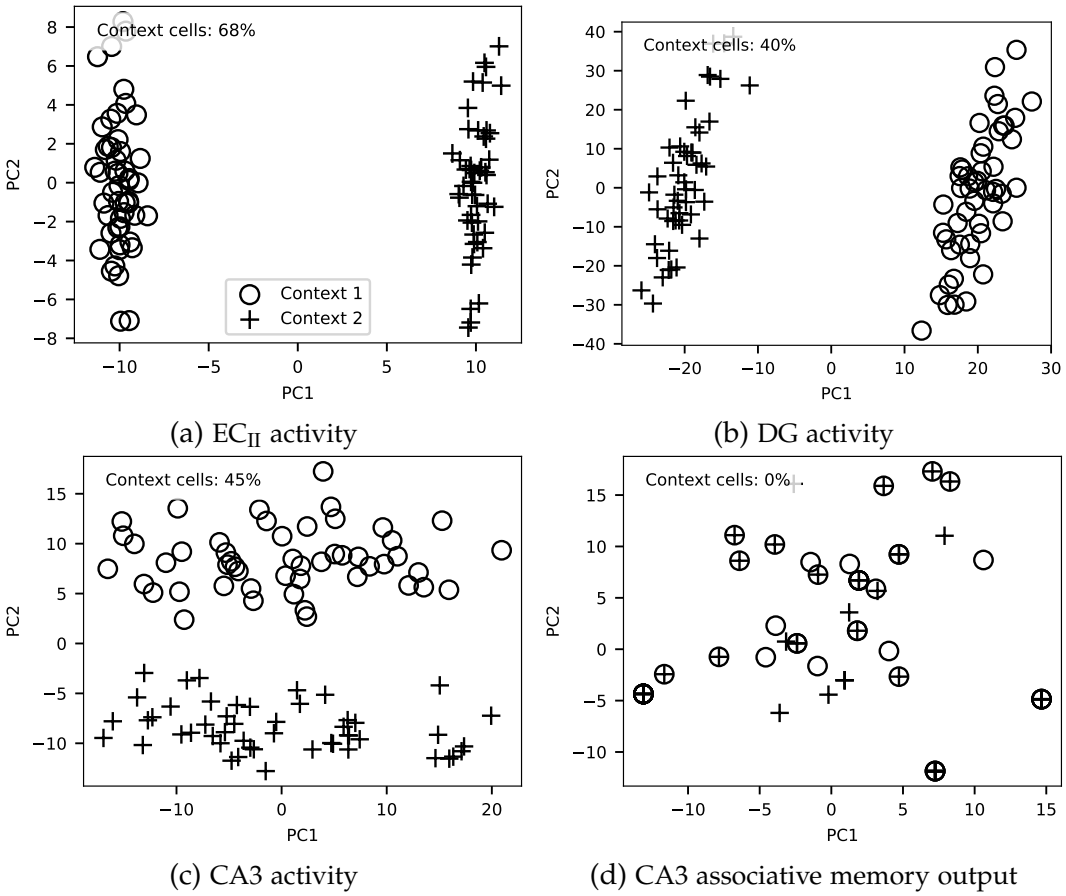


(e) Performance of the model when the associative memory is in CA1 (blue).



(f) Performance of the model when the associative memory is in CA1 (blue) and there is no monosynaptic pathway.

Figure 7.3: Behavioral results. Gray dashed lines indicate the baseline performance of 50%.



(a) EC_{II} activity

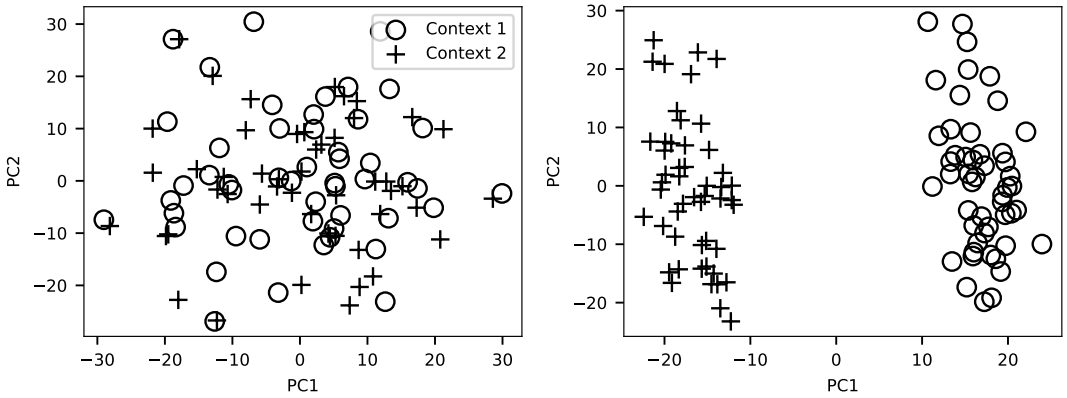
(b) DG activity

(c) CA3 activity

(d) CA3 associative memory output

Figure 7.4: Principal component analysis of activity in different layers. Here, EC_{II} is the target and CA3 is the memory.

causal influence on behavioral performance and the contextual separation of activity, a model with EC_{II} as the target and DG as the associative memory was used. After the test phase, 1664 context cells were identified in DG (46%), and were then inactivated during an additional test phase by clamping their activity to 0 in all 50 test tasks. With this protocol, the contexts are no longer separated in the principal component space (Figure 7.5a), as compared to when 1664 neurons are randomly selected for inactivation instead (Figure 7.5b). Furthermore, while random inactivation yielded 72% correct answers (similar to results of Figure 7.3a), performance dropped to 52% when context cells were inactivated.



(a) Inactivation of 1664 context cells in DG (b) Inactivation of 1664 random cells in DG

Figure 7.5: Principal component analysis of DG with EC_{II} as a target and DG as a memory, with context cell and random cell inactivation.

7.3 Discussion

A task requiring one-shot episodic memory was used to study context-guided retrieval in a simple hippocampus-inspired network. The task was introduced by Peters et al. (2013), used in the computational study of Pilly et al. (2018), and is very similar to the Think/No-Think paradigm used to study cognitive control over memory in humans (Anderson et al. 2016). The behavioral results presented here reveal that modern Hopfield networks can encode and retrieve contextual memories. The fast associative memory is necessary to perform the task when stimuli are only presented once. Furthermore, it is found that the position of the target of contextual information is a key factor in correctly performing the task. The locus of the associative memory is also important. In fact, performance seems to be governed by the distance between these two elements, such that the target must be close, but not downstream, the associative memory layer. Thus, the model that follows the standard framework of hippocampal processing and the model of prefrontal-hippocampal interactions proposed by Howard Eichenbaum (2017b) does not perform well. This is due to the associative memory in CA3 being too far from prefrontal projections to superficial layers of the entorhinal cortex. While there is a direct connection from the entorhinal cortex to CA3 in the anatomy, which could improve

performance, it is most often thought to be silenced during encoding in the standard framework.

Some alternative models are much better at performing the task. This is for example the case of models whose associative memory is located in a layer directly targeted by the prefrontal context. Alternatively, models with associative memory in DG and CA1 work well with the entorhinal cortex as the target. This is due to DG and CA1 being connected to the entorhinal cortex through the trisynaptic and monosynaptic pathways respectively. Another factor influencing performance is whether the networks learn to reconstruct context during the pretraining phase. Models that take context into account in the reconstruction error perform better. This is not surprising, as backpropagation then favors the transmission of context information across more layers to reconstruct it in the output layer. Since the hippocampus has been hypothesized to learn to autoencode its input (Santos-Pata et al. 2021; Ketz et al. 2013), and since it both receives input from and projects back to the prefrontal cortex, it is reasonable to assume that context reconstruction is a training objective of the hippocampus.

Beyond behavioral performance, the model has been found to encode context explicitly, as suggested by the separation of contexts in the principal component space and the discovery of context-modulated cells. Most importantly, these context cells have been found to be necessary to separate contexts and perform the task correctly. Context information must indeed be encoded in the input of the associative memory for it to encode and recall memories contextually. Context neurons are reminiscent of splitter cells discovered in the hippocampus (Wood et al. 2000; L. M. Frank et al. 2000; Duvelle et al. 2023), which split place representation according to the past or future trajectory. In a sense, the recent past and expected future constitute valuable context information to the hippocampus. The context cells found in this work can thus be considered splitter cells without explicit temporal and spatial components.

8 One-Shot Sequence Learning

8.1 Introduction

In this chapter, the most important contribution of this thesis is introduced: the Consequence model. With the renewal of interest in Hopfield networks and the advent of powerful modern variants, it is worth reevaluating the way the hippocampus stores memories. The standard framework of hippocampal memory processing postulates that memories are stored in recurrent collaterals of CA3, and this view was strongly influenced by Hopfield networks which were originally recurrent. The best associative memory models are now feedforward and implemented in two-layer feedforward networks that successfully compute the similarity between the query and stored memories, separate similarity scores to only recall specific memories, and project the resulting values back to the sensory space. Since these models are made of separate key and value matrices, they can store both autoassociations and heteroassociations, and thus learn sequences. The question then arises as to how the hippocampus could perform such operations.

8.2 Requirements for Hippocampal Associative Memory

A biological neural network implementing a two-layer associative memory must satisfy certain requirements (Table 8.1). First, in order for memory re-

call to be specific as is the case for episodic memory, similarity scores must be strongly separated. If memory coding is local, this means that the activity of the population in which memories are stored must be sparse during retrieval. Furthermore, it must also have enough neurons to encode many memories. The theoretical capacity of a population can be approximated as the number of possible combinations of active cells. The dentate gyrus, CA3 and CA1 contain 1,200,000, 250,000 and 390,000 neurons respectively, in the rat hippocampus (Cutsuridis et al. 2019). Furthermore, approximately 0.8%, 3% and 9% of dentate gyrus, CA3 and CA1 cells are active at a time (Neher et al. 2015). Hence, the dentate gyrus is likely to have a much bigger capacity than the other subfields, which makes it a good candidate for storing memories. A mechanism to recruit neurons is also needed to store novel information. Neurogenesis is a unique property of the dentate gyrus and could enable the storage of new memories.

The population reconstructing memories, or in other words the one storing values, also needs to meet requirements. This output layer must be targeted by the memory population and also contain the features of the value to encode, for example with a residual connection from the sensory input. It then needs to adjust synaptic weights to minimize the discrepancy between the information reconstructed from memory and the actual information to store. For sequence learning, heteroassociations needs to be performed successively, thus the output either needs to be connected back to the input, or directly to the memory neurons. CA3 is well suited for the role of ouput layer, as it receives information from the dentate gyrus, and the hilar backprojection connects it back to dentate granule cells.

The associative memory must recall information according to incoming sensory inputs and contextual information. As shown previously, the dentate gyrus is directly targeted by the entorhinal cortex which receives top-down prefrontal cues. Furthermore, CA3 is recurrently connected and can thus accumulate information over time to contextually guide sequence retrieval and disambiguate overlapping trajectories. These elements converge to a comprehensive model of how the hippocampus could implement a two-layer associative memory that can store and retrieve sequences contex-

tually (Figure 8.1, upper left model). Sensory input is sent from EC_{II} to DG, which compares it to stored memories and selects a sparse subset of relevant neurons. Memories are then decoded by CA3 and compared to the sensory input sent through the residual connection from the entorhinal cortex. This constitutes the autoassociative circuit. In the CA3-DG backprojection, associations can be learned between CA3 activity and the neurons of DG active in the following time step, as proposed by Lisman, Talamini, et al. (2005) in the Socratic model. As CA3 accumulates information over multiple time steps, the heteroassociative keys stored in this pathway are conditioned on the recent past.

Alternative implementations are compared in Table 8.1 and shown in Figure 8.1. The EC_{II}-CA3-DG circuit could implement the associative memory similarly. However, CA3 is less sparse and contains less neurons than DG. Furthermore, there is no report of neurogenesis in CA3 and it is not clear whether DG could keep track of the temporal context. Finally, it has been shown in the preceding chapter that prefrontal information sent through the entorhinal cortex fails to reach CA3, at least without the direct EC_{II}-CA3 connection. Furthermore, it does not reach CA1 in absence of the monosynaptic pathway. The EC_{II}-CA3-CA1, DG-CA3-CA1, CA3-CA1-EC_V and EC_{III}-CA1-EC_V options all lack an intrahippocampal backprojection necessary for sequence learning. On top of that, EC_{II}-CA3-CA1, DG-CA3-CA1 and EC_{III}-CA1-EC_V do not have a clear residual connection, as EC_{III} stores information of a different kind than EC_{II} and EC_V (Hafting, Fyhn, Bonnevie, et al. 2008). It is a possibility, however, that the CA3-CA1-EC_V circuit could learn to reconstruct the input from EC_{II} in EC_V, since these two entorhinal layers represent similar information. Yet, this model could not learn sequences in absence of intrahippocampal backprojection.

8.3 Comparison of Sequence Learning Models

The EC_{II}-DG-CA3 circuit underlies the proposed Consequence model (standing for contextual sequence). It exploits the unique characteristics of the

Features	**EC_{II}-DG-CA3**	EC_{II}-CA3-DG	EC_{II}-CA3-CA1	DG-CA3-CA1	CA3-CA1-EC_V	EC_{III}-CA1-EC_V
Capacity	**+++**	+	+	+	++	++
Neurogenesis	✓					
Residual Connection	**+++**	+++	+	+	++	+
Backprojection	✓	✓				
Contextual Modulation	**++**	+(+)	+(+)	+	+	++
Temporal Context	✓	?				

Table 8.1: Candidate circuits (Input-Memory-Output) for implementing episodic memory using a two-layer associative memory within the hippocampus. Capacity refers to the number of possible combinations of active neurons in the memory population. Neurogenesis and contextual modulation also refer to the memory population layer. The residual connection is a way to convey target information to the output layer. Backprojection from the output to the input or to the memory, and temporal context formation in the heteroassociative input, are both necessary for contextual sequence learning. Finally, contextual modulation refers to the ability for prefrontal information to bias retrieval from the memory. According to these criteria, the EC_{II}-DG-CA3 circuit (bold) is the best candidate for implementing episodic memory with a modern associative memory in the hippocampus.

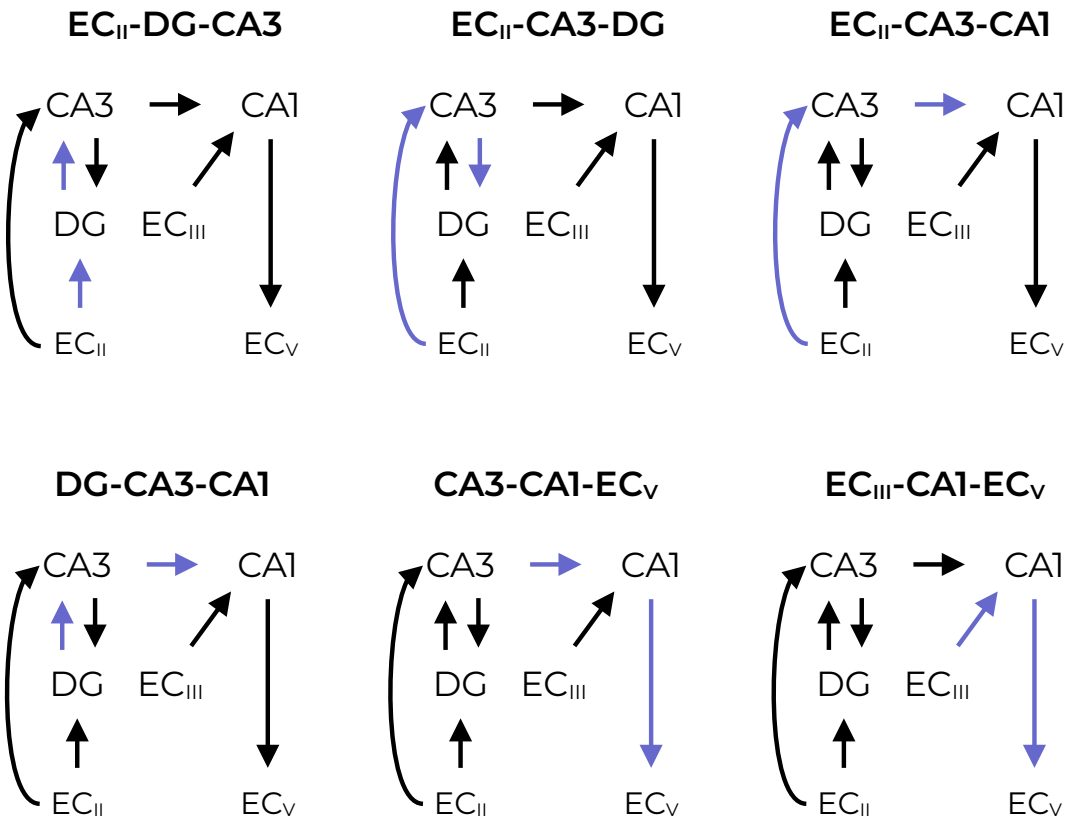


Figure 8.1: Candidate circuits (Input-Memory-Output) for implementing episodic memory using a two-layer associative memory within the hippocampus. Associative memory connections are shown in blue.

dentate gyrus to implement a key-value associative memory which learns to associate the current state of the episode to its successor. Like in the Socratic model, the CA3-to-DG backprojection learns heteroassociations while the EC_{II}-to-DG connection can implement autoassociations. Inspired by the sparsity of granule cells, the activation function of DG is used to separate memories. The resulting DG activation is projected back to CA3 with a value matrix that learns to reproduce EC_{II}-driven patterns of CA3 activity. By predicting successive successors, Consequence can recall complete sequences of experienced states.

This new model will be compared to the standard framework of sequence learning in the hippocampus (Levy 1996), which will be called the Collateral model in reference to the role of CA3 recurrent connections. Consequence will also be compared to the Socratic model (Lisman, Talamini, et al. 2005). In order to encode sequences of complex information, these models are embedded within a deep learning network of the whole hippocampus which learns in an autoencoding fashion. As this network learns information slowly, like semantic memory in the complementary learning systems theory (Section 3.3.1), it is called the semantic pathway. It should be noted that this name serves as a conceptual shortcut rather than a formal technical term. Each model can be divided into an encoding and decoding semantic pathway, and a fast learning episodic pathway (Figure 8.2).

8.3.1 Semantic Pathways

The semantic pathway of all models is a deep learning neural network trained to encode and decode information for episodic storage and retrieval in the episodic pathway, with a learning rate of 2×10^{-4}. The number of neurons in each region is set as the number of neurons in the analog rat region, as retrieved from Cutsuridis et al. (2019), scaled down by a factor $N = 5 \times 10^{-4}$ (Table 7.1). The activation function is the softmax function in DG (for comparison with Consequence; see Section 8.3.2) and ReLU in other layers. During training, information flows from Input to Output.

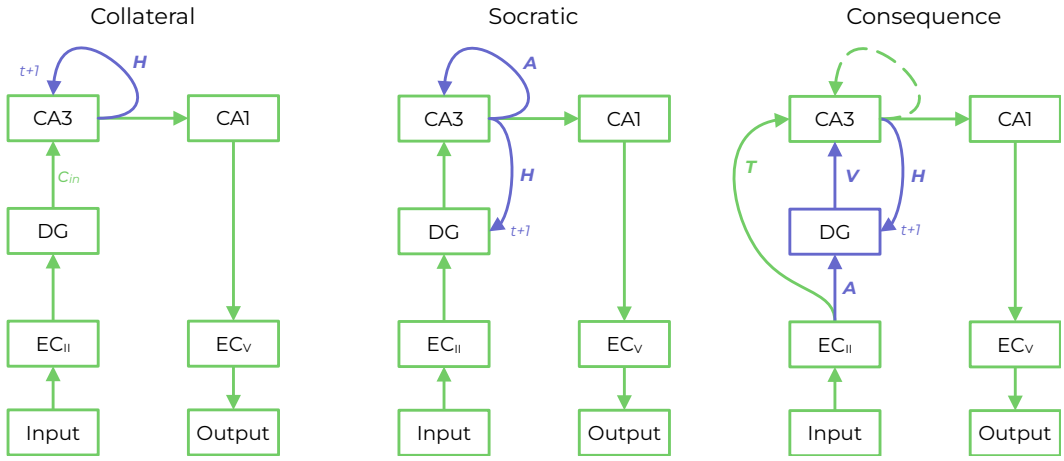


Figure 8.2: Candidate models of sequence learning. In all three models, the semantic pathway (green) develops by minimizing the discrepancy between Input(t) and its reconstruction Output(t) through backpropagation. The fast episodic pathway (blue) learns to recall sequences of CA3 activity. Left: architecture of the Collateral model. This episodic network consists of a single heteroassociative CA3-to-CA3 connection $\boldsymbol{H}$, the target of which is the CA3 pattern of the next time step. The density of DG-to-CA3 connections is determined by probability c_{in}, as partial connectivity was proposed as a mechanism for history-dependent modulation of sequence recall in the Collateral model Levy (1996). Center: architecture of the Socratic model. In this candidate, the CA3-to-CA3 connection $\boldsymbol{A}$ is autoassociative and is used to complete the current pattern of CA3. Heteroassociations are performed by the CA3-to-DG backprojection $\boldsymbol{H}$, the goal of which is to reconstruct the next DG pattern. The recalled pattern is then "cleaned up" using $\boldsymbol{A}$ in order to avoid the concatenation of heteroassociation errors. Right: architecture of the Consequence sequence learner. In addition to using the same heteroassociative mechanism as Socratic, the episodic network of Consequence employs autoassociative EC_{II}-to-DG ($\boldsymbol{A}$) and DG-to-CA3 ($\boldsymbol{V}$) projections. $\boldsymbol{V}$ is analog to the value matrix of a key-value memory network, in that it is used to decode memory patterns, which can come from the environment through EC_{II} (pattern completion) or from CA3 (sequence completion). Instead of flowing through DG, information in the semantic network goes from EC_{II} directly to CA3 through matrix $\boldsymbol{T}$. This skip connection is crucial for training the episodic network of Consequence, as $\boldsymbol{V}$ learns to mimick the output of connection $\boldsymbol{T}$ which serves as a target. Since the CA3-to-CA3 recurrence is not necessary for simple sequence learning in the Consequence model, multiple implementations are explored for enabling history-dependent sequence recall. In all models, the semantic network is used to reconstruct Output from recalled CA3 patterns.

A reconstruction error is computed and used to update connections of the network using the backpropagation algorithm (Rumelhart et al. 1986; Linnainmaa 1976). The semantic pathways of the candidate models differ in three ways. First, in the Collateral model, DG does not fully connect to CA3. A Bernoulli sampling process is used such that each CA3 neuron has probability c_{in} to receive connections from DG, while the entries of the DG-to-CA3 weight matrix corresponding to nonselected CA3 neurons are zeroed at each forward propagation. This has been proposed to enable contextual modulation of sequence learning, as neurons not receiving external input but connected to other CA3 neurons through the recurrent connections encode contextual representations (Levy 1996). Secondly, the CA3-to-CA3 recurrence of CONSEQUENCE is not part of the episodic pathway like in other models. Memories in this model are instead encoded in the dentate gyrus. This leaves room for other kinds of processing. In fact, it will become clear later that recurrent collaterals of CA3 can be dedicated to encoding a history of past states to disambiguate sequences with common bouts like the model of Levy (1996). The last difference is the skip connection $\boldsymbol{T}$ of CONSEQUENCE. The dentate gyrus is indeed bypassed and only used by the episodic network. This is done for three reasons. First, while this connection is not always ascribed to a specific function in computational models of the hippocampus, it is present in the biology as reviewed earlier. Secondly, it is expected that error gradients flow better from CA3 directly back to EC_{II} than through DG if the activity of the latter is sparse with strong lateral inhibition. While DG activity profile has been proposed to subserve functions like pattern separation, it is not easily included in deep learning architectures. For example, if the activation function is set to a softmax with high steepness β, the window of nonzero derivative is small as examplified in Figure 8.3, which can lead to a problem analog to vanishing gradient (Bengio et al. 1994). Using a softmax as the activation function for DG therefore makes training upstream layers harder. By skipping DG in the semantic pathway, we can allow it to be very sparse for efficient episodic processing without affecting backpropagation training. Last but

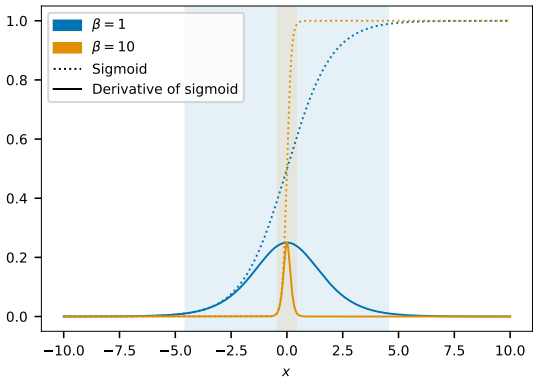


Figure 8.3: Gradient of the sigmoid function vanishes with steepness β. Area where gradient falls under $\epsilon = 1\text{e}{-2}$ is highlighted for each β value.

not least, the skip connection $\boldsymbol{T}$ acts as a target for teaching the DG-to-CA3 projection of the episodic pathway, as detailed below.

8.3.2 Fast Episodic Pathway of Consequence

In the Consequence model, DG implements a key-value memory network with two sets of keys (autoassociative keys $\boldsymbol{A}$ and heteroassociative keys $\boldsymbol{H}$) and a set of values $\boldsymbol{V}$ which are learned using the biologically plausible rules proposed by Tyulmankov et al. (2021). In order to write keys, a local third factor is sequentially active for each postsynaptic DG neuron j:

$$\gamma_j(t) = \begin{cases} 1 & \text{if } t = j \bmod N_{\text{DG}} \\ 0 & \text{otherwise} \end{cases} \tag{8.1}$$

where N_{DG} is the number of DG neurons. This factor is combined to a scalar global learning rate η_e for determining the learning rate of each synapse:

$$\eta^A_{ij} = \eta^H_{ij} = \eta_e \gamma_j, \tag{8.2}$$

which is then used to update $\boldsymbol{A}$ and $\boldsymbol{H}$ weights:

$$\boldsymbol{A}(t+1) = (1 - \boldsymbol{\eta}^A) \odot \boldsymbol{A}(t) + \boldsymbol{\eta}^A \odot (\mathrm{EC}_{\mathrm{II}}(t)\mathbf{1}) \tag{8.3}$$

$$\boldsymbol{H}(t+1) = (1 - \boldsymbol{\eta}^H) \odot \boldsymbol{H}(t) + \boldsymbol{\eta}^H \odot (\mathrm{CA3}(t-1)\mathbf{1}) \tag{8.4}$$

where $\mathbf{1} \equiv (1, 1, ..., 1)$. Note that vectors are regularized with the L2 norm before encoding and retrieving them to ensure there is no bias in favor of certain memories. This is done for the other candidate models too. After storing the normalized keys, the new weights are used to compute the state of DG:

$$\mathrm{DG}^{in}(t) = \begin{cases} \boldsymbol{A}^\top(t+1)\mathrm{EC}_{\mathrm{II}}(t) & \text{for encoding and pattern recall} \\ \boldsymbol{H}^\top(t+1)\mathrm{CA3}(t-1) & \text{for sequence recall} \end{cases} \tag{8.5}$$

$$\mathrm{DG}(t) = \mathrm{softmax}(\beta_{\mathrm{DG}}\mathrm{DG}^{in}(t)) \tag{8.6}$$

Just as in some modern Hopfield networks (Ramsauer et al. 2020), β_{DG} is the inverse temperature controlling the amount of memory blending. As β_{DG} approaches infinity, DG activity sparsifies since the softmax gets closer and closer to a max function giving a result resembling a one-hot vector (i.e. with one entry being one and the others being zero), given that all input entries are different. The predicted pattern is then the selected column of $\boldsymbol{V}$. Conversely with $\beta_{\mathrm{DG}} = 0$, the output of the softmax is a uniform distribution that results in a prediction that is an average of all values. During encoding, the state of DG is then used to update the value matrix:

$$\boldsymbol{V}(t+1) = (1 - \eta_e \mathrm{DG}(t)\mathbf{1}) \odot \boldsymbol{A}(t) + \eta_e \mathrm{DG}(\boldsymbol{T}^\top \mathrm{EC}_{\mathrm{II}}(t))^\top \tag{8.7}$$

$$\tag{8.8}$$

where $\boldsymbol{T}^\top \mathrm{EC}_{\mathrm{II}}(t)$ is the input from $\mathrm{EC}_{\mathrm{II}}$ to CA3 which serves as a target matrix for $\boldsymbol{V}$. In other words, $\boldsymbol{V}$ learns to replace the input from the environment for recalling CA3 states.

Depending on the processing phase, the input to CA3 either comes from the entorhinal cortex or DG recall:

$$\mathrm{CA3}^{in}(t) = \begin{cases} \boldsymbol{V}^\top \mathrm{DG}(t) & \text{during recall} \\ \boldsymbol{T}^\top \mathrm{EC}_{\mathrm{II}}(t) & \text{during encoding.} \end{cases} \tag{8.9}$$

The state of CA3 is computed as:

$$\mathrm{CA3}(t) = \mathcal{R}\big(\mathrm{CA3}^{in}(t)\big) \tag{8.10}$$

where $\mathcal{R}$ is the recurrence function of CA3 which will be described in Section 8.3.5.

Once the network has been trained and memories have been stored, Equations 8.6 and 8.10 can be chained to recall a static pattern (using entorhinal input to DG) or a sequence transition (using the CA3-to-DG backprojection). Full sequences can be retrieved by computing Equation 8.10 multiple times. The reinstantiated CA3 states can be decoded using the semantic pathway in order to recover the sequence of incoming sensory (or cortical) stimuli. Now that the Consequence model has been described, the episodic pathway of the other candidate models will be outlined.

8.3.3 Fast Episodic Pathway of the Collateral Model

The first candidate to which Consequence is compared is a model inspired by the standard framework. In the standard framework, memories are implemented in CA3 recurrent collaterals. While this connection is often proposed to encode autoassociations, this does not allow for the storage of sequences. Here, the CA3-to-CA3 connection encodes heteroassociations like in the model of Levy (1996): it learns to reconstruct $\mathrm{CA3}(t+1)$ from $\mathrm{CA3}(t)$. To this end, heteroassociative weights are updated with a postsynaptic rule (Levy 1982):

$$H_{ij}(t) = H_{ij}(t-1) + \eta_e \mathrm{CA3}_j(t)\big[\mathrm{CA3}_i(t-1) - H_{ij}(t-1)\big] \tag{8.11}$$

where H_{ij} is the synaptic weight from neuron i to j and η_e is the rate of episodic learning. The activity of CA3 is:

$$\mathrm{CA3}(t) = \mathrm{ReLU}\Big(\big[(\boldsymbol{M}_{in} \odot \boldsymbol{W})^\top \mathrm{DG}(t) + (\boldsymbol{M}_r \odot \boldsymbol{H})^\top \mathrm{CA3}(t-1)\big]\Big) \quad (8.12)$$

where $\boldsymbol{M}_{in}$ and $\boldsymbol{M}_r$ are connection masks, $\odot$ is the Hadamard (element-wise) product and $\boldsymbol{W}$ is the DG-to-CA3 weight matrix. While $\boldsymbol{M}_{in}$ is the result of the neuron-wise sampling procedure described in section 8.3.1, the recurrent mask $\boldsymbol{M}_r$ is constructed with a synaptic-wise process. Each entry has probability c_r of being 1 and is set to 0 otherwise. Note that contrary to other models, the episodic pathway contributes to information processing even in the pretraining phase in which the semantic pathway develops. Weights $\boldsymbol{H}$ are thus initialized with nonzero values and left unchanged until memories are encoded in the model. This is necessary to train the semantic network to cope with CA3 neurons that do not receive external inputs (i.e. $\mathrm{col}_j(\boldsymbol{M}_{in}) = \vec{0}$) and would otherwise only become active once the semantic pathway is trained and memories are encoded.

8.3.4 Fast Episodic Pathway of Socratic

The Socratic model employs two associative connections of different kinds. The CA3-to-DG backprojection learns to predict $\mathrm{DG}(t+1)$ from $\mathrm{CA3}(t)$ using Equation 8.4 of Consequence with a learning rate $\eta^H_{ij} = \eta_{e_1}\gamma_j$ (to distinguish η_{e_1} from the CA3-to-CA3 rate η_{e_2} introduced below).

The second associative connection is the CA3-to-CA3 recurrence which is used to encode autoassociations and cleanup the result of the heteroassociation in order to prevent the concatenation of errors. It is updated similarly to the connection of the Collateral model:

$$A_{ij}(t) = A_{ij}(t-1) + \eta_{e_2}\mathrm{CA3}_j(t)\big[\mathrm{CA3}_i(t) - A_{ij}(t)\big] \quad (8.13)$$

with a second episodic learning rate η_{e_2}.

Unlike the Collateral model, the episodic network of Socratic needs not be active during the training of the semantic pathway. Thus, the activity of

DG is governed by external inputs during training and for the recall of a single autoassociation, and driven by CA3 during heteroassociative recall:

$$\mathrm{DG}^{in}(t) = \begin{cases} \boldsymbol{W}_{\mathrm{EC_{II}}\to\mathrm{DG}}^{\top}\mathrm{EC_{II}}(t) & \text{for encoding and pattern recall} \\ \boldsymbol{H}^{\top}\mathrm{CA3}(t-1) & \text{for sequence recall} \end{cases} \tag{8.14}$$

$$\mathrm{DG}(t) = \mathrm{softmax}(\beta_{\mathrm{DG}}\mathrm{DG}^{in}) \tag{8.15}$$

Similarly, the autoassociative recurrence is only used during recall:

$$\mathrm{CA3}^{in}(t) = \boldsymbol{W}_{\mathrm{DG}\to\mathrm{CA3}}^{\top}\mathrm{DG}(t) \tag{8.16}$$

$$\mathrm{CA3}(t) = \begin{cases} \mathrm{ReLU}(\mathrm{CA3}^{in}(t)) & \text{for encoding} \\ \mathrm{ReLU}(\boldsymbol{A}^{\top}\mathrm{CA3}^{in}(t) + \mathrm{CA3}^{in}(t)) & \text{for recall} \end{cases} \tag{8.17}$$

$$\tag{8.18}$$

8.3.5 Temporal Context Formation in Consequence

With discrete sequence learning as a primary function, the hippocampus must learn how the input states follow one another. A naive approach would be to learn a set of associations between current states and subsequent states. However, this approach fails to account for the learning of temporal dependencies. Therefore, the naive approach is incapable of disambiguating sequences that share the same states.

A common solution to this problem is to use a representation of the temporal context as additional information to make a prediction about the next state. In neural networks, this is typically implemented with recurrent connections that convey information over time (Elman 1990). With its recurrent collaterals, CA3 is well suited to play this role, and the recurrence function $\mathcal{R}$ of Equation 8.10 can be:

$$\mathcal{R}(\boldsymbol{x}_t) = \tanh(\boldsymbol{x}_t + \boldsymbol{R}^{\top}\mathrm{CA3}_{t-1} + \boldsymbol{b}) \tag{8.19}$$

where $\boldsymbol{R}$ are recurrent weights and $\boldsymbol{b}$ is the bias vector. A simpler alternative is to model CA3 as a leaky feedforward layer:

$$\mathcal{R}(\boldsymbol{x_t}) = \alpha \text{ReLU}(\boldsymbol{x}_t + \boldsymbol{b}) + (1 - \alpha)\text{CA3}_{t-1} \tag{8.20}$$

where $\alpha \in [0, 1]$ is the leak rate. Both leaky and recurrent weights versions of CA3 are initialized without recurrence in the first step of each sequence. In principle, these two recurrent functions could enable the model to disambiguate overlapping sequences by taking past information into account for episodic storage and retrieval.

8.4 Methods

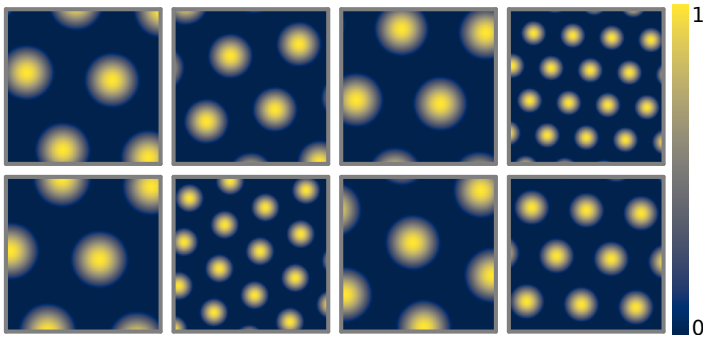
8.4.1 Datasets

Two datasets are used throughout this study. The binary dataset consists of sequences of random binary patterns. Each bit is either 1 with probability .35 and 0 with probability .65, as motivated by average activity levels of the entorhinal cortex reported by Neher et al. (2015). When using this dataset, the binary cross-entropy is used to train the semantic pathway.

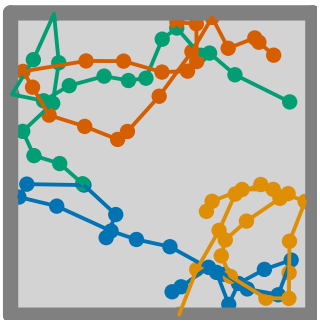
The spatial dataset consists of sequences of spatial positions represented by 30 cells with grid activity profiles. The sequences and grid cell responses are generated with the RatInABox Python library (George, Cothi, Clopath, et al. 2022). The firing of a grid cell is modelled as:

$$F_i(t) = \frac{1}{3}\max\left(0, \sum_{j=0}^{2} \cos\left(2\pi \frac{\boldsymbol{x}(t) \cdot \boldsymbol{e}_{\theta_i + j\pi/3}}{\lambda_i} + \phi_i\right)\right) \tag{8.21}$$

where $\phi_i \in [0, 2\pi]$ and $\theta_i \in [0, 2\pi]$ are the phase and orientation respectively, λ_i is the scale, $\boldsymbol{x}(t)$ is the current position and $\boldsymbol{e}_\theta$ is the unit vector pointing in the direction θ. Phases, orientations and scales are sampled uniformly. Example grid fields and trajectories generated with RatInABox



(a) Response profile of eight example grid cells with varying scale, orientation and phase.



(b) Four example sequences of 20 positions.

Figure 8.4: Examples of biologically realistic data generated with RatInABox (George, Cothi, Clopath, et al. 2022).

are shown in Figure 8.4. The mean square error loss is used to train the semantic pathway with this dataset.

8.4.2 Hyperparameter Optimization

The models have some hyperparameters that need to be tuned for fair comparison. These include β_{DG}, η_e and α for Consequence, β_{DG}, η_{e1} and η_{e2} for Socratic, and β_{DG}, c_{in}, c_r and η_e for the Collateral model. The parameters are thus optimized using Gaussian processes with the scikit-optimize library (Head et al. 2021). First, 100 parameter sets are generated with a Halton sequence (Halton 1960), then tested. After this initial phase, 100 additional evaluations are made guided by the Bayesian surrogate model. Each evaluation is made with 3 different runs.

8.5 Experiments

8.5.1 One-shot Sequence Learning

Here the ability of the candidate models to learn sequences in one-shot is tested. In the pretraining phase, the semantic pathways are trained to auto-encode 1000 independent spatial positions or binary patterns during

50 epochs. Grid scales are uniformly sampled between 0.1 and 0.5. The spatial environment is an empty 1×1 meter box. Then, 25 sequences of 20 steps are presented once for encoding in the episodic pathways. Spatial sequences are initialized and generated randomly. After this one-shot exposure phase, the ability of the models to recall the sequences is tested. In order to do that, the first half of each sequence is given as a cue, during which the episodic pathway is silenced. Then, the episodic pathway is used to recall the second half of the sequence in absence of additional external input. Performance is calculated as the R^2 score of the regression between the recalled and actual second half of the sequence. Similarly, the hyperparameter optimization process seeks to minimize the reconstruction loss of the second half of the sequences. Note that performance and loss in the spatial dataset are computed in the grid cell space. Here, the simple leaky recurrent function was used in CONSEQUENCE.

The results of the hyperparameter optimization procedures are shown in Figures 8.5, 8.6, 8.7, 8.8, 8.9 and 8.10. For CONSEQUENCE, the most important parameter is β_{DG} in that it influences performance the most (Figures 8.5 and 8.6). It must be large to properly separate memories during retrieval. The optimization of η_e reveals a slightly more complex pattern. At first sight, it seems that the model works best with slower episodic learning rates, which is at odds with the one-shot learning regime. However, when β_{DG} is large, the best performance is obtained with fast learning rates. Furthermore, CONSEQUENCE performs best with leak rate $\alpha = 1$. When it comes to the SOCRATIC model, η_{e_2} (learning rate of CA3 recurrent collaterals) and β_{DG} are the most important, even though the model did not perform well overall (Figures 8.7 and 8.8). Finally, the Collateral model was most influenced by the density of recurrent weights c_r, but did not perform as well as CONSEQUENCE either (Figures 8.9 and 8.10). Performance of the three optimized models is shown in Figure 8.11. CONSEQUENCE performs best in both spatial and binary datasets. In comparison, the performance of the collateral and SOCRATIC models did not reach baseline despite hyperparameter optimization. In order to test the robustness of CONSEQUENCE, Gaussian noise is added to the recall cues. Since the sequences have values

between 0 and 1, a Gaussian noise of $\xi = 1$ is very strong. The model nonetheless performs well (Figure 8.12).

8.5.2 Place Cells and Remapping

Now that Consequence has been shown to perform well, it is worth investigating how biologically plausible the representations it employs are. When spatial sequences are encoded in the model, the response profile of recruited dentate gyrus cells resembles that of hippocampal place cells (Figure 8.13a; left). When placed in a novel environment (an adjacent box of equal shape) in which no sequences were stored, the place field of dentate gyrus cells remaps (Figure 8.13a; right). It has been shown that remapping in the biological hippocampus reflects attractor dynamics, such that continuously morphing between a circle and a square environment results in abrupt and simultaneous changes in the response profile of place cells (Wills et al. 2005; Figure 8.14). In order to test whether this can occur in the dentate gyrus of Consequence, it is assumed that the shape of the environment is (at least partially) represented in external regions and provided as input to the hippocampus. In the model, this is done by providing semantic information through 30 lateral entorhinal cortex cells, in addition to the encoding of space of the 30 grid cells. Semantic vectors A and B are created independently, such that $A_i, B_i \sim U(-1, 1)$, with $i = 1, 2, ..., 30$. Prior knowledge of these vectors in the first environment is encoded by associating either A or B to each pretraining position, and to each position of the sequences learned. Then, the response profile of dentate gyrus is constructed when morphing between vectors A and B in the novel adjacent environment (Figure 8.13b). Neurons of the Consequence dentate gyrus remap abruptly and synchronously as a result of morphing. Similarity analyses are also reminiscent of attractor dynamics observed by Wills et al. (2005), as shown in Figure 8.15.

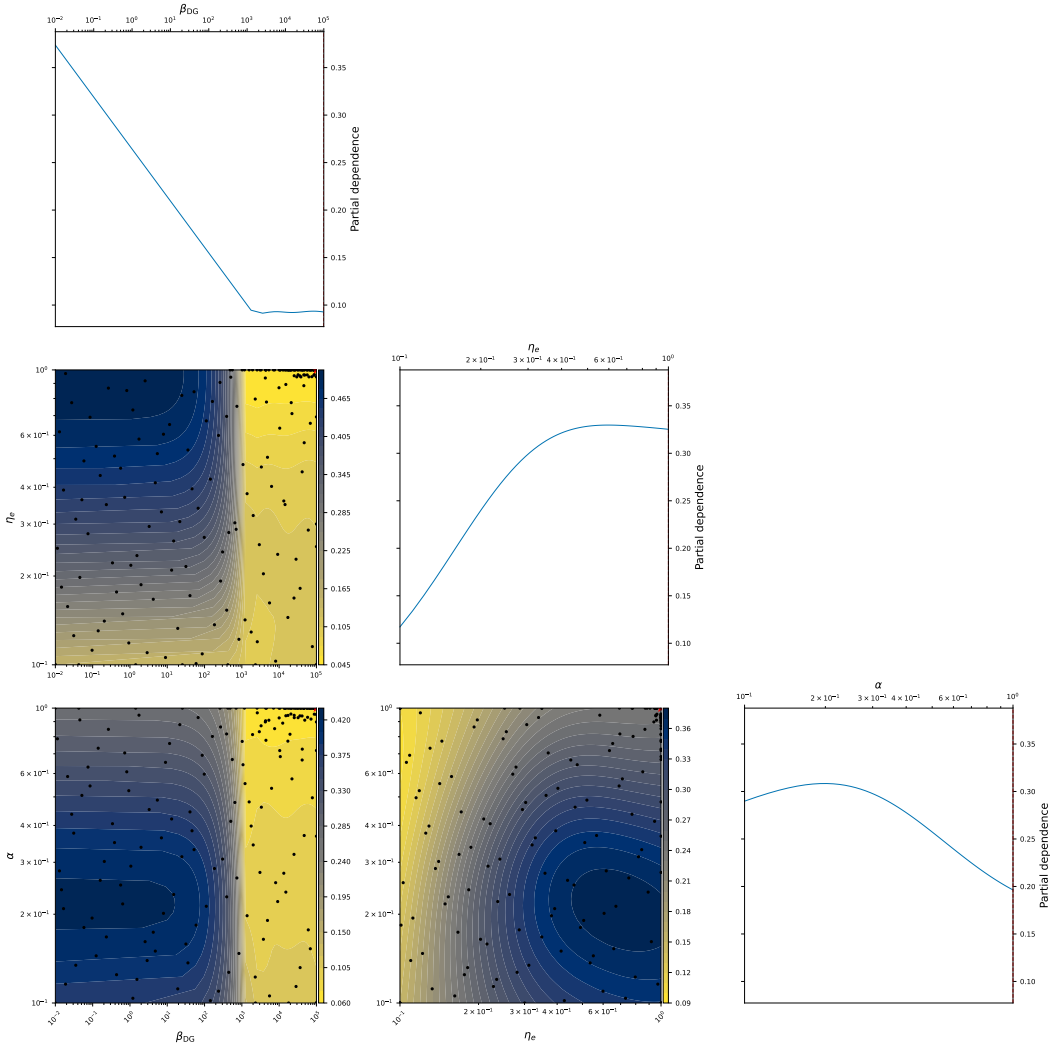


Figure 8.5: Results of hyperparameter optimization of the CONSEQUENCE model with the spatial dataset. Partial dependence plots show how individual hyperparameters influence the loss function by marginalizing over the values of all other parameters. Similarly, contour plots show interactions of parameter pairs by marginalizing over the remaining variables.

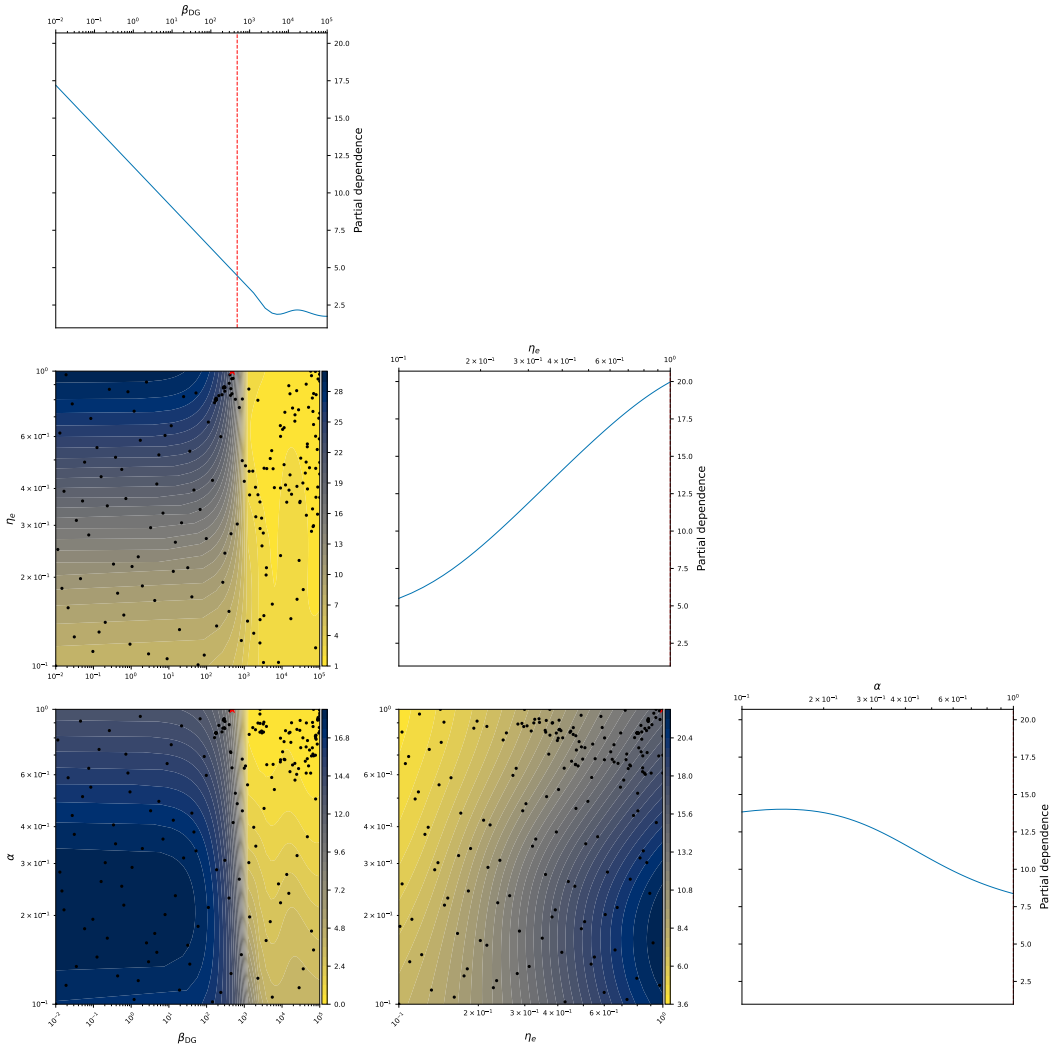


Figure 8.6: Results of hyperparameter optimization of the CONSEQUENCE model with the binary dataset. See Figure 8.5 for details regarding the plots.

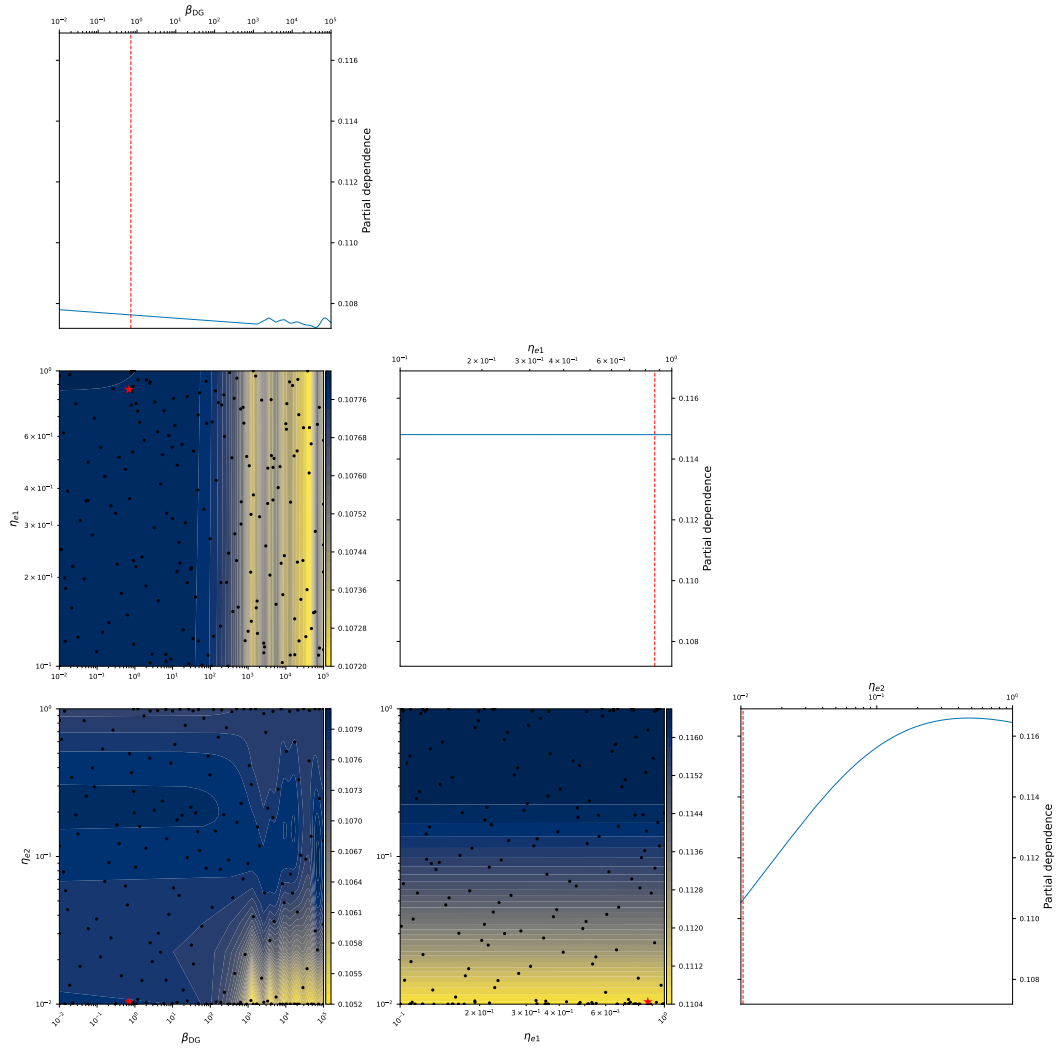


Figure 8.7: Results of hyperparameter optimization of the SOCRATIC model with the spatial dataset. See Figure 8.5 for details regarding the plots.

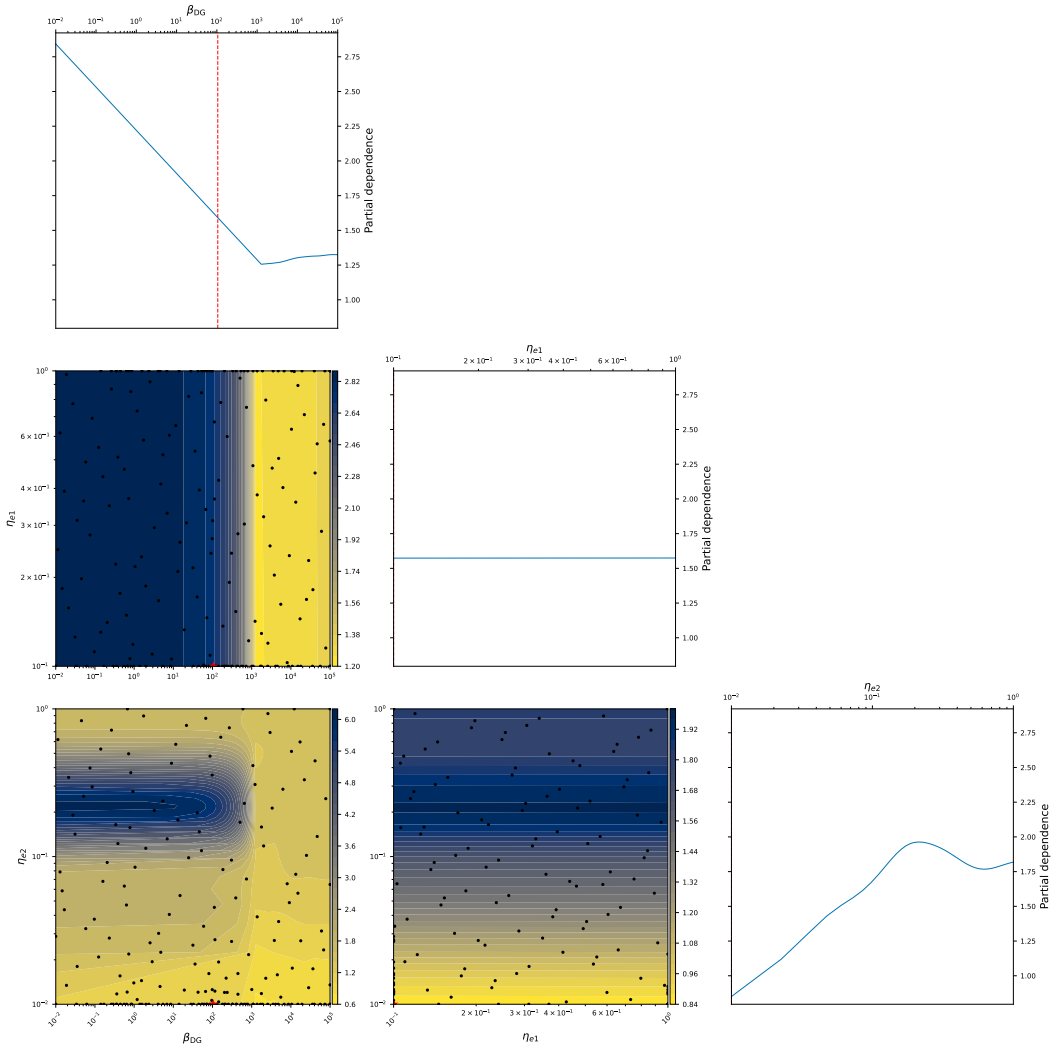


Figure 8.8: Results of hyperparameter optimization of the SOCRATIC model with the binary dataset. See Figure 8.5 for details regarding the plots.

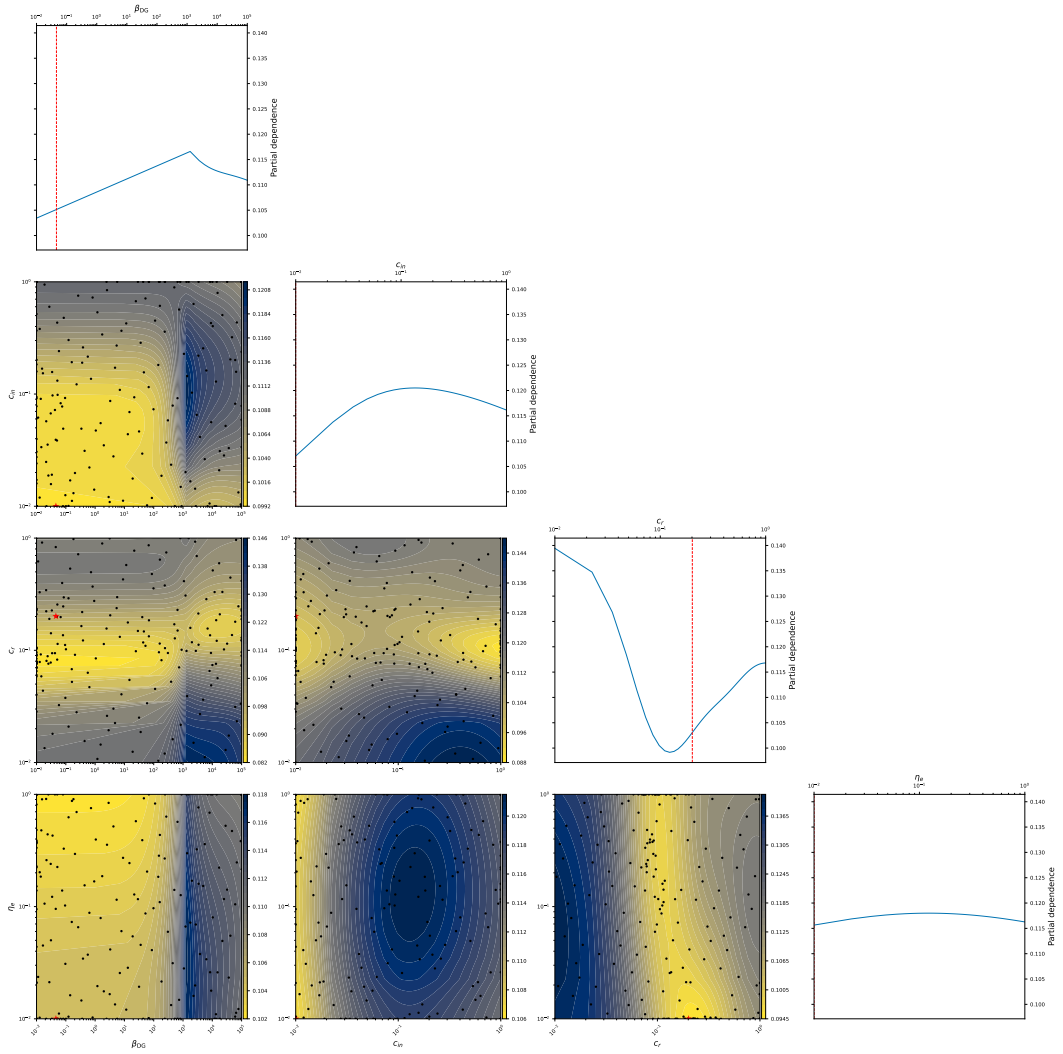


Figure 8.9: Results of hyperparameter optimization of the Collateral model with the spatial dataset. See Figure 8.5 for details regarding the plots.

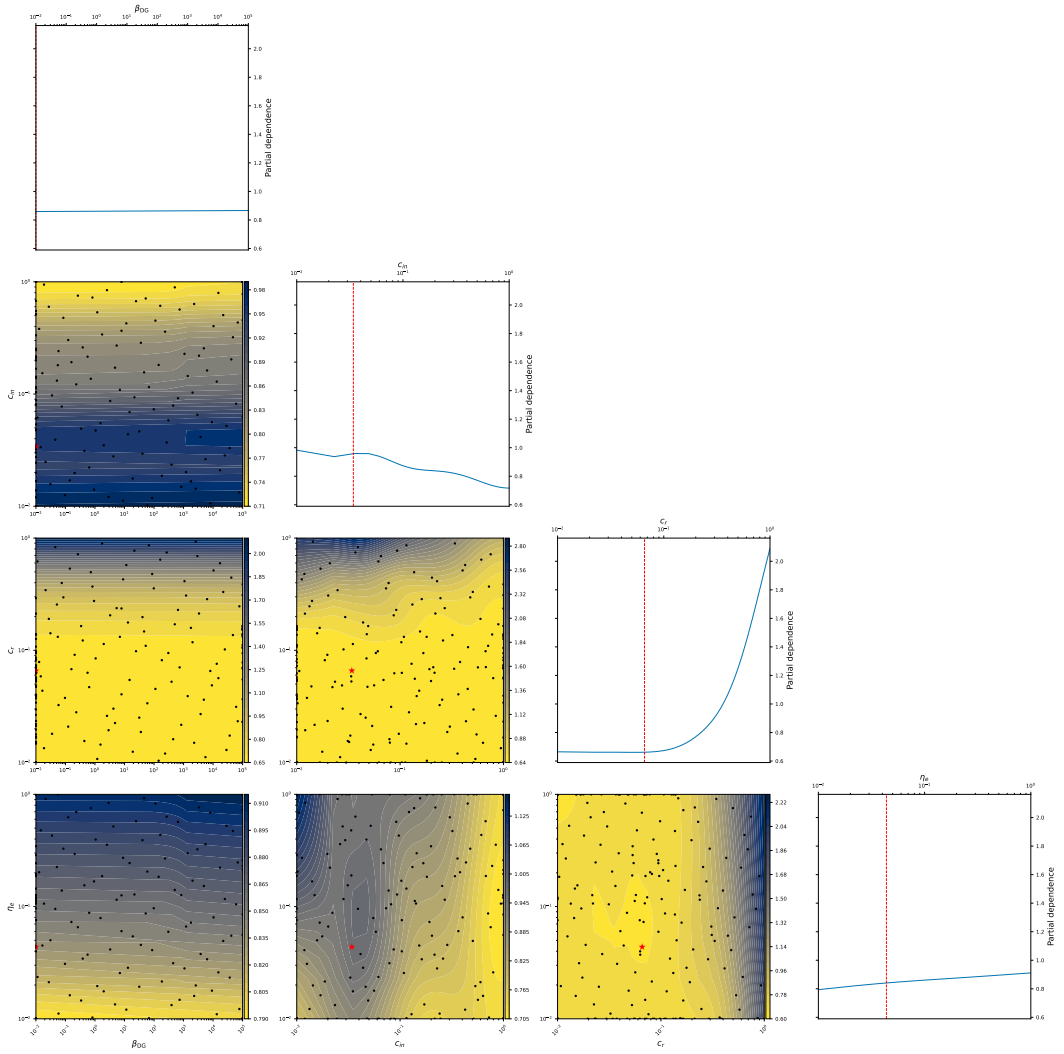
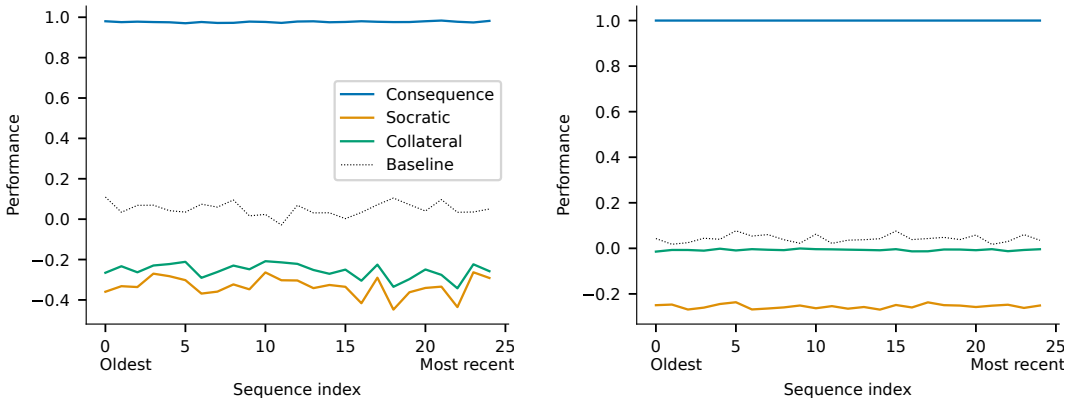
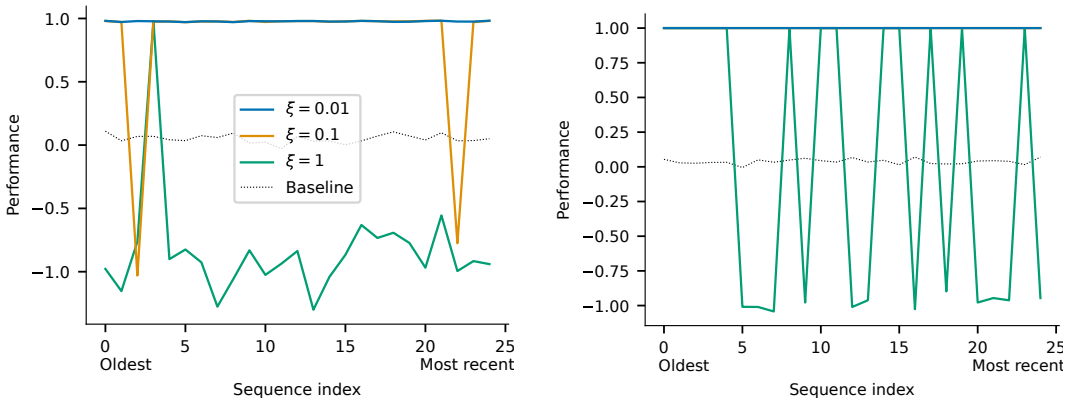


Figure 8.10: Results of hyperparameter optimization of the Collateral model with the binary dataset. See Figure 8.5 for details regarding the plots.



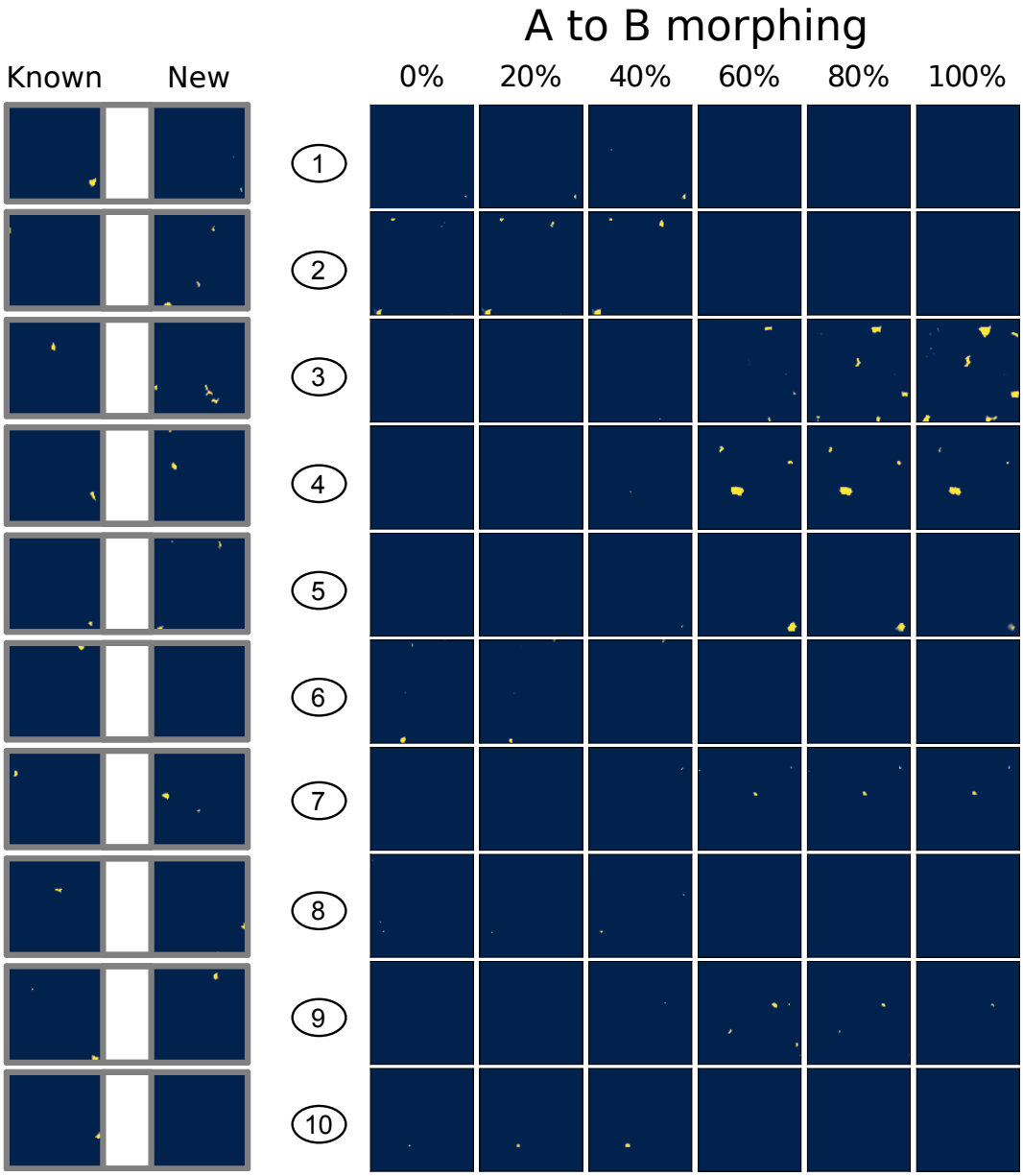
(a) Spatial dataset (b) Binary dataset

Figure 8.11: Performance of optimized candidate models. Performance of a model that would always output the average of the dataset is shown as baseline.



(a) Spatial dataset (b) Binary dataset

Figure 8.12: Performance of CONSEQUENCE with various levels of cue noise. Gaussian noise of standard deviation ξ is added to the input during recall.



(a) Response of example dentate gyrus cells in known and new environments.

(b) Response of dentate gyrus cells in new environment when morphing between A and B.

Figure 8.13: Place response profiles of dentate gyrus cells of the CONSEQUENCE model.

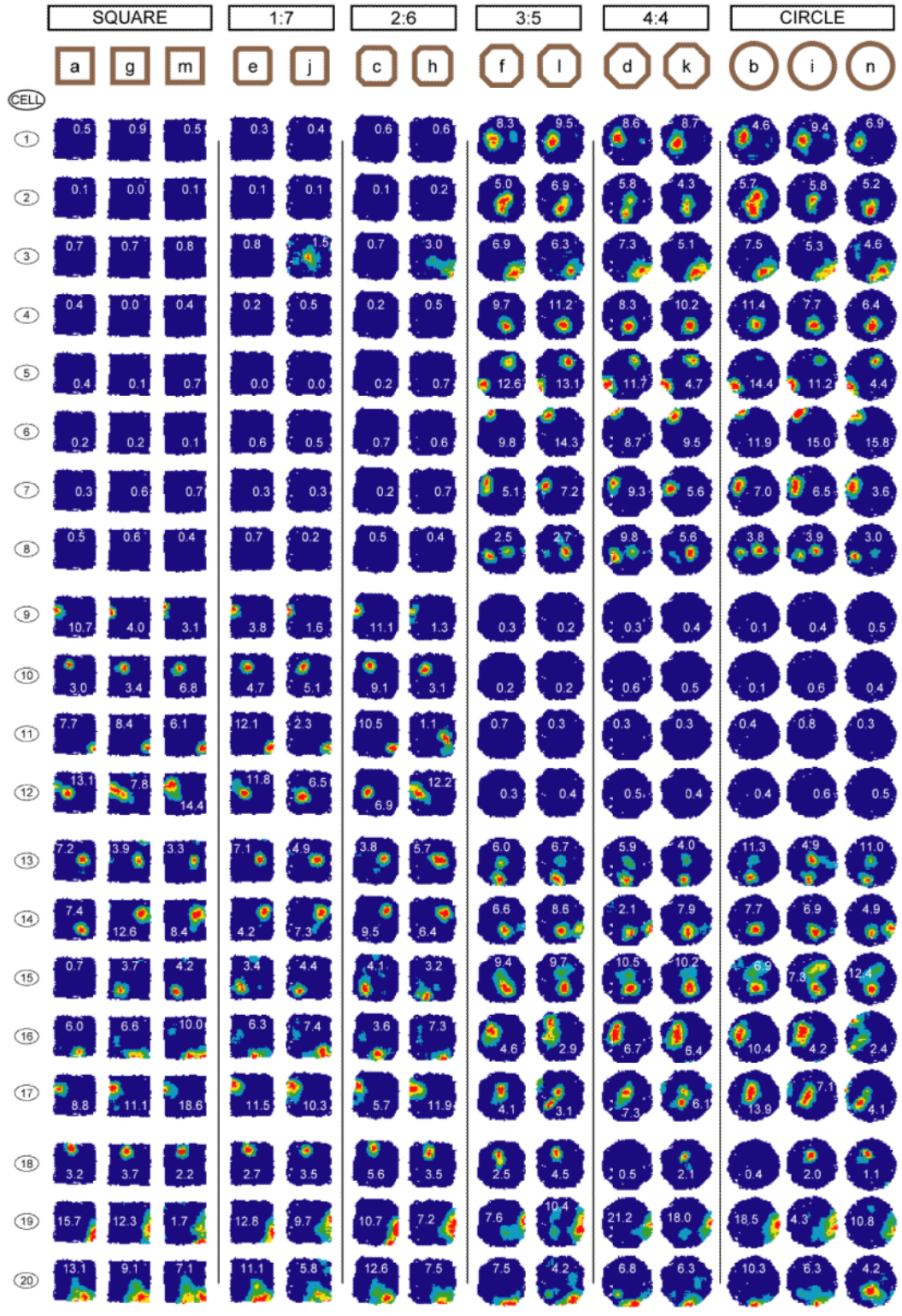
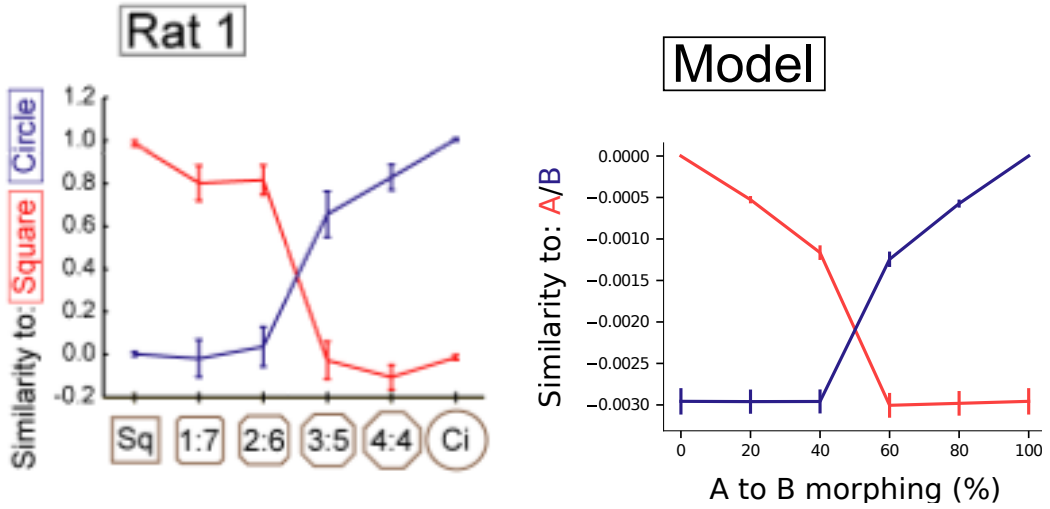


Figure 8.14: Abrupt and coherent expression of square or circle representation during probe trials in intermediate octagonal environments in Rat 4. The 17 of 20 place cells simultaneously recorded from rat 4 with different (remapped) firing patterns in the square and the circle almost all switch from the square-like to circle-like pattern between the 2:6 and 3:5 octagons. 8 cells had fields in the circle but not the square (cells 1-8); 4 in the square but not the circle (9-12); 5 fired in both but in different places (13-17), and 3 did not reach criterion for remapping (18-20). Taken from Wills et al. (2005).



(a) Coordinated shift in square-to-circle switch point between the first and second octagon series. The plot shows the similarity of place cells' firing patterns in probe trials of varying shape to their firing patterns in square (red) or circle (blue) baseline trials (mean and standard error across cells). Taken from Wills et al. (2005).

(b) Similarity between dentate place fields of CONSEQUENCE during morphing and baseline trial of A (red) or B (blue). Similarity is computed as the negative mean squared error. Plotted are the mean and standard error across cells.

Figure 8.15: Analysis of place fields during morphing in the biological and simulated hippocampus.

8.5.3 Disambiguation of Overlapping Sequences

The sequences encoded so far were not generated to overlap, so there is no need to keep track of past states such that a leak rate of $\alpha = 1$ was found to work best (Figures 8.5 and 8.6). However, rodents can also perform alternation tasks in which the hippocampus encodes the recent past, as suggested by the presence of retrospective splitter cells (Wood et al. 2000; L. M. Frank et al. 2000), to predict the location of future rewards. In order to model this process with Consequence, trajectories of 10 positions are generated in an eight-shaped maze. 50 trajectories start at the bottom left and end at the top right, and 50 trajectories start at the bottom right and end at the top left (Figure 8.16; left). The number of DG neurons is multiplied by 1.7 to ensure that the episodic pathway has sufficient capacity to encode these sequences. Importantly, all trajectories pass through a common arm and meet in the center before separating. This is done to prevent the model from predicting future positions based on the current state. For accurate prediction of the second half of the sequences, the model must keep a trace of where it comes from.

The model is pretrained with positions spanning the environment as described before. Grid scales are uniformly sampled between 0.5 and 1. A Gaussian process regression model with radial basis function kernel is trained to decode the positions retrieved by the model[1]. Example decoded sequences are shown in Figure 8.16 (right).

Figure 8.17 shows the performance of the model in this task. With $\alpha = 0.5$, the model ends its recall in the correct arm in 95% and 72% of trials with a cue noise of $\xi = 0.1$ and $\xi = 0.3$ respectively. Performance approaches 50% with higher α as the model looses track of where it comes from, and approaches 0% with $\alpha < 0.2$ because recall does not reach any upper arm if CA3 is updating too slowly. Performance is reflected in the separation of sequences in principal component space (Figure 8.18). While the center is encoded differently according to the starting arm with $\alpha = 0.5$

[1] This process is detailed in https://github.com/RatInABox-Lab/RatInABox/blob/main/demos/decoding_position_example.ipynb

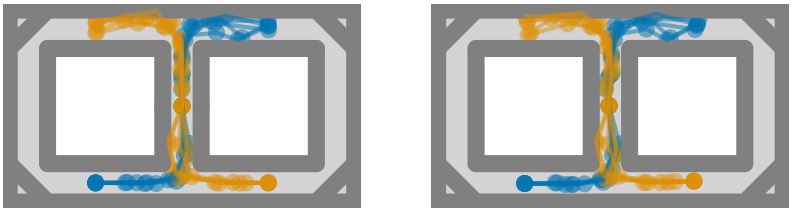


Figure 8.16: 10 example actual (left) and decoded (right) trajectories of 10 positions generated in the eight maze with RatInABox (George, Cothi, Clopath, et al. 2022). Blue trajectories start in the lower left arm and end in the upper right arm. Orange trajectories start in the lower right arm and end in the upper left arm. All trajectories meet at the center before diverging.

(Figure 8.18d), it is not separated with $\alpha = 1$ (Figure 8.18b). In comparison, the average performance of the model with recurrent CA3 weights (Equation 8.19) is $82\% \pm 19$ and $59\% \pm 13$ with $\xi = 0.1$ and $\xi = 0.3$ respectively. The state of the recurrent neural network is initialized without the recurrent connection during the first step of each sequence. Therefore, the recurrent weights are not used nor updated when training the semantic pathway with individual positions (sequences of length 1). To test whether the autoencoding objective function can indeed recruit recurrent weights, an additional pretraining phase is performed. A long random sequence of 2509 steps is generated and transformed into 2500 sequences of 10 steps using a sliding window mimicking theta phase precession. A horizontally flipped duplicate of these sequences is created to avoid any bias towards right or left arms. During that second pretraining phase, the recurrent weights of CA3 are trained to autoencode spatial positions of the 5000 sequences before the model encodes the 100 test sequences. 50 runs are performed and compared to 50 control runs lacking the second pretraining phase. With $\xi = 0.1$, the outcomes from the second pretraining test (M = 0.81, SD = 0.03) and the control test (M = 0.85, SD = 0.14) indicate that pretraining CA3 with sequences does not significantly change the average performance ($t(49) = 1.93$, $p = .06$). The Brown-Forsythe test indicates that second pretraining induces changes in variance (statistic = 15.55; $p < .001$).

Similarly, using $\xi = 0.3$, the results of the second pretraining test (M = 0.59, SD = 0.05) and the control test (M = 0.58, SD = 0.11) suggest that pretraining CA3 does not significantly alter the mean performance ($t(49) = -0.64$, $p = .53$), while changing variance (statistic = 19.78; $p < .001$). An additional experiment is carried out to ensure that the recurrent weights can be learned and used to perform the task rather than just decreasing the variance in performance. The initial recurrent weights are divided by 50 to penalize the model with weights left random. In this case, with $\xi = 0.1$, the results obtained with random (M = 0.52, SD = 0.06) and trained weights (M = 0.6, SD = 0.04) indicate that training recurrent weights is indeed beneficial to performance ($t(49) = -7.57$, $p < 0.001$). With $\xi = 0.3$, the results of random (M = 0.49, SD = 0.06) and trained weights (M = 0.52, SD = 0.05) lead to the same conclusion ($t(49) = -2.75$, $p < 0.01$).

So far, eight-maze results have proven that the model can store and recall many experienced overlapping trajectories. In order to show that the model can also perform the spatial alternation task in one-shot, a new experiment is conducted in which a single sequence for each starting arm is encoded. Consequence is used with recurrent weights (initial weights left unchanged) and the second pretraining phase introduced in the previous paragraph. The 98 remaining sequences cue the network during the test phase. 69% of these sequences end in the correct arm despite not being encoded. With leaky CA3 ($\alpha = 0.5$), 92% of the unexperienced sequences end correctly.

8.6 Discussion

In this chapter, the possibility that the hippocampus implements a two-layer associative memory was explored. While it has been proposed several times that the hippocampus could implement one (Krotov and J. Hopfield 2020; Chandra et al. 2023; Spens and Burgess 2023), the precise anatomical mapping had not been discovered. After reviewing all possible circuits, a new model was proposed in which the dentate gyrus plays a central role.

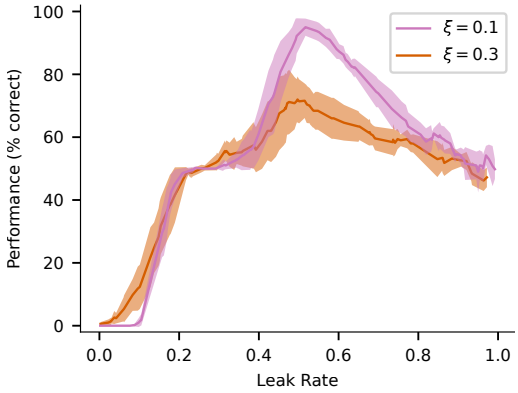
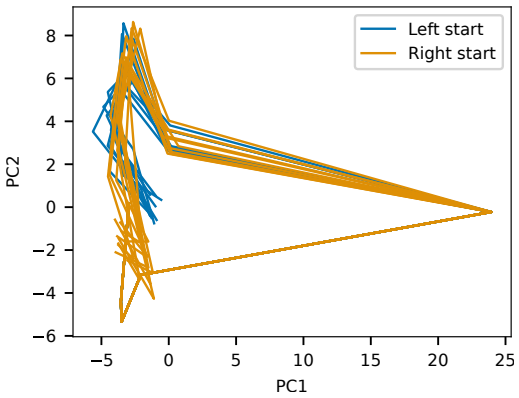


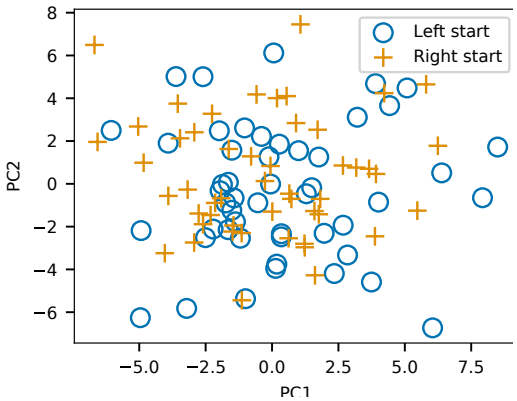
Figure 8.17: Performance of Consequence with leaky CA3 in the eight maze task. Gaussian noise of standard deviation ξ is added to the input during recall. A trial is correct if the final decoded position is in the upper right arm for trajectories starting on the left and in the upper left arm for trajectories starting on the right. The mean and standard deviation of 5 runs are shown.

With its large number of neurons and unique properties like sparse activity and continuous neurogenesis, this region is well suited to store properly separated memories. Furthermore, it is ideally situated with projections to and from CA3, and direct afferent connections from the entorhinal cortex. In this new model, CA3 also plays a central role, as it teaches the dentate gyrus to reconstruct memories with a distinct projections from entorhinal neurons. It also sends information back to the dentate gyrus, closing an ideal loop for sequence learning. Finally, recurrent collaterals can carry information over time to disambiguate overlapping sequences.

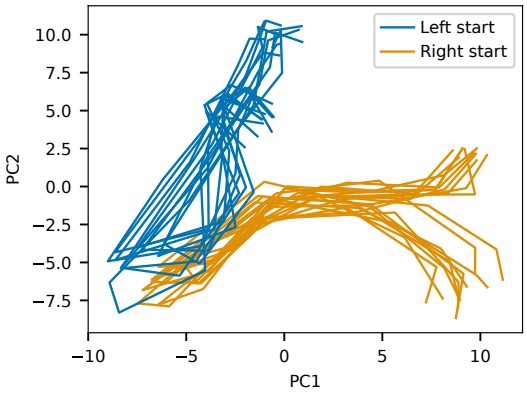
The proposed model learns sequences perfectly. Implementations of the models of Levy (1996) and Lisman, Talamini, et al. (2005) were proposed that matched the implementation of Consequence as closely as possible, while respecting the details of the original publications. The new model outperformed its predecessors in one-shot learning of spatial trajectories and random sequences. The Socratic model has inspired the development of Consequence in its use of the CA3-to-DG backprojection. However, no implementation of Socratic has previously been published and it failed to learn the tasks in the present work. This is likely due to a tradeoff between a high beta allowing the separation of memories and a lower beta allowing



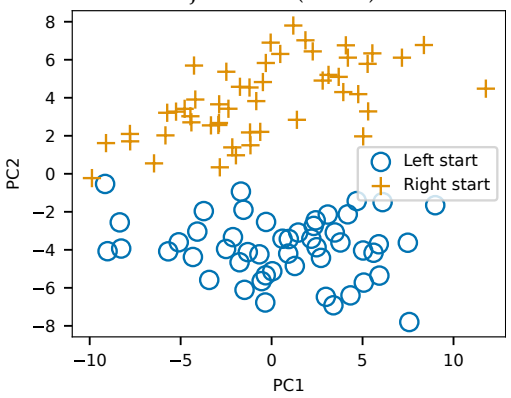
(a) Principal component analysis of the full trajectories ($\alpha = 1$).



(b) Principal component analysis restricted to the center position which is shared across all trajectories ($\alpha = 1$).



(c) Principal component analysis of the full trajectories ($\alpha = 0.5$).



(d) Principal component analysis restricted to the center position which is shared across all trajectories ($\alpha = 0.5$).

Figure 8.18: Principal component analyses of CA3 states in the eight maze task ($\xi = 0.1$). The analyses are made on 100 trajectories (50 for each starting arm) and 20 trajectories are shown.

the error gradients of the semantic pathway training to go up the afferent connections to the dentate gyrus. This limitation is overcome by the connection of the entorhinal cortex to CA3 in the Consequence model. The Collateral model of Levy (1996) did not perform well either. While it could learn simple sequences in original publications, the general idea of storing sequences has been shown to be suboptimal (Melchior et al. 2019). As they cannot implement a separation function (Millidge et al. 2022), recurrent connections are not suffient to learn information in one-shot. The more recent CRISP model of sequence learning in the hippocampus (Cheng 2013) could also be compared to the proposed model, as it has been shown to outperform the Collateral model (Melchior et al. 2019). However, its implementation would require training CA3 separately and would not fit straightforwardly into the semantic pathway.

Consequence has also been shown to be able to recall sequential memories from noisy cues. Furthermore, biological properties emerged from the storage of spatial trajectories. Activation of granular cells in which memories have been stored resembles the receptive fields of hippocampal place cells, analogous to the model of Benna and Fusi (2021). More strikingly, the cells remap in new unexperienced environments. This remapping has been shown to follow similar attractor dynamics to hippocampal place cells (Wills et al. 2005). Place cells have not yet been found in the semantic pathway, but since it is very similar to the autoencoder model of Santos-Pata et al. (2021), it is expected that they emerge from hyperparameter tuning or small adjustments. While the semantic pathway is not the main focus of the present work, this topic is under active development. Finally, the model successfully learned the eight maze task in which similar sequence memories must be disambiguated according to the recent past. Retrospective splitting representations (Duvelle et al. 2023) emerged, as suggested by the differential encoding of merging sequences. The model performs well with random recurrent CA3 connections too, which is consistent with the success of reservoir computing (Schrauwen et al. 2007), but the performance is more variable over runs than with leaky CA3. Pretraining recurrent weights with phase precessing sequences reduces variance without

altering average performance when initial weights are left unchanged, and improves performance when the magnitude of initial weights is reduced. With contextual modulation similar to that investigated in the previous chapter, prospective splitter cells could probably be observed too. The prefrontal cortex could indeed contextually bias the hippocampus to recall trajectories leading to a certain state, which should be reflected in the representation of space. The Consequence model has also been shown to be capable of efficiently using its experience. The model can indeed perform the spatial alternation task of the eight-maze in one-shot, after experiencing one sequence starting in each lower arm only.

The operations of the Consequence model are biologically plausible. The semantic pathway is trained with backpropagation, and it has been argued that the hippocampus is capable of implementing such error-driven learning (Santos-Pata et al. 2021; Y. Zheng et al. 2022; Ketz et al. 2013). The similarity function is the dot product which is simply implemented by synaptic transmission in artificial neural networks. The softmax function is very efficient to separate memories, and it has recently been shown that it can be implemented in biological neural networks too (Snow and Orchard 2022). This implementation involves interneurons, of which the biological dentate gyrus is populated with. It remains to be determined whether interneuron computation in the dentate gyrus is indeed linked to softmax calculation, as suggested by the steep sigmoidal response of dentate gyrus neurons to entorhinal inputs (Renata Bartesaghi et al. 1995), or whether the dentate gyrus implements a slightly different separation function through its sparsity. The episodic pathway of Consequence is trained using a learning rule that was recently proposed as a biologically plausible way of encoding memories in modern associative memory networks (Tyulmankov et al. 2021). The authors showed that storing keys with random local third factor is more plausible but less efficient than sequentially activating the factor of each neuron as done here (Equation 8.1). With random selection of neurons for memory encoding, stored information can be arbitrarily overwritten. The continuous birth of highly plastic granule cells in the dentate gyrus seems to reflect a biological implementation of the sequential factor,

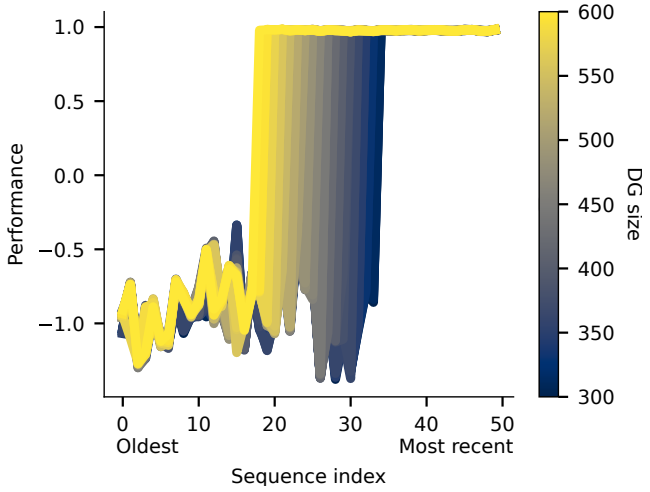


Figure 8.19: Performance of Consequence as a function of the number of neurons in DG in the spatial dataset. The number of successfully stored sequences scales linearly with the size of DG.

in which newborn neurons are recruited to store new information and then become less plastic to preserve the memory. The current capacity of Consequence is linear in the number of granule cells, as each neuron stores one autoassociation and one heteroassociation (Figure 8.19). In the biological hippocampus, dentate gyrus activity is sparse, but involves more than a single neuron. The dentate gyrus neurons in the model therefore represent populations of neurons, but these populations must be independent to avoid confusing memories.

It has been suggested that the dentate gyrus is necessary to recall recent memories before information gets consolidated in CA3 collaterals (Hainmueller and Bartos 2020). It is also a prediction of the Consequence model that the (initial) storage of one-shot experience is dependent on the integrity of granule cells. Once the unique experience has been captured by the dentate gyrus, replays could be generated to encode the information in CA3. CA3 recurrences could thus be trained in many-shot fashion, with a learning rule that also enables CA3 to keep track of the temporal context. Hence, the Consequence model is compatible with the view that there are complementary learning systems within the hippocampus but, unlike the model of Schapiro, Turk-Browne, M. M. Botvinick, et al. (2017), CA3 is a sta-

tistical learner in the new model. There is mixed evidence regarding the role of neurogenesis in learning, due to the diversity of tasks and methods. Another prediction of Consequence is that the birth of new granule cells is necessary for real one-shot learning in the hippocampus, even though fast learning can potentially occur in CA3. This is compatible with reports that neurogenesis in the dentate gyrus is increased by novelty (J. Brown et al. 2003) and necessary for learning temporal associations (Shors et al. 2001). As the dentate gyrus

Overall, the Consequence model constitutes a powerful alternative to the standard framework postulating that episodic learning mainly involves CA3 recurrent collaterals, a proposal that lacks empirical evidence (Cheng 2013) and is computationnally suboptimal (Millidge et al. 2022; Melchior et al. 2019).

9 General Discussion

This thesis began with three seemingly simple questions revolving around Marcel Proust's famous novel excerpt:

- *How can the simple flavor of a madeleine transport us through time and space, allowing us to vividly revisit a specific moment from our past?*
- *What makes this particular event stand out amidst the myriad of others?*
- *In what way is this memory intended to guide behavior?*

The review of the state of the art and the computational work carried out have been structured around these questions. First, a cognitive science approach was taken to describe episodic memory in contrast to other cognitive functions (Chapter 3). Episodic memory is associative, consisting of associations between elements. Crucially, these associations can link elements separated in time while preserving temporal order. Episodic memory is known as explicit memory because its content has privileged access to consciousness, even though it may operate unconsciously. It differs from semantic memory in the specificity of its memories. Since episodic memory often learns in one-shot, it encodes the characteristics of each situation, rather than extracting abstract regularities from multiple experiences, as semantic memory does. These two memories are complementary. The first allows us to integrate new information very quickly, while the second allows us to generalize our knowledge to new situations. Initial episode encoding requires the hippocampus. The involvement of this structure in episodic recall varies from theory to theory, depending in particular on how old the memory is.

Chapter 4 provided a more in-depth overview of the biological implementation of episodic memory, focusing on the hippocampus. It has been shown that the hippocampus is made up of several subfields. One that stands out is the dentate gyrus with characteristic properties like adult neurogenesis, sparsity, and steep sigmoidal response. It is also the most populated region of the hippocampus. The medial temporal lobe employs specific representations, including spatial ones with entorhinal grid cells and hippocampal place cells. They serve as structural scaffolding for the encoding of episodes. These cells also reflect memory operations during different oscillation regimes. In the theta-gamma regime, the firing of grid and place cells is organized in phase precessing sequences. During sharp-wave ripples, experienced spatial trajectories are replayed at a faster pace. This process underlies the transfer of information from episodic memory to semantic memory. It has also been shown that the standard framework of hippocampal computation posits that memories are stored in recurrent collaterals of area CA3. Several criticisms of this model have been outlined, one of which is the lack of emphasis on sequence learning. Furthermore, the Hopfield network on which the framework is based on, is attracting renewed interest. Modern, high-performance versions that integrate into deep learning have been proposed, but they do not involve recurrent connections. Overall, this calls for a new model of the hippocampus that is able to learn sequences in one-shot, and modern Hopfield networks can provide computational tools for its development.

In the first review chapter (Chapter 3), cognitive control has also been introduced as the ability to learn explicit rules and modulate inappropriate automatic responses. The interactions with episodic memory are threefold. First, episodic memory can replay experiences to train cognitive control offline. Secondly, episodic memory can supply information to the cognitive controller online. For example, it can inform the controller about the appropriateness of actions by extracting information from situations experienced once or only a few times. Finally, cognitive control can in turn modulate episodic memory retrieval to guide memory search and obtain contextually relevant information. These interactions have been described

in Chapter 5 as resulting from interactions between the prefrontal cortex and the medial temporal lobe. As with other cortical areas, entorhinal neurons exchange information with the prefrontal cortex. When memories are replayed in the hippocampus, information can train the prefrontal cortex offline. There is also a direct connection from CA1 to the medial prefrontal cortex, which can provide episodic memory context for online decision-making. Conversely, hippocampal neurons encode prefrontal information sent by the entorhinal cortex. As a result, hippocampal representations reflect the current context. Prospective splitter cells disambiguate spatial positions according to future trajectory. It is likely that they emerge because of the prefrontal cortex sending planning information. Similarly, retrospective splitter cells disambiguate spatial positions according to past trajectory. These emerge internally, as the hippocampus keeps track of the temporal context. This way, recall is biased towards memories that are relevant with respect to the current temporal context and ongoing goals. Prefrontal-hippocampal interactions are scheduled by subcortical structures like the nucleus reuniens of the thalamus, such that the prefrontal cortex controls the hippocampus to bias memory retrieval, and the hippocampus influences the prefrontal cortex when episodic context is provided.

Some machine learning algorithms are inspired by these interactions, mostly by the hippocampus-to-prefrontal direction. One of them is neural episodic control (Pritzel et al. 2017). It comprises a differentiable neural dictionary supposedly playing the role of the hippocampus to guide actions, but this parallel has so far been made on a purely functional basis. It is unclear whether the hippocampus stores experiences in the same way as the dictionary. Chapter 6 presents a formal proof that the differentiable neural dictionary belongs to the same framework as Hopfield networks: the recently proposed universal Hopfield framework (Millidge et al. 2022). Just as the dictionary, models of this framework recall memories by performing similarity, separation and projection operations. The implications are twofold. It has been shown that the dictionary can recall arbitrary information like images. The second implication is that neural episodic control could potentially be improved by using other instances

of the framework. This novel connection between reinforcement learning and associative memory paves the way for further research. It also adds a whole new dimension to a hippocampal model based on Hopfield networks, like the one presented in Chapter 8, as it can easily be integrated into a reinforcement learning paradigm.

Most machine learning algorithms of memory-augmented control ignore the contextual modulation of memory operations by the prefrontal cortex. In Chapter 7, this topic was studied. It was demonstrated that modern Hopfield networks can learn contextual memories in one-shot. For the first time, a model of the hippocampus was modulated with context information passing through multiple layers, instead of targeting memory directly. The ability to separate semantic representations and to recall the correct memory contextually was assessed as a function of the position of the Hopfield network and the prefrontal target. The prefrontal cortex can influence memory recall if memory is not too far downstream the targeted layer. Explicitly teaching the network to reconstruct context increases the distance from which the context can influence memory. This study provides design principles for the development of reinforcement learning models that can use their knowledge more efficiently. In addition, the results further call into question the standard framework, as targeting the entorhinal cortex is inefficient when memories are stored in CA3. The model performs much better when memories are stored in the dentate gyrus or CA1.

The Consequence model, standing for contextual sequence, was presented in Chapter 8. All operations of the model are biologically plausible. Since it is based on a modern Hopfield network located in the dentate gyrus, it can be integrated into reinforcement learning and modulated through prefrontal connections to the entorhinal cortex. Of all the possible implementations of a modern Hopfield network in the hippocampus, Consequence meets the most criteria (table 8.1). It largely outperforms alternative models for one-shot sequence learning. Furthermore, it can account for biological findings, such as the formation of place cells whose fields remap in novel and morphing environments. Finally, it is able to learn overlapping sequences in one-shot too, by keeping track of the tem-

poral context in the same way as hippocampal retrospective splitter cells. This allows the model to perform the spatial alternation task after a single exposure. All in all, these results argue in favor of bringing together the study of the hippocampus and the development of modern Hopfield networks. Y. Zheng et al. (2022) called for *correcting the hebbian mistake* by using error-driven algorithms, similarly, the success of feedforward associative memory models sheds new light on the functioning of the hippocampus and questions the original proposal of J. J. Hopfield (1982).

Beyond scientific contributions, a lot of code was produced to perform the experiments. The code of Millidge et al. (2022) was extended to test the capacity of the differentiable neural dictionary in associative memory tasks[1]. The code of Sarrico et al. (2019) was augmented with new similarity and separation functions[2]. A testbed was created to investigate the contextual modulation of modern Hopfield networks in the hippocampus[3]. Finally, another testbed was developed to benchmark and optimize models on a series of tasks, including the new Consequence model[4].

Perspectives

One of the aims of this thesis was to improve machine learning algorithms with functions such as episodic memory and cognitive control. We urgently need to develop models capable of hindsight and aligned with our values. Just as *deep learning needs a prefrontal cortex* (Russin et al. 2020), it probably needs a hippocampus that bidirectionally interacts with its prefrontal cortex too (Chateau-Laurent and Alexandre 2021a).

The prospects opened up by the Consequence model are manifold. The first step will be to tune the model to bring out place cells elsewhere than in the dentate gyrus, which is currently being done. Another direction is to study the model's generative capacities. The softmax function used

[1] https://github.com/HugoChateauLaurent/DND_AssociativeMemory

[2] https://github.com/HugoChateauLaurent/NEC_UHN

[3] https://github.com/HugoChateauLaurent/ContextMHN

[4] https://github.com/HugoChateauLaurent/OneShotSequenceLearning

to separate memories outputs a probability distribution which could be sampled instead of being used as is. Memory recall would then be stochastic, and the generative capacities of the hippocampus could be explored. Hippocampal contribution to imagination indeed relies on the ability to generate unexperienced yet plausible sequences of events. Consequence could model this by sampling from the softmax, and the β parameter would control the extent to which the model generates sequences that match its experience or, on the contrary, generates novel transitions. This is in line with the proposal that the study of the hippocampus can help in the development of artificial creativity (Chateau-Laurent and Alexandre 2021b). Another important step for the longevity of the model would be to propose a more precise description of the ordering of processing modes and how they fit into hippocampal oscillation patterns. For the sake of simplicity, the semantic pathway was trained prior to the encoding of memories in the episodic pathway, but how the model as a whole develops with experience will need to be clarified in the future.

Ideally, the elements studied separately here should be integrated in a common architecture (Figure 9.1). Consequence could store and retrieve experiences resembling current observations to feed a PFC-like LSTM, like in Fortunato et al. (2019). The hippocampus would receive projections from the LSTM to modulate encoding and retrieval in order for the LSTM to receive helpful information. More biologically plausible models of the prefrontal cortex could also be used (e.g. Dagar 2023; W. H. Alexander and J. W. Brown 2015; O'Reilly and M. J. Frank 2006). With this kind of architecture, it would be possible to test whether hippocampal neurons projecting to the prefrontal cortex have broad spatial representations (ventral) and neurons targeted by the prefrontal cortex have small spatial representations (dorsal), as predicted by Howard Eichenbaum (2017b). Interestingly, Consequence would add a temporal dimension to episodic control, as it stores sequences. Just as Dreamer (Hafner et al. 2019) generates model-based imaginative sequences, Consequence could generate prospective episodic sequences in collaboration with the prefrontal cortex. This would bring machine learning closer to the construction system of the brain (Hassabis and Maguire

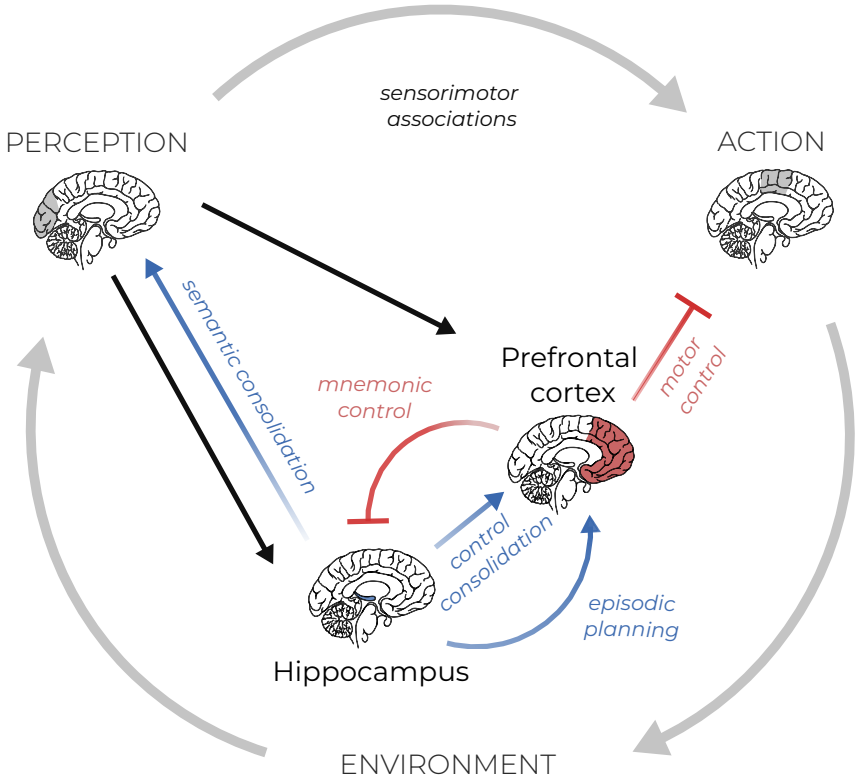


Figure 9.1: Proposed architecture for generative episodic control. Action is selected through sensorimotor associations controlled by the prefrontal cortex. The prefrontal controller interacts with the hippocampus to generate prospective episodic sequences. Offline, the hippocampus is used to consolidate new information in semantic memory and train the prefrontal cortex with simulated situations.

2009). To encode all the specificity of memories and generalize efficiently at the same time, the hippocampus should not be a bottleneck like in neural episodic control, but rather be complemented by a semantic memory processing sensory inputs and selecting actions in a model-free way. Replays generated in the hippocampus would be used to train the controller and consolidate semantic memory offline.

It now appears more feasible to provide answers to the three initial questions posed in this thesis. The reason the narrator of *Swann's Way* is transported to the distant past is that the madeleine's flavor acts as a key connected to specific neurons. These neurons, once selectively chosen among irrelevant ones, reactivate the value of the association and trigger a cascade of information that replaces cortical representations with those dedicated to a singular experience: Sunday mornings at his aunt Léonie's house in Combray. How is this recollection meant to influence behavior? It does not, it is just a madeleine.

Bibliography

Agster, Kara L and Rebecca D Burwell (2009). "Cortical efferents of the perirhinal, postrhinal, and entorhinal cortices of the rat". *Hippocampus* 19:12, pp. 1159–1186.

Agster, Kara L, Norbert J Fortin, and Howard Eichenbaum (2002). "The hippocampus and disambiguation of overlapping sequences". *Journal of Neuroscience* 22:13, pp. 5760–5768.

Aguilera, Matthieu, Vincent Douchamps, Demian Battaglia, and Romain Goutagny (2022). "How Many Gammas? Redefining Hippocampal Theta-Gamma Dynamic During Spatial Learning". *Frontiers in behavioral neuroscience* 16, p. 811278.

Alexander, William H and Joshua W Brown (2015). "Hierarchical error representation: a computational model of anterior cingulate and dorsolateral prefrontal cortex". *Neural Computation* 27:11, pp. 2354–2410.

Allen, Gary V and David A Hopkins (1989). "Mamillary body in the rat: topography and synaptology of projections from the subicular complex, prefrontal cortex, and midbrain tegmentum". *Journal of Comparative Neurology* 286:3, pp. 311–336.

Alvernhe, Alice, Etienne Save, and Bruno Poucet (2011). "Local remapping of place cell firing in the Tolman detour task". *European Journal of Neuroscience* 33:9, pp. 1696–1705.

Ambrose, R Ellen, Brad E Pfeiffer, and David J Foster (2016). "Reverse replay of hippocampal place cells is uniquely modulated by changing reward". *Neuron* 91:5, pp. 1124–1136.

Andersen, Per (1975). "Organization of hippocampal neurons and their interconnections". In: *The Hippocampus: Volume 1: Structure and Development*. Springer, pp. 155–175.

Anderson, Michael C, Jamie G Bunce, and Helen Barbas (2016). "Prefrontal–hippocampal pathways underlying inhibitory control over memory". *Neurobiology of learning and memory* 134, pp. 145–161.

Andrianova, Lilya, Steliana Yanakieva, Gabriella Margetts-Smith, Shivali Kohli, Erica S Brady, John P Aggleton, and Michael T Craig (2023). "No evidence from complementary data sources of a direct glutamatergic projection from the mouse anterior cingulate area to the hippocampal formation". *Elife* 12, e77364.

Apergis-Schoute, John, Aline Pinto, and Denis Paré (2006). "Ultrastructural organization of medial prefrontal inputs to the rhinal cortices". *European Journal of Neuroscience* 24:1, pp. 135–144.

Aronov, Dmitriy, Rhino Nevers, and David W Tank (2017). "Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit". *Nature* 543:7647, pp. 719–722.

Atkinson, Richard C and Richard M Shiffrin (1968). "Human memory: A proposed system and its control processes". In: *Psychology of learning and motivation*. Vol. 2. Elsevier, pp. 89–195.

Augusto-Oliveira, Marcus, Gabriela PF Arrifano, João O Malva, and Maria Elena Crespo-Lopez (2019). "Adult hippocampal neurogenesis in different taxonomic groups: Possible functional similarities and striking controversies". *Cells* 8:2, p. 125.

Backus, Alexander R, Jan-Mathijs Schoffelen, Szabolcs Szebényi, Simon Hanslmayr, and Christian F Doeller (2016). "Hippocampal-prefrontal theta oscillations support memory integration". *Current Biology* 26:4, pp. 450–457.

Baddeley, Alan (2000). "The episodic buffer: a new component of working memory?" *Trends in cognitive sciences* 4:11, pp. 417–423.

Baddeley, Alan D and J Graham (1974). "Working memory". *The psychology of learning and motivation* 8, pp. 47–89.

Badre, David and Mark D'Esposito (2007). "Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex". *Journal of cognitive neuroscience* 19:12, pp. 2082–2099.

Baeg, EH, YB Kim, K Huh, I Mook-Jung, HT Kim, and MW Jung (2003). "Dynamics of population code for working memory in the prefrontal cortex". *Neuron* 40:1, pp. 177–188.

Bakker, Hanna den, Marie Van Dijck, Jyh-Jang Sun, and Fabian Kloosterman (2023). "Sharp-wave-ripple-associated activity in the medial prefrontal cortex supports spatial rule switching". *Cell Reports* 42:8.

Banino, Andrea, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. (2018). "Vector-based navigation using grid-like representations in artificial agents". *Nature* 557:7705, pp. 429–433.

Barnes, Carol A, Bruce L McNaughton, Shed JY Mizumori, Brian W Leonard, and Lei-Huey Lin (1990). "Chapter Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing". *Progress in brain research* 83, pp. 287–300.

Barnes, Jean M and Benton J Underwood (1959). "" Fate" of first-list associations in transfer theory." *Journal of experimental psychology* 58:2, p. 97.

Bartesaghi, R, M Migliore, and T Gessi (2006). "Input–output relations in the entorhinal cortex–dentate–hippocampal system: Evidence for a nonlinear transfer of signals". *Neuroscience* 142:1, pp. 247–265.

Bartesaghi, Renata, Tiziana Gessi, and Michele Migliore (1995). "Input-output relations in the entorhinal-hippocampal-entorhinal loop: Entorhinal cortex and dentate gyrus". *Hippocampus* 5:5, pp. 440–451.

Battaglia, Francesco P, Gary R Sutherland, and Bruce L McNaughton (2004). "Local sensory cues and place cell directionality: additional evidence of prospective coding in the hippocampus". *Journal of Neuroscience* 24:19, pp. 4541–4550.

Bauer, Richard H and Joaquin M Fuster (1976). "Delayed-matching and delayed-response deficit from cooling dorsolateral prefrontal cortex in monkeys." *Journal of comparative and physiological psychology* 90:3, p. 293.

Becker, Suzanna (2005). "A computational principle for hippocampal learning and neurogenesis". *Hippocampus* 15:6, pp. 722–738.

Benchenane, Karim, Adrien Peyrache, Mehdi Khamassi, Patrick L Tierney, Yves Gioanni, Francesco P Battaglia, and Sidney I Wiener (2010). "Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning". *Neuron* 66:6, pp. 921–936.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". *IEEE transactions on neural networks* 5:2, pp. 157–166.

Benna, Marcus K and Stefano Fusi (2021). "Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence". *Proceedings of the National Academy of Sciences* 118:51, e2018422118.

Bentley, Jon Louis (1975). "Multidimensional binary search trees used for associative searching". *Communications of the ACM* 18:9, pp. 509–517.

Berners-Lee, Alice, Xiaojing Wu, and David J Foster (2021). "Prefrontal cortical neurons are selective for non-local hippocampal representations during replay and behavior". *Journal of Neuroscience* 41:27, pp. 5894–5908.

Bernier, Brian E, Anthony F Lacagnina, Adam Ayoub, Francis Shue, Boris V Zemelman, Franklin B Krasne, and Michael R Drew (2017). "Dentate gyrus contributes to retrieval as well as encoding: evidence from context fear conditioning, recall, and extinction". *Journal of Neuroscience* 37:26, pp. 6359–6371.

Beukers, Andre O, Timothy J Buschman, Jonathan D Cohen, and Kenneth A Norman (2021). "Is activity silent working memory simply episodic memory?" *Trends in cognitive sciences* 25:4, pp. 284–293.

Bian, Xin-Lan, Cheng Qin, Cheng-Yun Cai, Ying Zhou, Yan Tao, Yu-Hui Lin, Hai-Yin Wu, Lei Chang, Chun-Xia Luo, and Dong-Ya Zhu (2019). "Anterior cingulate cortex to ventral hippocampus circuit mediates contextual fear generalization". *Journal of Neuroscience* 39:29, pp. 5728–5739.

Bieri, Kevin Wood, Katelyn N Bobbitt, and Laura Lee Colgin (2014). "Slow and fast gamma rhythms coordinate different spatial coding modes in hippocampal place cells". *Neuron* 82:3, pp. 670–681.

Blundell, Charles, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis (2016). "Model-free episodic control". *arXiv preprint arXiv:1606.04460*.

Botvinick, Matthew, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis (2019). "Reinforcement learning, fast and slow". *Trends in cognitive sciences* 23:5, pp. 408–422.

Brankačk, Jurij, Mark Stewart, and Steven E Fox (1993). "Current source density analysis of the hippocampal theta rhythm: associated sustained potentials and candidate synaptic generators". *Brain research* 615:2, pp. 310–327.

Brasted, Peter J and Steven P Wise (2004). "Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum". *European Journal of Neuroscience* 19:3, pp. 721–740.

Brockman, Greg, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba (2016). "Openai gym". *arXiv preprint arXiv:1606.01540*.

Brown, Jason, Christiana M Cooper-Kuhn, Gerd Kempermann, Henriette Van Praag, Jürgen Winkler, Fred H Gage, and H Georg Kuhn (2003). "Enriched environment and physical activity stimulate hippocampal but not olfactory bulb neurogenesis". *European journal of Neuroscience* 17:10, pp. 2042–2046.

Bruck, Jehoshua (1990). "On the convergence properties of the Hopfield model". *Proceedings of the IEEE* 78:10, pp. 1579–1585.

Burak, Yoram and Ila R Fiete (2009). "Accurate path integration in continuous attractor network models of grid cells". *PLoS computational biology* 5:2, e1000291.

Burwell, Rebecca D and David G Amaral (1998). "Cortical afferents of the perirhinal, postrhinal, and entorhinal cortices of the rat". *Journal of comparative neurology* 398:2, pp. 179–205.

Butler, James L and Ole Paulsen (2014). "The hippocampal cacophony: multiple layers of communication". *Neuron* 84:2, pp. 251–253.

Buzsáki, György (1989). "Two-stage model of memory trace formation: a role for "noisy" brain states". *Neuroscience* 31:3, pp. 551–570.

Buzsáki, György and David Tingley (2018). "Space and time: the hippocampus as a sequence generator". *Trends in cognitive sciences* 22:10, pp. 853–869.

Buzsáki, György and Brendon O Watson (2012). "Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease." *Dialogues in clinical neuroscience*.

Cassel, Jean-Christophe, Anne Pereira De Vasconcelos, Michaël Loureiro, Thibault Cholvin, John C Dalrymple-Alford, and Robert P Vertes (2013). "The reuniens and rhomboid nuclei: neuroanatomy, electrophysiological characteristics and behavioral implications". *Progress in neurobiology* 111, pp. 34–52.

Caucheteux, Charlotte and Jean-Rémi King (2022). "Brains and algorithms partially converge in natural language processing". *Communications biology* 5:1, p. 134.

Cazé, Romain, Mehdi Khamassi, Lise Aubin, and Benoît Girard (2018). "Hippocampal replays under the scrutiny of reinforcement learning models". *Journal of neurophysiology* 120:6, pp. 2877–2896.

Cazin, Nicolas (2018). "A replay driven model of spatial sequence learning in the hippocampus-prefrontal cortex network using reservoir computing". PhD thesis. Université de Lyon.

Chandra, Sarthak, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete (2023). "High-capacity flexible hippocampal associative and episodic memory enabled by prestructured"spatial"representations". *bioRxiv*, pp. 2023–11.

Chateau-Laurent, Hugo and Frédéric Alexandre (2021a). "Augmenting machine learning with flexible episodic memory". In: *13th International Joint Conference on Computational Intelligence*.

– (2021b). "Towards a Computational Cognitive Neuroscience Model of Creativity". In: *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, pp. 67–71.

Chaudhry, Hamza Tahir, Jacob A Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan (2023). "Long Sequence Hopfield Memory". *arXiv preprint arXiv:2306.04532*.

Cheng, Sen (2013). "The CRISP theory of hippocampal function in episodic memory". *Frontiers in neural circuits* 7, p. 88.

Chudasama, Yogita, Victoria M Doobay, and Yuchen Liu (2012). "Hippocampal-prefrontal cortical circuit mediates inhibitory response control in the rat". *Journal of Neuroscience* 32:32, pp. 10915–10924.

Cipolotti, Lisa, Tim Shallice, Dennis Chan, Nick Fox, Rachel Scahill, Gail Harrison, John Stevens, and Peter Rudge (2001). "Long-term retrograde amnesia... the crucial role of the hippocampus". *Neuropsychologia* 39:2, pp. 151–172.

Clark, Andy (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA.

Clark, Andy and David Chalmers (1998). "The extended mind". *analysis* 58:1, pp. 7–19.

Cohen, Jonathan D and Randall C O'Reilly (1996). "A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory". *Prospective memory: Theory and applications*, pp. 267–295.

Cohen, Neal J, Jennifer Ryan, Caroline Hunt, Lorene Romine, Tracey Wszalek, and Courtney Nash (1999). "Hippocampal system and declarative (relational) memory: summarizing the data from functional neuroimaging studies". *Hippocampus* 9:1, pp. 83–98.

Colgin, Laura Lee (2015). "Do slow and fast gamma rhythms correspond to distinct functional states in the hippocampal network?" *Brain research* 1621, pp. 309–315.

Colgin, Laura Lee, Tobias Denninger, Marianne Fyhn, Torkel Hafting, Tora Bonnevie, Ole Jensen, May-Britt Moser, and Edvard I Moser (2009). "Frequency of gamma oscillations routes flow of information in the hippocampus". *Nature* 462:7271, pp. 353–357.

Cowen, Stephen L, Glen A Davis, and Douglas A Nitz (2012). "Anterior cingulate neurons in the rat map anticipated effort and reward to their associated action sequences". *Journal of neurophysiology* 107:9, pp. 2393–2407.

Cravens, Catherine J, Noelia Vargas-Pinto, Kimberly M Christian, and Kazu Nakazawa (2006). "CA3 NMDA receptors are crucial for rapid and automatic representation of context memory". *European Journal of Neuroscience* 24:6, pp. 1771–1780.

Cueva, Christopher J and Xue-Xin Wei (2018). "Emergence of grid-like representations by training recurrent neural networks to perform spatial localization". *arXiv preprint arXiv:1803.07770*.

Cutsuridis, Vassilis, Bruce P Graham, Stuart Cobb, and Imre Vida (2019). *Hippocampal microcircuits: a computational modeler's resource book*. Springer.

Dagar, Snigdha (2023). "Computational modeling of Cognitive Control for rule-guided behavior". PhD thesis. Université de Bordeaux.

Dayan, Peter (1993). "Improving generalization for temporal difference learning: The successor representation". *Neural computation* 5:4, pp. 613–624.

De Lavilléon, Gaetan, Marie Masako Lacroix, Laure Rondi-Reig, and Karim Benchenane (2015). "Explicit memory creation during sleep demonstrates a causal role of place cells in navigation". *Nature neuroscience* 18:4, pp. 493–495.

Degonda, Nadia, Christian RA Mondadori, Simone Bosshardt, Conny F Schmidt, Peter Boesiger, Roger M Nitsch, Christoph Hock, and Katharina Henke (2005). "Implicit associative learning engages the hippocampus and interacts with explicit associative learning". *Neuron* 46:3, pp. 505–520.

Dehaene, Stanislas, Jean-Pierre Changeux, Lionel Naccache, Jérôme Sackur, and Claire Sergent (2006). "Conscious, preconscious, and subliminal processing: a testable taxonomy". *Trends in cognitive sciences* 10:5, pp. 204–211.

Demircigil, Mete, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet (2017). "On a model of associative memory with huge storage capacity". *Journal of Statistical Physics* 168, pp. 288–299.

Denny, Christine A, Mazen A Kheirbek, Eva L Alba, Kenji F Tanaka, Rebecca A Brachman, Kimberly B Laughman, Nicole K Tomm, Gergely F Turi, Attila Losonczy, and René Hen (2014). "Hippocampal memory traces are differentially modulated by experience, time, and adult neurogenesis". *Neuron* 83:1, pp. 189–201.

Deshmukh, Sachin S and James J Knierim (2013). "Influence of local objects on hippocampal representations: Landmark vectors and memory". *Hippocampus* 23:4, pp. 253–267.

Diamantaki, Maria, Markus Frey, Philipp Berens, Patricia Preston-Ferrer, and Andrea Burgalossi (2016). "Sparse activity of identified dentate granule cells during spatial exploration". *Elife* 5, e20252.

Diba, Kamran and György Buzsáki (2007). "Forward and reverse hippocampal place-cell sequences during ripples". *Nature neuroscience* 10:10, pp. 1241–1242.

Dickinson, Anthony and Bernard Balleine (2002). "The role of learning in the operation of motivational systems". *Stevens' handbook of experimental psychology* 3, pp. 497–533.

Dobbins, Ian G, Heather Foley, Daniel L Schacter, and Anthony D Wagner (2002). "Executive control during episodic retrieval: multiple prefrontal processes subserve source memory". *Neuron* 35:5, pp. 989–996.

Dolleman-van der Weel, Margriet J, Amy L Griffin, Hiroshi T Ito, Matthew L Shapiro, Menno P Witter, Robert P Vertes, and Timothy A Allen (2019). "The nucleus reuniens of the thalamus sits at the nexus of a hippocampus and medial prefrontal cortex circuit enabling memory and behavior". *Learning & Memory* 26:7, p. 191.

Dolleman-Van Der Weel, MJ and MP Witter (1996). "Projections from the nucleus reuniens thalami to the entorhinal cortex, hippocampal field CA1, and the subiculum in the rat arise from different populations of neurons". *Journal of comparative neurology* 364:4, pp. 637–650.

Douchamps, Vincent, Ali Jeewajee, Pam Blundell, Neil Burgess, and Colin Lever (2013). "Evidence for encoding versus retrieval scheduling in the hippocampus by theta phase and acetylcholine". *Journal of Neuroscience* 33:20, pp. 8689–8704.

Drieu, Céline and Michaël Zugaro (2019). "Hippocampal sequences during exploration: mechanisms and functions". *Frontiers in cellular neuroscience* 13, p. 232.

Dupret, David, Joseph O'neill, Barty Pleydell-Bouverie, and Jozsef Csicsvari (2010). "The reorganization and reactivation of hippocampal maps predict spatial memory performance". *Nature neuroscience* 13:8, pp. 995–1002.

Dusek, Jeffery A and Howard Eichenbaum (1997). "The hippocampus and memory for orderly stimulus relations". *Proceedings of the National Academy of Sciences* 94:13, pp. 7109–7114.

Duvelle, Éléonore, Roddy M Grieves, and Matthijs AA Van der Meer (2023). "Temporal context and latent state inference in the hippocampal splitter signal". *ELife* 12, e82357.

Ego-Stengel, Valérie and Matthew A Wilson (2010). "Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat". *Hippocampus* 20:1, pp. 1–10.

Egorov, Alexei V, Bassam N Hamam, Erik Fransén, Michael E Hasselmo, and Angel A Alonso (2002). "Graded persistent activity in entorhinal cortex neurons". *Nature* 420:6912, pp. 173–178.

Eichenbaum, H, M Kuperstein, A Fagan, and J Nagode (1987). "Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task". *Journal of Neuroscience* 7:3, pp. 716–732.

Eichenbaum, Howard (2017a). "Memory: organization and control". *Annual review of psychology* 68, pp. 19–45.

– (2017b). "Prefrontal–hippocampal interactions in episodic memory". *Nature Reviews Neuroscience* 18:9, pp. 547–558.

Eichenbaum, Howard, Paul Dudchenko, Emma Wood, Matthew Shapiro, and Heikki Tanila (1999). "The hippocampus, memory, and place cells: is it spatial memory or a memory space?" *Neuron* 23:2, pp. 209–226.

Eichenbaum, Howard, Andrew P Yonelinas, and Charan Ranganath (2007). "The medial temporal lobe and recognition memory". *Annu. Rev. Neurosci.* 30, pp. 123–152.

Elman, Jeffrey L (1990). "Finding structure in time". *Cognitive science* 14:2, pp. 179–211.

Ergorul, Ceren and Howard Eichenbaum (2004). "The hippocampus and memory for "what", "where", and "when"". *Learning & Memory* 11:4, pp. 397–405.

Farovik, Anja, Laura M Dupont, Miguel Arce, and Howard Eichenbaum (2008). "Medial prefrontal cortex supports recollection, but not familiarity, in the rat". *Journal of Neuroscience* 28:50, pp. 13428–13434.

Feng, Ting, Delia Silva, and David J Foster (2015). "Dissociation between the experience-dependent development of hippocampal theta sequences and single-trial phase precession". *Journal of Neuroscience* 35:12, pp. 4890–4902.

Fernández-Ruiz, Antonio, Azahara Oliva, Eliezyer Fermino de Oliveira, Florbela Rocha-Almeida, David Tingley, and György Buzsáki (2019). "Long-duration hippocampal sharp wave ripples improve memory". *Science* 364:6445, pp. 1082–1086.

Fernández-Ruiz, Antonio, Azahara Oliva, Gergő A Nagy, Andrew P Maurer, Antal Berényi, and György Buzsáki (2017). "Entorhinal-CA3 dual-input control of spike timing in the hippocampus by theta-gamma coupling". *Neuron* 93:5, pp. 1213–1226.

Fortin, Norbert J, Kara L Agster, and Howard Eichenbaum (2002). "Critical role of the hippocampus in memory for sequences of events". *Nature neuroscience* 5:5, pp. 458–462.

Fortunato, Meire, Melissa Tan, Ryan Faulkner, Steven Hansen, Adrià Puigdomènech Badia, Gavin Buttimore, Charles Deck, Joel Z Leibo, and Charles Blundell (2019). "Generalization of reinforcement learners with working and episodic memory". *Advances in neural information processing systems* 32.

Foster, David J and Matthew A Wilson (2006). "Reverse replay of behavioural sequences in hippocampal place cells during the awake state". *Nature* 440:7084, pp. 680–683.

Francisco, Varela, Thompson Evan, and Rosch Eleanor (1991). "The embodied mind: Cognitive science and human experience". *Massachusetts Institute of Technology*.

Frank, Loren M, Emery N Brown, and Matthew Wilson (2000). "Trajectory encoding in the hippocampus and entorhinal cortex". *Neuron* 27:1, pp. 169–178.

Frankland, Paul W and Bruno Bontempi (2005). "The organization of recent and remote memories". *Nature reviews neuroscience* 6:2, pp. 119–130.

Fransén, Erik, Babak Tahvildari, Alexei V Egorov, Michael E Hasselmo, and Angel A Alonso (2006). "Mechanism of graded persistent cellular activity of entorhinal cortex layer v neurons". *Neuron* 49:5, pp. 735–746.

Fried, Itzhak, Katherine A MacDonald, and Charles L Wilson (1997). "Single neuron activity in human hippocampus and amygdala during recognition of faces and objects". *Neuron* 18:5, pp. 753–765.

Fuhs, Mark C and David S Touretzky (2006). "A spin glass model of path integration in rat medial entorhinal cortex". *Journal of Neuroscience* 26:16, pp. 4266–4276.

– (2007). "Context learning in the rodent hippocampus". *Neural computation* 19:12, pp. 3173–3215.

Funahashi, Shintaro, Charles J Bruce, and Patricia S Goldman-Rakic (1989). "Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex". *Journal of neurophysiology* 61:2, pp. 331–349.

– (1993). "Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic scotomas". *Journal of Neuroscience* 13:4, pp. 1479–1497.

Fuster, Joaquin M and Garrett E Alexander (1971). "Neuron activity related to short-term memory". *Science* 173:3997, pp. 652–654.

Gaal, Simon van, Floris P De Lange, and Michael X Cohen (2012). "The role of consciousness in cognitive control and decision making". *Frontiers in human neuroscience* 6, p. 121.

Gabrieli, John DE, Suzanne Corkin, Susan F Mickel, and John H Growdon (1993). "Intact acquisition and long-term retention of mirror-tracing skill in Alzheimer's disease and in global amnesia." *Behavioral neuroscience* 107:6, p. 899.

Gardner, Richard J, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser (2022). "Toroidal topology of population activity in grid cells". *Nature* 602:7895, pp. 123–128.

George, Tom M, William de Cothi, Claudia Clopath, Kimberly Stachenfeld, and Caswell Barry (2022). "RatInABox: A toolkit for modelling locomotion and neuronal activity in continuous environments". *bioRxiv*, pp. 2022–08.

George, Tom M, William de Cothi, Kimberly L Stachenfeld, and Caswell Barry (2023). "Rapid learning of predictive maps with STDP and theta phase precession". *Elife* 12, e80663.

Gershman, Samuel J and Nathaniel D Daw (2017). "Reinforcement learning and episodic memory in humans and animals: an integrative framework". *Annual review of psychology* 68, pp. 101–128.

Gershman, Samuel J, Christopher D Moore, Michael T Todd, Kenneth A Norman, and Per B Sederberg (2012). "The successor representation and temporal context". *Neural Computation* 24:6, pp. 1553–1568.

Gerstner, Wulfram, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea (2018). "Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules". *Frontiers in neural circuits* 12, p. 53.

Gilbert, Daniel T and Timothy D Wilson (2007). "Prospection: Experiencing the future". *Science* 317:5843, pp. 1351–1354.

Gilbert, Paul E, Raymond P Kesner, and Inah Lee (2001). "Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1". *Hippocampus* 11:6, pp. 626–636.

Girardeau, Gabrielle, Karim Benchenane, Sidney I Wiener, György Buzsáki, and Michaël B Zugaro (2009). "Selective suppression of hippocampal ripples impairs spatial memory". *Nature neuroscience* 12:10, pp. 1222–1223.

Giustino, Thomas F and Stephen Maren (2015). "The role of the medial prefrontal cortex in the conditioning and extinction of fear". *Frontiers in behavioral neuroscience* 9, p. 298.

Goldman-Rakic, Patricia S (1996). "The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351:1346, pp. 1445–1453.

Gonzalo-Ruiz, A, A Alonso, JM Sanz, and RR Llinas (1992). "Afferent projections to the mammillary complex of the rat, with special reference to those from surrounding hypothalamic regions". *Journal of Comparative Neurology* 321:2, pp. 277–299.

Goyal, Anirudh, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, et al. (2022). "Retrieval-augmented reinforcement learning". In: *International Conference on Machine Learning*. PMLR, pp. 7740–7765.

Graves, Alex, Greg Wayne, and Ivo Danihelka (2014). "Neural turing machines". *arXiv preprint arXiv:1410.5401*.

Graves, Alex, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. (2016). "Hybrid computing using a neural network with dynamic external memory". *Nature* 538:7626, pp. 471–476.

Gridchyn, Igor, Philipp Schoenenberger, Joseph O'Neill, and Jozsef Csicsvari (2020). "Assembly-specific disruption of hippocampal replay leads to selective memory deficit". *Neuron* 106:2, pp. 291–300.

Guanella, Alexis, Daniel Kiper, and Paul Verschure (2007). "A model of grid cells based on a twisted torus topology". *International journal of neural systems* 17:04, pp. 231–240.

Guise, Kevin G and Matthew L Shapiro (2017). "Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding". *Neuron* 94:1, pp. 183–192.

Gupta, Anoopum S, Matthijs AA Van Der Meer, David S Touretzky, and A David Redish (2010). "Hippocampal replay is not a simple function of experience". *Neuron* 65:5, pp. 695–705.

Hafner, Danijar, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi (2019). "Dream to control: Learning behaviors by latent imagination". *arXiv preprint arXiv:1912.01603*.

Hafting, Torkel, Marianne Fyhn, Tora Bonnevie, May-Britt Moser, and Edvard I Moser (2008). "Hippocampus-independent phase precession in entorhinal grid cells". *Nature* 453:7199, pp. 1248–1252.

Hafting, Torkel, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser (2005). "Microstructure of a spatial map in the entorhinal cortex". *Nature* 436:7052, pp. 801–806.

Hahnloser, Richard HR, Alexay A Kozhevnikov, and Michale S Fee (2003). "Erratum: An ultra-sparse code underlies the generation of neural sequences in a songbird". *Nature* 421:6920, pp. 294–294.

Hainmueller, Thomas and Marlene Bartos (2020). "Dentate gyrus circuits for encoding, retrieval and discrimination of episodic memories". *Nature Reviews Neuroscience* 21:3, pp. 153–168.

Hallock, Henry L, Arick Wang, and Amy L Griffin (2016). "Ventral midline thalamus is critical for hippocampal–prefrontal synchrony and spatial working memory". *Journal of Neuroscience* 36:32, pp. 8372–8389.

Halton, John H (1960). "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals". *Numerische Mathematik* 2, pp. 84–90.

Hannula, Deborah E and Anthony J Greene (2012). "The hippocampus reevaluated in unconscious learning and memory: at a tipping point?" *Frontiers in human neuroscience* 6, p. 80.

Hannula, Deborah E and Charan Ranganath (2009). "The eyes have it: hippocampal activity predicts expression of memory in eye movements". *Neuron* 63:5, pp. 592–599.

Hannula, Deborah E, Jennifer D Ryan, Daniel Tranel, and Neal J Cohen (2007). "Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia". *Journal of cognitive neuroscience* 19:10, pp. 1690–1705.

Hassabis, Demis, Dharshan Kumaran, and Eleanor A Maguire (2007). "Using imagination to understand the neural basis of episodic memory". *Journal of neuroscience* 27:52, pp. 14365–14374.

Hassabis, Demis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire (2007). "Patients with hippocampal amnesia cannot imagine

new experiences". *Proceedings of the National Academy of Sciences* 104:5, pp. 1726–1731.

Hassabis, Demis and Eleanor A Maguire (2009). "The construction system of the brain". *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:1521, pp. 1263–1271.

Hasselmo, Michael E, Clara Bodelón, and Bradley P Wyble (2002). "A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning". *Neural computation* 14:4, pp. 793–817.

Hasselmo, Michael E and Howard Eichenbaum (2005). "Hippocampal mechanisms for the context-dependent retrieval of episodes". *Neural networks* 18:9, pp. 1172–1190.

Hasselmo, Michael E and Chantal E Stern (2006). "Mechanisms underlying working memory for novel information". *Trends in cognitive sciences* 10:11, pp. 487–493.

Hayden, Benjamin Y and Michael L Platt (2010). "Neurons in anterior cingulate cortex multiplex information about reward and action". *Journal of Neuroscience* 30:9, pp. 3339–3346.

Head, Tim, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi (2021). *scikit-optimize/scikit-optimize*. Version v0.9.0. DOI: 10.5281/zenodo.5565057. URL: https://doi.org/10.5281/zenodo.5565057.

Hebb, Donald Olding (1949). *The organisation of behaviour: a neuropsychological theory*. Science Editions New York.

Heckers, Stephan, Martin Zalesak, Anthony P Weiss, Tali Ditman, and Debra Titone (2004). "Hippocampal activation during transitive inference in humans". *Hippocampus* 14:2, pp. 153–162.

Henke, Katharina (2010). "A model for memory systems based on processing modes rather than consciousness". *Nature Reviews Neuroscience* 11:7, pp. 523–532.

Henke, Katharina, Christian RA Mondadori, Valerie Treyer, Roger M Nitsch, Alfred Buck, and Christoph Hock (2003). "Nonconscious formation and reactivation of semantic associations by way of the medial temporal lobe". *Neuropsychologia* 41:8, pp. 863–876.

Henke, Katharina, Valerie Treyer, Eva Turi Nagy, Stefan Kneifel, Max Dürsteler, Roger M Nitsch, and Alfred Buck (2003). "Active hippocampus during nonconscious memories". *Consciousness and cognition* 12:1, pp. 31–48.

Henze, Darrell A, Nathaniel N Urban, and German Barrionuevo (2000). "The multifarious hippocampal mossy fiber pathway: a review". *Neuroscience* 98:3, pp. 407–427.

Heys, James G, Krsna V Rangarajan, and Daniel A Dombeck (2014). "The functional micro-organization of grid cells revealed by cellular-resolution imaging". *Neuron* 84:5, pp. 1079–1090.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". *Neural computation* 9:8, pp. 1735–1780.

Holroyd, Clay B and Nick Yeung (2012). "Motivation of extended behaviors by anterior cingulate cortex". *Trends in cognitive sciences* 16:2, pp. 122–128.

Hölscher, Christian, Roger Anwyl, and Michael J Rowan (1997). "Stimulation on the positive phase of hippocampal theta rhythm induces long-term potentiation that can be depotentiated by stimulation on the negative phase in area CA1 in vivo". *Journal of Neuroscience* 17:16, pp. 6470–6477.

Hoover, Walter B and Robert P Vertes (2007). "Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat". *Brain Structure and Function* 212, pp. 149–179.

Hopfield, John J (1982). "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the national academy of sciences* 79:8, pp. 2554–2558.

– (1984). "Neurons with graded response have collective computational properties like those of two-state neurons." *Proceedings of the national academy of sciences* 81:10, pp. 3088–3092.

Howard, Marc W and Michael J Kahana (2002). "A distributed representation of temporal context". *Journal of mathematical psychology* 46:3, pp. 269–299.

Hyman, James M, Eric A Zilli, Amanda M Paley, and Michael E Hasselmo (2005). "Medial prefrontal cortex cells show dynamic modulation with

the hippocampal theta rhythm dependent on behavior". *Hippocampus* 15:6, pp. 739–749.

Hyman, James M, Eric A Zilli, Amanda M Paley, and Michael E Hasselmo (2010). "Working memory performance correlates with prefrontal-hippocampal theta interactions but not with prefrontal neuron firing rates". *Frontiers in integrative neuroscience* 4, p. 1162.

Ising, Ernst (1925). "Beitrag zur theorie des ferromagnetismus". *Zeitschrift für Physik* 31:1, pp. 253–258.

Ito, Hiroshi T, Edvard I Moser, and May-Britt Moser (2018). "Supramammillary nucleus modulates spike-time coordination in the prefrontal-thalamo-hippocampal circuit during navigation". *Neuron* 99:3, pp. 576–587.

Ito, Hiroshi T, Sheng-Jia Zhang, Menno P Witter, Edvard I Moser, and May-Britt Moser (2015). "A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation". *Nature* 522:7554, pp. 50–55.

Itskov, Vladimir, Carina Curto, Eva Pastalkova, and György Buzsáki (2011). "Cell assembly sequences arising from spike threshold adaptation keep track of time in the hippocampus". *Journal of Neuroscience* 31:8, pp. 2828–2834.

Jadhav, Shantanu P, Caleb Kemere, P Walter German, and Loren M Frank (2012). "Awake hippocampal sharp-wave ripples support spatial memory". *Science* 336:6087, pp. 1454–1458.

Jadhav, Shantanu P, Gideon Rothschild, Demetris K Roumis, and Loren M Frank (2016). "Coordinated excitation and inhibition of prefrontal ensembles during awake hippocampal sharp-wave ripple events". *Neuron* 90:1, pp. 113–127.

Janowsky, Jeri S, Arthur P Shimamura, and Larry R Squire (1989). "Source memory impairment in patients with frontal lobe lesions". *Neuropsychologia* 27:8, pp. 1043–1056.

Jay, Thérèse M, Jacques Glowinski, and Anne-Marie Thierry (1989). "Selectivity of the hippocampal projection to the prelimbic area of the prefrontal cortex in the rat". *Brain research* 505:2, pp. 337–340.

Jay, Thérèse M and Menno P Witter (1991). "Distribution of hippocampal CA1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of Phaseolus vulgaris-leucoagglutinin". *Journal of Comparative Neurology* 313:4, pp. 574–586.

Jayachandran, Maanasa, Stephanie B Linley, Maximilian Schlecht, Stephen V Mahler, Robert P Vertes, and Timothy A Allen (2019). "Prefrontal pathways provide top-down control of memory for sequences of events". *Cell reports* 28:3, pp. 640–654.

Jensen, Kristopher T, Guillaume Hennequin, and Marcelo G Mattar (2023). "A recurrent network model of planning explains hippocampal replay and human behavior". *bioRxiv*, pp. 2023–01.

Jensen, Ole, John E Lisman, et al. (1996). "Novel lists of 7+/-2 known items can be reliably stored in an oscillatory short-term memory network: interaction with long-term memory". *Learn Mem* 3:2-3, pp. 257–263.

Jensen, Ole and John E Lisman (1996). "Theta/gamma networks with slow NMDA channels learn sequences and encode episodic memory: role of NMDA channels in recall." *Learning & Memory* 3:2-3, pp. 264–278.

– (2000). "Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding". *Journal of neurophysiology* 83:5, pp. 2602–2609.

– (2005). "Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer". *Trends in neurosciences* 28:2, pp. 67–72.

Ji, Daoyun and Matthew A Wilson (2007). "Coordinated memory replay in the visual cortex and hippocampus during sleep". *Nature neuroscience* 10:1, pp. 100–107.

Jocham, Gerhard, Tilmann A Klein, and Markus Ullsperger (2011). "Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices". *Journal of Neuroscience* 31:5, pp. 1606–1613.

Jones, Matthew W and Matthew A Wilson (2005). "Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task". *PLoS biology* 3:12, e402.

Jung, Min W, Sidney I Wiener, and Bruce L McNaughton (1994). "Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat". *Journal of Neuroscience* 14:12, pp. 7347–7356.

Jung, MW and BL McNaughton (1993). "Spatial selectivity of unit activity in the hippocampal granular layer". *Hippocampus* 3:2, pp. 165–182.

Kanerva, Pentti (1988). *Sparse distributed memory*. MIT press.

Kassab, Randa and Frédéric Alexandre (2018). "Pattern separation in the hippocampus: distinct circuits under different conditions". *Brain Structure and Function* 223:6, pp. 2785–2808.

Keinath, Alexandra T, Andrés Nieto-Posadas, Jennifer C Robinson, and Mark P Brandon (2020). "DG–CA3 circuitry mediates hippocampal representations of latent information". *Nature communications* 11:1, p. 3026.

Kennerley, Steven W, Timothy EJ Behrens, and Jonathan D Wallis (2011). "Double dissociation of value computations in orbitofrontal and anterior cingulate neurons". *Nature neuroscience* 14:12, pp. 1581–1589.

Kesner, Raymond P, Paul E Gilbert, and Lindsay A Barua (2002). "The role of the hippocampus in memory for the temporal order of a sequence of odors." *Behavioral neuroscience* 116:2, p. 286.

Ketz, Nicholas, Srinimisha G Morkonda, and Randall C O'Reilly (2013). "Theta coordinated error-driven learning in the hippocampus". *PLoS computational biology* 9:6, e1003067.

Killian, Nathaniel J, Michael J Jutras, and Elizabeth A Buffalo (2012). "A map of visual space in the primate entorhinal cortex". *Nature* 491:7426, pp. 761–764.

Kishimoto, Yasushi, Kazu Nakazawa, Susumu Tonegawa, Yutaka Kirino, and Masanobu Kano (2006). "Hippocampal CA3 NMDA receptors are crucial for adaptive timing of trace eyeblink conditioned response". *Journal of Neuroscience* 26:5, pp. 1562–1570.

Kjelstrup, Kirsten Brun, Trygve Solstad, Vegard Heimly Brun, Torkel Hafting, Stefan Leutgeb, Menno P Witter, Edvard I Moser, and May-Britt Moser (2008). "Finite scale of spatial representation in the hippocampus". *Science* 321:5885, pp. 140–143.

Koechlin, Etienne, Chrystele Ody, and Frédérique Kouneiher (2003). "The architecture of cognitive control in the human prefrontal cortex". *Science* 302:5648, pp. 1181–1185.

Komorowski, Robert W, Carolyn G Garcia, Alix Wilson, Shoai Hattori, Marc W Howard, and Howard Eichenbaum (2013). "Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts". *Journal of Neuroscience* 33:18, pp. 8079–8087.

Konkel, Alex and Neal J Cohen (2009). "Relational memory and the hippocampus: representations and methods". *Frontiers in neuroscience* 3, p. 679.

Kouneiher, Frédérique, Sylvain Charron, and Etienne Koechlin (2009). "Motivation and cognitive control in the human prefrontal cortex". *Nature neuroscience* 12:7, pp. 939–945.

Krotov, Dmitry (2023). "A new frontier for Hopfield networks". *Nature Reviews Physics*, pp. 1–2.

Krotov, Dmitry and John Hopfield (2020). "Large associative memory problem in neurobiology and machine learning". *arXiv preprint arXiv:2008.06996*.

Krotov, Dmitry and John J Hopfield (2016). "Dense associative memory for pattern recognition". *Advances in neural information processing systems* 29.

Kubota, Kisou and Hiroaki Niki (1971). "Prefrontal cortical unit activity and delayed alternation performance in monkeys." *Journal of neurophysiology* 34:3, pp. 337–347.

Kudrimoti, Hemant S, Carol A Barnes, and Bruce L McNaughton (1999). "Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics". *Journal of Neuroscience* 19:10, pp. 4090–4101.

Kumaran, Dharshan, Demis Hassabis, and James L McClelland (2016). "What learning systems do intelligent agents need? Complementary learning systems theory updated". *Trends in cognitive sciences* 20:7, pp. 512–534.

Kumaran, Dharshan, Hans Ludwig Melo, and Emrah Duzel (2012). "The emergence and representation of knowledge about social and nonsocial hierarchies". *Neuron* 76:3, pp. 653–666.

Kunec, Steve, Michael E Hasselmo, and Nancy Kopell (2005). "Encoding and retrieval in the CA3 region of the hippocampus: a model of theta-phase separation". *Journal of neurophysiology* 94:1, pp. 70–82.

Lassalle, Jean-Michel, Thierry Bataille, and Hélene Halley (2000). "Reversible inactivation of the hippocampal mossy fiber synapses in mice impairs spatial learning, but neither consolidation nor memory retrieval, in the Morris navigation task". *Neurobiology of learning and memory* 73:3, pp. 243–257.

Lau, Hakwan C and Richard E Passingham (2007). "Unconscious activation of the cognitive control system in the human prefrontal cortex". *Journal of Neuroscience* 27:21, pp. 5805–5811.

Le, Hung, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, and Svetha Venkatesh (2021). "Model-based episodic memory induces dynamic hybrid controls". *Advances in Neural Information Processing Systems* 34, pp. 30313–30325.

Lee, Albert K and Matthew A Wilson (2002). "Memory of sequential experience in the hippocampus during slow wave sleep". *Neuron* 36:6, pp. 1183–1194.

Lengyel, Máté and Peter Dayan (2007). "Hippocampal contributions to control: the third way". *Advances in neural information processing systems* 20.

Lenz, Wilhelm (1920). "Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern". *Z. Phys.* 21, pp. 613–615.

Levy, William B (1982). "Associative encoding at synapses". In: *Proceedings of the Fourth Annual Conference of Cognitive Science Society*, pp. 135–136.

– (1996). "A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks". *Hippocampus* 6:6, pp. 579–590.

Li, Minghua, Jessica L Kinney, Yu-Qiu Jiang, Daniel K Lee, Qiwen Wu, Daehoon Lee, Wen-Cheng Xiong, and Qian Sun (2023). "Hypothalamic supramammillary nucleus selectively excites hippocampal CA3 interneurons to suppress CA3 pyramidal neuron activity". *Journal of Neuroscience* 43:25, pp. 4612–4624.

Lim, Seung-Lark, John P O'Doherty, and Antonio Rangel (2011). "The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention". *Journal of Neuroscience* 31:37, pp. 13214–13223.

Lin, Long-Ji (1992). "Self-improving reactive agents based on reinforcement learning, planning and teaching". *Machine learning* 8, pp. 293–321.

Linnainmaa, Seppo (1976). "Taylor expansion of the accumulated rounding error". *BIT Numerical Mathematics* 16:2, pp. 146–160.

Lisman, John E (1999). "Relating hippocampal circuitry to function: recall of memory sequences by reciprocal dentate–CA3 interactions". *Neuron* 22:2, pp. 233–242.

Lisman, John E and Anthony A Grace (2005). "The hippocampal-VTA loop: controlling the entry of information into long-term memory". *Neuron* 46:5, pp. 703–713.

Lisman, John E and Marco AP Idiart (1995). "Storage of 7 ± 2 short-term memories in oscillatory subcycles". *Science* 267:5203, pp. 1512–1515.

Lisman, John E and Ole Jensen (2013). "The theta-gamma neural code". *Neuron* 77:6, pp. 1002–1016.

Lisman, John E and Nonna A Otmakhova (2001). "Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine". *Hippocampus* 11:5, pp. 551–568.

Lisman, John E, Lucia M Talamini, and Antonino Raffone (2005). "Recall of memory sequences by interaction of the dentate and CA3: a revised model of the phase precession". *Neural Networks* 18:9, pp. 1191–1201.

Long, Michael A and Michale S Fee (2008). "Using temperature to analyse temporal dynamics in the songbird motor pathway". *Nature* 456:7219, pp. 189–194.

Lopez-Rojas, Jeffrey and Michael R Kreutz (2016). "Mature granule cells of the dentate gyrus - Passive bystanders or principal performers in hippocampal function?" *Neuroscience & Biobehavioral Reviews* 64, pp. 167–174.

Lu, Qihong, Uri Hasson, and Kenneth A Norman (2022). "A neural network model of when to retrieve and encode episodic memories". *elife* 11, e74445.

MacDonald, Christopher J, Kyle Q Lepage, Uri T Eden, and Howard Eichenbaum (2011). "Hippocampal "time cells" bridge the gap in memory for discontiguous events". *Neuron* 71:4, pp. 737–749.

MacKay, Donald G, Deborah M Burke, and Rachel Stewart (1998). "HM's language production deficits: Implications for relations between memory, semantic binding, and the hippocampal system". *Journal of Memory and Language* 38:1, pp. 28–69.

Manns, Joseph R and Howard Eichenbaum (2006). "Evolution of declarative memory". *Hippocampus* 16:9, pp. 795–808.

Marr, D (1971). "Simple memory: a theory for archicortex". *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 262:841, pp. 23–81.

Marr, David and W Thomas Thach (1991). "A theory of cerebellar cortex". *From the Retina to the Neocortex: Selected Papers of David Marr*, pp. 11–50.

Mattar, Marcelo G and Nathaniel D Daw (2018). "Prioritized memory access explains planning and hippocampal replay". *Nature neuroscience* 21:11, pp. 1609–1617.

Maurer, Andrew P, Shea R VanRhoads, Gary R Sutherland, Peter Lipa, and Bruce L McNaughton (2005). "Self-motion and the origin of differential spatial scaling along the septo-temporal axis of the hippocampus". *Hippocampus* 15:7, pp. 841–852.

McClelland, James L (2013). "Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory." *Journal of Experimental Psychology: General* 142:4, p. 1190.

McClelland, James L, Bruce L McNaughton, and Randall C O'Reilly (1995). "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." *Psychological review* 102:3, p. 419.

McCloskey, Michael and Neal J Cohen (1989). "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation*. Vol. 24. Elsevier, pp. 109–165.

McHugh, Thomas J, Matthew W Jones, Jennifer J Quinn, Nina Balthasar, Roberto Coppari, Joel K Elmquist, Bradford B Lowell, Michael S Fanselow, Matthew A Wilson, and Susumu Tonegawa (2007). "Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network". *Science* 317:5834, pp. 94–99.

McNaughton, Bruce L, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser (2006). "Path integration and the neural basis of the'cognitive map'". *Nature Reviews Neuroscience* 7:8, pp. 663–678.

McNaughton, Bruce L and Richard GM Morris (1987). "Hippocampal synaptic enhancement and information storage within a distributed memory system". *Trends in neurosciences* 10:10, pp. 408–415.

Mehta, Mayank R, Carol A Barnes, and Bruce L McNaughton (1997). "Experience-dependent, asymmetric expansion of hippocampal place fields". *Proceedings of the National Academy of Sciences* 94:16, pp. 8918–8921.

Melchior, Jan, Mehdi Bayati, Amir Azizi, Sen Cheng, and Laurenz Wiskott (2019). "A hippocampus model for online one-shot storage of pattern sequences". *arXiv preprint arXiv:1905.12937*.

Melo, Luckeciano C (2022). "Transformers are meta-reinforcement learners". In: *international conference on machine learning*. PMLR, pp. 15340–15359.

Menzel, Randolf (2023). "Navigation and dance communication in honeybees: a cognitive perspective". *Journal of Comparative Physiology A*, pp. 1–13.

Miller, Earl K and Jonathan D Cohen (2001). "An integrative theory of prefrontal cortex function". *Annual review of neuroscience* 24:1, pp. 167–202.

Miller, Michael H and J Orbach (1972). "Retention of spatial alternation following frontal lobe resections in stump-tailed macaques". *Neuropsychologia* 10:3, pp. 291–298.

Millidge, Beren, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz (2022). "Universal hopfield networks: A general frame-

work for single-shot associative memory models". In: *International Conference on Machine Learning*. PMLR, pp. 15561–15583.

Milner, Brenda (1962). "Les troubles de la memoire accompagnant des lesions hippocampiques bilaterales". *Physiologie de l'hippocampe* 107, pp. 257–272.

Milner, Brenda, Suzanne Corkin, and H-L Teuber (1968). "Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM". *Neuropsychologia* 6:3, pp. 215–234.

Miyashita, Teiko, Stepan Kubik, Nahideh Haghighi, Oswald Steward, and John F Guzowski (2009). "Rapid activation of plasticity-associated gene transcription in hippocampal neurons provides a mechanism for encoding of one-trial experience". *Journal of Neuroscience* 29:4, pp. 898–906.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. (2015). "Human-level control through deep reinforcement learning". *nature* 518:7540, pp. 529–533.

Momennejad, Ida and Marc W Howard (2018). "Predicting the future with multi-scale successor representations". *BioRxiv*, p. 449470.

Moore, Andrew W and Christopher G Atkeson (1993). "Prioritized sweeping: Reinforcement learning with less data and less time". *Machine learning* 13, pp. 103–130.

Moscovitch, Morris (1992). "Memory and working-with-memory: A component process model based on modules and central systems". *Journal of cognitive neuroscience* 4:3, pp. 257–267.

Moser, May-Britt, Edvard I Moser, Elma Forrest, Per Andersen, and RG40869 Morris (1995). "Spatial learning with a minislab in the dorsal hippocampus." *Proceedings of the National Academy of Sciences* 92:21, pp. 9697–9701.

Muller, Robert U and John L Kubie (1987). "The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells". *Journal of Neuroscience* 7:7, pp. 1951–1968.

Nádasdy, Zoltán, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki (1999). "Replay and time compression of recurring spike

sequences in the hippocampus". *Journal of Neuroscience* 19:21, pp. 9497–9507.

Nadel, Lynn and Morris Moscovitch (1997). "Memory consolidation, retrograde amnesia and the hippocampal complex". *Current opinion in neurobiology* 7:2, pp. 217–227.

Nadel, Lynn, A Samsonovich, Lee Ryan, and Morris Moscovitch (2000). "Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results". *Hippocampus* 10:4, pp. 352–368.

Nakamura, Nozomu H, Masaki Fukunaga, Keith T Akama, Tomoko Soga, Sonoko Ogawa, and Constantine Pavlides (2010). "Hippocampal cells encode places by forming small anatomical clusters". *Neuroscience* 166:3, pp. 994–1007.

Nakazawa, Kazu, Linus D Sun, Michael C Quirk, Laure Rondi-Reig, Matthew A Wilson, and Susumu Tonegawa (2003). "Hippocampal CA3 NMDA receptors are crucial for memory acquisition of one-time experience". *Neuron* 38:2, pp. 305–315.

Navawongse, Rapeechai and Howard Eichenbaum (2013). "Distinct pathways for rule-based retrieval and spatial mapping of memory representations in hippocampal neurons". *Journal of Neuroscience* 33:3, pp. 1002–1013.

Nayebi, Aran, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, et al. (2021). "Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks". *Advances in Neural Information Processing Systems* 34, pp. 12167–12179.

Nee, Derek Evan and Joshua W Brown (2012). "Rostral–caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory". *Neuroimage* 63:3, pp. 1285–1294.

Neher, Torsten, Sen Cheng, and Laurenz Wiskott (2015). "Memory storage fidelity in the hippocampal circuit: the role of subregions and input statistics". *PLoS computational biology* 11:5, e1004250.

Niki, Hiroaki and Masataka Watanabe (1979). "Prefrontal and cingulate unit activity during timing behavior in the monkey". *Brain research* 171:2, pp. 213–224.

Noonan, MP, RB Mars, and MFS Rushworth (2011). "Distinct roles of three frontal cortical areas in reward-guided behavior". *Journal of Neuroscience* 31:40, pp. 14399–14412.

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

O'Mara, Shane (2005). "The subiculum: what it does, what it might do, and what neuroanatomy has yet to tell us". *Journal of anatomy* 207:3, pp. 271–282.

O'Mara, Shane M, Sean Commins, Michael Anderson, and John Gigg (2001). "The subiculum: a review of form, physiology and function". *Progress in neurobiology* 64:2, pp. 129–155.

O'Neill, Pia-Kelsey, Joshua A Gordon, and Torfi Sigurdsson (2013). "Theta oscillations in the medial prefrontal cortex are modulated by spatial working memory and synchronize with the hippocampus through its ventral subregion". *Journal of Neuroscience* 33:35, pp. 14211–14224.

O'Reilly, Randall C and Michael J Frank (2006). "Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia". *Neural computation* 18:2, pp. 283–328.

O'Reilly, Randall C, Michael J Frank, Thomas E Hazy, and Brandon Watz (2007). "PVLV: the primary value and learned value Pavlovian learning algorithm." *Behavioral neuroscience* 121:1, p. 31.

O'Reilly, Randall C and James L McClelland (1994). "Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off". *Hippocampus* 4:6, pp. 661–682.

Park, Seongmin A, Maryam Zolfaghar, Jacob Russin, Douglas S Miller, Randall C O'Reilly, and Erie D Boorman (2023). "The representational geometry of cognitive maps under dynamic cognitive control". *bioRxiv*, pp. 2023–02.

Park, Sungmo, Emily E Kramer, Valentina Mercaldo, Asim J Rashid, Nathan Insel, Paul W Frankland, and Sheena A Josselyn (2016). "Neuronal allocation to a hippocampal engram". *Neuropsychopharmacology* 41:13, pp. 2987–2993.

Pasquier, Daniel A and Fernando Reinoso-Suarez (1976). "Direct projections from hypothalamus to hippocampus in the rat demonstrated by retrograde transport of horseradish peroxidase". *Brain Research* 108:1, pp. 165–169.

Pavlides, Constantine and Jonathan Winson (1989). "Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes". *Journal of neuroscience* 9:8, pp. 2907–2918.

Persuh, Marjan, Eric LaRock, and Jacob Berger (2018). "Working memory and consciousness: The current state of play". *Frontiers in human neuroscience* 12, p. 78.

Peters, Gregory J, Christopher N David, Madison D Marcus, and David M Smith (2013). "The medial prefrontal cortex is critical for memory retrieval and resolving interference". *Learning & memory* 20:4, pp. 201–209.

Peyrache, Adrien, Mehdi Khamassi, Karim Benchenane, Sidney I Wiener, and Francesco P Battaglia (2009). "Replay of rule-learning related neural patterns in the prefrontal cortex during sleep". *Nature neuroscience* 12:7, pp. 919–926.

Pezzulo, Giovanni, Caleb Kemere, and Matthijs AA Van Der Meer (2017). "Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition". *Annals of the New York Academy of Sciences* 1396:1, pp. 144–165.

Pfeiffer, Brad E and David J Foster (2013). "Hippocampal place-cell sequences depict future paths to remembered goals". *Nature* 497:7447, pp. 74–79.

Pilly, Praveen, Michael Howard, and Rajan Bhattacharyya (2018). "Modeling contextual modulation of memory associations in the hippocampus". *Frontiers in Human Neuroscience* 12, p. 442.

Place, Ryan, Anja Farovik, Marco Brockmann, and Howard Eichenbaum (2016). "Bidirectional prefrontal-hippocampal interactions support context-guided memory". *Nature neuroscience* 19:8, pp. 992–994.

Posner, Michael I, Charles R Snyder, and R Solso (2004). "Attention and cognitive control". *Cognitive psychology: Key readings* 205, pp. 55–85.

Preston, Alison R and Howard Eichenbaum (2013). "Interplay of hippocampus and prefrontal cortex in memory". *Current biology* 23:17, R764–R773.

Preston, Alison R, Yael Shrager, Nicole M Dudukovic, and John DE Gabrieli (2004). "Hippocampal contribution to the novel use of relational information in declarative memory". *Hippocampus* 14:2, pp. 148–152.

Pritzel, Alexander, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell (2017). "Neural episodic control". In: *International conference on machine learning*. PMLR, pp. 2827–2836.

Proffitt, Dennis R (2006). "Embodied perception and the economy of action". *Perspectives on psychological science* 1:2, pp. 110–122.

Purves, Dale, Elizabeth M Brannon, Roberto Cabeza, Scott A Huettel, Kevin S LaBar, Michael L Platt, and Marty G Woldorff (2013). *Principles of cognitive neuroscience*. Vol. 83. 3. Sinauer Associates Sunderland.

Quirk, Gregory J, Robert U Muller, and John L Kubie (1990). "The firing of hippocampal place cells in the dark depends on the rat's recent experience". *Journal of Neuroscience* 10:6, pp. 2008–2017.

Rajasethupathy, Priyamvada, Sethuraman Sankaran, James H Marshel, Christina K Kim, Emily Ferenczi, Soo Yeun Lee, Andre Berndt, Charu Ramakrishnan, Anna Jaffe, Maisie Lo, et al. (2015). "Projections from neocortex mediate top-down control of memory retrieval". *Nature* 526:7575, pp. 653–659.

Ramsauer, Hubert, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. (2020). "Hopfield networks is all you need". *arXiv preprint arXiv:2008.02217*.

Reser, Jared Edward (2022). "A Cognitive Architecture for Machine Consciousness and Artificial Superintelligence: Updating Working Memory Iteratively". *arXiv preprint arXiv:2203.17255*.

Reynolds, Jeremy R, Randall C O'Reilly, Jonathan D Cohen, and Todd S Braver (2012). "The function and organization of lateral prefrontal cortex: a test of competing hypotheses". *PloS one* 7:2, e30284.

Ritter, Samuel, Jane Wang, Zeb Kurth-Nelson, Siddhant Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick (2018). "Been there, done that: Meta-learning with episodic recall". In: *International conference on machine learning*. PMLR, pp. 4354–4363.

Rosenbaum, R Shayna, Sandra Priselac, Stefan Köhler, Sandra E Black, Fuqiang Gao, Lynn Nadel, and Morris Moscovitch (2000). "Remote spatial memory in an amnesic person with extensive bilateral hippocampal lesions". *Nature neuroscience* 3:10, pp. 1044–1048.

Rumelhart, David E (1990). "Brain style computation: Learning and generalization". In: *An introduction to neural and electronic networks*, pp. 405–420.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". *nature* 323:6088, pp. 533–536.

Rushworth, Matthew FS, Mark J Buckley, Timothy EJ Behrens, Mark E Walton, and David M Bannerman (2007). "Functional organization of the medial frontal cortex". *Current opinion in neurobiology* 17:2, pp. 220–227.

Rushworth, MFS, Mark E Walton, Steve W Kennerley, and DM Bannerman (2004). "Action sets and decisions in the medial frontal cortex". *Trends in cognitive sciences* 8:9, pp. 410–417.

Russin, Jacob, Randall C O'Reilly, and Yoshua Bengio (2020). "Deep learning needs a prefrontal cortex". *Work Bridging AI Cogn Sci* 107:603-616, p. 1.

Sabariego, Marta, Antonia Schönwald, Brittney L Boublil, David T Zimmerman, Siavash Ahmadi, Nailea Gonzalez, Christian Leibold, Robert E Clark, Jill K Leutgeb, and Stefan Leutgeb (2019). "Time cells in the hippocampus are neither dependent on medial entorhinal cortex inputs nor necessary for spatial working memory". *Neuron* 102:6, pp. 1235–1248.

Sakurai, Y (2002). "Coding of auditory temporal and pitch information by hippocampal individual cells and cell assemblies in the rat". *Neuroscience* 115:4, pp. 1153–1163.

Sankar, Remya (2022). "Dual pathway architecture underlying vocal learning in songbirds". Theses. Université de Bordeaux. URL: https://hal.science/tel-03910872.

Santos-Pata, Diogo, Adrián F Amil, Ivan Georgiev Raikov, César Rennó-Costa, Anna Mura, Ivan Soltesz, and Paul FMJ Verschure (2021). "Entorhinal mismatch: A model of self-supervised learning in the hippocampus". *Iscience* 24:4.

Sarrico, Marta, Kai Arulkumaran, Andrea Agostinelli, Pierre Richemond, and Anil Anthony Bharath (2019). "Sample-efficient reinforcement learning with maximum entropy mellowmax episodic control". *arXiv preprint arXiv:1911.09615*.

Schacter, Daniel L, Donna Rose Addis, and Randy L Buckner (2007). "Remembering the past to imagine the future: the prospective brain". *Nature reviews neuroscience* 8:9, pp. 657–661.

Schacter, Daniel L, Donna Rose Addis, Demis Hassabis, Victoria C Martin, R Nathan Spreng, and Karl K Szpunar (2012). "The future of memory: remembering, imagining, and the brain". *Neuron* 76:4, pp. 677–694.

Schaeffer, Rylan, Mikail Khona, and Ila Fiete (2022). "No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit". *Advances in Neural Information Processing Systems* 35, pp. 16052–16067.

Schapiro, Anna C, Lauren V Kustner, and Nicholas B Turk-Browne (2012). "Shaping of object representations in the human medial temporal lobe based on temporal regularities". *Current biology* 22:17, pp. 1622–1627.

Schapiro, Anna C, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman (2017). "Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning". *Philosophical Transactions of the Royal Society B: Biological Sciences* 372:1711, p. 20160049.

Schapiro, Anna C, Nicholas B Turk-Browne, Kenneth A Norman, and Matthew M Botvinick (2016). "Statistical learning of temporal community structure in the hippocampus". *Hippocampus* 26:1, pp. 3–8.

Scharfman, Helen E (2007). "The CA3 "backprojection" to the dentate gyrus". *Progress in brain research* 163, pp. 627–637.

Schaul, Tom, John Quan, Ioannis Antonoglou, and David Silver (2015). "Prioritized experience replay". *arXiv preprint arXiv:1511.05952*.

Schmidt-Hieber, Christoph, Peter Jonas, and Josef Bischofberger (2004). "Enhanced synaptic plasticity in newly generated granule cells of the adult hippocampus". *Nature* 429:6988, pp. 184–187.

Schnider, Armin and Radek Ptak (1999). "Spontaneous confabulators fail to suppress currently irrelevant memory traces". *Nature neuroscience* 2:7, pp. 677–681.

Schomburg, Erik W, Antonio Fernández-Ruiz, Kenji Mizuseki, Antal Berényi, Costas A Anastassiou, Christof Koch, and György Buzsáki (2014). "Theta phase segregation of input-specific gamma patterns in entorhinal-hippocampal networks". *Neuron* 84:2, pp. 470–485.

Schon, Karin, Randall E Newmark, Robert S Ross, and Chantal E Stern (2016). "A working memory buffer in parahippocampal regions: evidence from a load effect during the delay period". *Cerebral Cortex* 26:5, pp. 1965–1974.

Schrauwen, Benjamin, David Verstraeten, and Jan Van Campenhout (2007). "An overview of reservoir computing: theory, applications and implementations". In: *Proceedings of the 15th european symposium on artificial neural networks. p. 471-482 2007*, pp. 471–482.

Scoville, William Beecher and Brenda Milner (1957). "Loss of recent memory after bilateral hippocampal lesions". *Journal of neurology, neurosurgery, and psychiatry* 20:1, p. 11.

Seress, Laszlo (2010). *Hippocampus and seahorse cropped*. URL: https://commons.wikimedia.org/wiki/File:Hippocampus_and_seahorse_cropped.JPG.

Sesack, Susan R, Ariel Y Deutch, Robert H Roth, and Benjamin S Bunney (1989). "Topographical organization of the efferent projections of the medial prefrontal cortex in the rat: an anterograde tract-tracing study with Phaseolus vulgaris leucoagglutinin". *Journal of Comparative Neurology* 290:2, pp. 213–242.

Shimamura, Arthur P, Paul J Jurica, Jennifer A Mangels, Felick B Gershberg, and Robert T Knight (1995). "Susceptibility to memory interfer-

ence effects following frontal lobe damage: Findings from tests of paired-associate learning". *Journal of cognitive neuroscience* 7:2, pp. 144–152.

Shors, Tracey J, George Miesegaes, Anna Beylin, Mingrui Zhao, Tracy Rydel, and Elizabeth Gould (2001). "Neurogenesis in the adult is involved in the formation of trace memories". *Nature* 410:6826, pp. 372–376.

Siapas, Athanassios G, Evgueniy V Lubenov, and Matthew A Wilson (2005). "Prefrontal phase locking to hippocampal theta oscillations". *Neuron* 46:1, pp. 141–151.

Skaggs, William E and Bruce L McNaughton (1996). "Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience". *Science* 271:5257, pp. 1870–1873.

Skaggs, William E, Bruce L McNaughton, Matthew A Wilson, and Carol A Barnes (1996). "Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences". *Hippocampus* 6:2, pp. 149–172.

Snow, Mallory A and Jeff Orchard (2022). "Biological softmax: Demonstrated in modern Hopfield networks". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44.

Solstad, Trygve, Edvard I Moser, and Gaute T Einevoll (2006). "From grid cells to place cells: a mathematical model". *Hippocampus* 16:12, pp. 1026–1031.

Sorscher, Ben, Gabriel Mel, Surya Ganguli, and Samuel Ocko (2019). "A unified theory for the origin of grid cells through the lens of pattern formation". *Advances in neural information processing systems* 32.

Spellman, Timothy, Mattia Rigotti, Susanne E Ahmari, Stefano Fusi, Joseph A Gogos, and Joshua A Gordon (2015). "Hippocampal–prefrontal input supports spatial encoding in working memory". *Nature* 522:7556, pp. 309–314.

Spens, Eleanor and Neil Burgess (2023). "A generative model of memory construction and consolidation". *bioRxiv*, pp. 2023–01.

Spreng, R Nathan, Raymond A Mar, and Alice SN Kim (2009). "The common neural basis of autobiographical memory, prospection, navigation,

theory of mind, and the default mode: a quantitative meta-analysis". *Journal of cognitive neuroscience* 21:3, pp. 489–510.

Squire, Larry R (2009). "Memory and brain systems: 1969–2009". *Journal of Neuroscience* 29:41, pp. 12711–12716.

Squire, Larry R and Pablo Alvarez (1995). "Retrograde amnesia and memory consolidation: a neurobiological perspective". *Current opinion in neurobiology* 5:2, pp. 169–177.

Stachenfeld, Kimberly L, Matthew M Botvinick, and Samuel J Gershman (2017). "The hippocampus as a predictive map". *Nature neuroscience* 20:11, pp. 1643–1653.

Stensola, Hanne, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I Moser (2012). "The entorhinal grid map is discretized". *Nature* 492:7427, pp. 72–78.

Stokes, Mark G (2015). ""Activity-silent"working memory in prefrontal cortex: a dynamic coding framework". *Trends in cognitive sciences* 19:7, pp. 394–405.

Strange, Bryan A, Andrew Duggins, William Penny, Raymond J Dolan, and Karl J Friston (2005). "Information theory, novelty and hippocampal responses: unpredicted or unpredictable?" *Neural Networks* 18:3, pp. 225–230.

Strange, Bryan A, Menno P Witter, Ed S Lein, and Edvard I Moser (2014). "Functional organization of the hippocampal longitudinal axis". *Nature Reviews Neuroscience* 15:10, pp. 655–669.

Sukhbaatar, Sainbayar, Jason Weston, Rob Fergus, et al. (2015). "End-to-end memory networks". *Advances in neural information processing systems* 28.

Sutton, Richard S and Andrew G Barto (2014). *Reinforcement learning: An introduction*. URL: https://inst.eecs.berkeley.edu/~cs188/sp20/assets/files/SuttonBartoIPRLBook2ndEd.pdf.

Szczepanski, Sara M and Robert T Knight (2014). "Insights into human behavior from lesions to the prefrontal cortex". *Neuron* 83:5, pp. 1002–1018.

Tahvildari, Babak, Angel A Alonso, and Charles W Bourque (2008). "Ionic basis of ON and OFF persistent activity in layer III lateral entorhinal cortical principal neurons". *Journal of neurophysiology* 99:4, pp. 2006–2011.

Tang, Wenbo, Justin D Shin, Loren M Frank, and Shantanu P Jadhav (2017). "Hippocampal-prefrontal reactivation during learning is stronger in awake compared with sleep states". *Journal of Neuroscience* 37:49, pp. 11789–11805.

Tayler, Kaycie K, Kazumasa Z Tanaka, Leon G Reijmers, and Brian J Wiltgen (2013). "Reactivation of neural ensembles during the retrieval of recent and remote memory". *Current Biology* 23:2, pp. 99–106.

Teyler, Timothy J and Pascal DiScenna (1986). "The hippocampal memory indexing theory." *Behavioral neuroscience* 100:2, p. 147.

Tolman, Edward C (1948). "Cognitive maps in rats and men." *Psychological review* 55:4, p. 189.

Toni, Nicolas, Diego A Laplagne, Chunmei Zhao, Gabriela Lombardi, Charles E Ribak, Fred H Gage, and Alejandro F Schinder (2008). "Neurons born in the adult dentate gyrus form functional synapses with target cells". *Nature neuroscience* 11:8, pp. 901–907.

Treves, Alessandro and Edmund T Rolls (1992). "Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network". *Hippocampus* 2:2, pp. 189–199.

– (1994). "Computational analysis of the role of the hippocampus in memory". *Hippocampus* 4:3, pp. 374–391.

Trinchero, Mariela F, Karina A Buttner, Jessica N Sulkes Cuevas, Silvio G Temprana, Paula A Fontanet, M Cristina Monzón-Salinas, Fernanda Ledda, Gustavo Paratcha, and Alejandro F Schinder (2017). "High plasticity of new granule cells in the aging hippocampus". *Cell reports* 21:5, pp. 1129–1139.

Trivers, Robert and Richard Dawkins (1976). *The selfish gene*.

Tse, Dorothy, Tomonori Takeuchi, Masaki Kakeyama, Yasushi Kajii, Hiroyuki Okuno, Chiharu Tohyama, Haruhiko Bito, and Richard GM Morris (2011). "Schema-dependent gene activation and memory encoding in neocortex". *Science* 333:6044, pp. 891–895.

Tulving, Endel (1972). "Episodic and Semantic Memory". *Organization of Memory*, p. 381.

Turk-Browne, Nicholas B, Brian J Scholl, Marvin M Chun, and Marcia K Johnson (2009). "Neural evidence of statistical learning: Efficient detection of visual regularities without awareness". *Journal of cognitive neuroscience* 21:10, pp. 1934–1945.

Tyulmankov, Danil, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang (2021). "Biological learning in key-value memory networks". *Advances in Neural Information Processing Systems* 34, pp. 22247–22258.

Van Gaal, Simon, K Richard Ridderinkhof, Johannes J Fahrenfort, H Steven Scholte, and Victor AF Lamme (2008). "Frontal cortex mediates unconsciously triggered inhibitory control". *Journal of Neuroscience* 28:32, pp. 8053–8062.

Van Gaal, Simon, K Richard Ridderinkhof, H Steven Scholte, and Victor AF Lamme (2010). "Unconscious activation of the prefrontal no-go network". *Journal of neuroscience* 30:11, pp. 4143–4150.

Van Kesteren, Marlieke TR, Dirk J Ruiter, Guillén Fernández, and Richard N Henson (2012). "How schema and novelty augment memory formation". *Trends in neurosciences* 35:4, pp. 211–219.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". *Advances in neural information processing systems* 30.

Vertes, Robert P (2002). "Analysis of projections from the medial prefrontal cortex to the thalamus in the rat, with emphasis on nucleus reuniens". *Journal of comparative neurology* 442:2, pp. 163–187.

– (2006). "Interactions among the medial prefrontal cortex, hippocampus and midline thalamus in emotional and cognitive processing in the rat". *Neuroscience* 142:1, pp. 1–20.

Vertes, Robert P, Walter B Hoover, Klara Szigeti-Buck, and Csaba Leranth (2007). "Nucleus reuniens of the midline thalamus: link between the medial prefrontal cortex and the hippocampus". *Brain research bulletin* 71:6, pp. 601–609.

Vertes, Robert P and James Timothy McKenna (2000). "Collateral projections from the supramammillary nucleus to the medial septum and hippocampus". *Synapse* 38:3, pp. 281–293.

Viskontas, Indre V, Mary Pat McAndrews, and Morris Moscovitch (2002). "Memory for famous people in patients with unilateral temporal lobe epilepsy and excisions." *Neuropsychology* 16:4, p. 472.

Vyleta, Nicholas P, Carolina Borges-Merjane, and Peter Jonas (2016). "Plasticity-dependent, full detonation at hippocampal mossy fiber–CA3 pyramidal neuron synapses". *Elife* 5, e17977.

Wallenstein, Gene V and Michael E Hasselmo (1997). "GABAergic modulation of hippocampal population activity: sequence learning, place field development, and the phase precession effect". *Journal of neurophysiology* 78:1, pp. 393–408.

Wallenstein, Gene V, Michael E Hasselmo, and Howard Eichenbaum (1998). "The hippocampus as an associator of discontiguous events". *Trends in neurosciences* 21:8, pp. 317–323.

Wang, Dong V and Satoshi Ikemoto (2016). "Coordinated interaction between hippocampal sharp-wave ripples and anterior cingulate unit activity". *Journal of Neuroscience* 36:41, pp. 10663–10672.

Wayne, Greg, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z Leibo, Adam Santoro, et al. (2018). "Unsupervised predictive memory in a goal-directed agent". *arXiv preprint arXiv:1803.10760*.

Weinrich, M, SP Wise, and K-H Mauritz (1984). "A neurophysiological study of the premotor cortex in the rhesus monkey". *Brain* 107:2, pp. 385–414.

White, Lisa M (1996). *Temporal difference learning: eligibility traces and the successor representation for actions.* Citeseer.

Whittington, James CR, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens (2020). "The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation". *Cell* 183:5, pp. 1249–1263.

Whittington, James CR, Joseph Warren, and Timothy EJ Behrens (2021). "Relating transformers to models and neural representations of the hippocampal formation". *arXiv preprint arXiv:2112.04035*.

Wills, Tom J, Colin Lever, Francesca Cacucci, Neil Burgess, and John O'Keefe (2005). "Attractor dynamics in the hippocampal representation of the local environment". *Science* 308:5723, pp. 873–876.

Willshaw, DJ, Peter Dayan, and Richard GM Morris (2015). "Memory, modelling and Marr: a commentary on Marr (1971) "Simple memory: a theory of archicortex"". *Philosophical Transactions of the Royal Society B: Biological Sciences* 370:1666, p. 20140383.

Wilson, Matthew A and Bruce L McNaughton (1993). "Dynamics of the hippocampal ensemble code for space". *Science* 261:5124, pp. 1055–1058.

Witt, Jessica K (2011). "Action's effect on perception". *Current Directions in Psychological Science* 20:3, pp. 201–206.

Witt, Jessica K and Dennis R Proffitt (2005). "See the ball, hit the ball: Apparent ball size is correlated with batting average". *Psychological science* 16:12, pp. 937–938.

Witt, Jessica K, Dennis R Proffitt, and William Epstein (2010). "When and how are spatial perceptions scaled?" *Journal of Experimental Psychology: Human Perception and Performance* 36:5, p. 1153.

Witt, Jessica K and Mila Sugovic (2010). "Performance and ease influence perceived speed". *Perception* 39:10, pp. 1341–1353.

Witter, Menno P, Thanh P Doan, Bente Jacobsen, Eirik S Nilssen, and Shinya Ohara (2017). "Architecture of the entorhinal cortex a review of entorhinal anatomy in rodents with some comparative notes". *Frontiers in systems neuroscience* 11, p. 46.

Witter, Menno P, Floris G Wouterlood, Pieterke A Naber, and Theo Van Haeften (2000). "Anatomical organization of the parahippocampal-hippocampal network". *Annals of the New York Academy of Sciences* 911:1, pp. 1–24.

Wood, Emma R, Paul A Dudchenko, R Jonathan Robitsek, and Howard Eichenbaum (2000). "Hippocampal neurons encode information about different types of memory episodes occurring in the same location". *Neuron* 27:3, pp. 623–633.

Wouterlood, FG, HWM Steinbusch, PGM Luiten, and JGJM Bol (1987). "Projection from the prefrontal cortex to histaminergic cell groups in the posterior hypothalamic region of the rat. Anterograde tracing with

Phaseolus vulgaris leucoagglutinin combined with immunocytochemistry of histidine decarboxylase". *Brain research* 406:1-2, pp. 330–336.

Wyble, Bradley P, Christiane Linster, and Michael E Hasselmo (2000). "Size of CA1-evoked synaptic potentials is related to theta rhythm phase in rat hippocampus". *Journal of Neurophysiology* 83:4, pp. 2138–2144.

Wyss, JM, LW Swanson, and WM Cowan (1979). "A study of subcortical afferents to the hippocampal formation in the rat". *Neuroscience* 4:4, pp. 463–476.

Xu, Wei and Thomas C Südhof (2013). "A neural circuit for memory specificity and generalization". *Science* 339:6125, pp. 1290–1295.

Yau, Suk-yu, Ang Li, Kwok-Fai So, et al. (2015). "Involvement of adult hippocampal neurogenesis in learning and forgetting". *Neural plasticity* 2015.

Young, James J and Matthew L Shapiro (2011). "Dynamic coding of goal-directed paths by orbital prefrontal cortex". *Journal of Neuroscience* 31:16, pp. 5989–6000.

Zheng, Chenguang, Kevin Wood Bieri, Yi-Tse Hsiao, and Laura Lee Colgin (2016). "Spatial sequence coding differs during slow and fast gamma rhythms in the hippocampus". *Neuron* 89:2, pp. 398–408.

Zheng, Yicong, Xiaonan L Liu, Satoru Nishiyama, Charan Ranganath, and Randall C O'Reilly (2022). "Correcting the hebbian mistake: Toward a fully error-driven hippocampus". *PLoS Computational Biology* 18:10, e1010589.

Zielinski, Mark C, Wenbo Tang, and Shantanu P Jadhav (2020). "The role of replay and theta sequences in mediating hippocampal-prefrontal interactions for memory and cognition". *Hippocampus* 30:1, pp. 60–72.

Zlotnik, Gregorio and Aaron Vansintjan (2019). "Memory: An extended definition". *Frontiers in psychology* 10, p. 2523.