Navigating Spatial, Temporal, and Semantic Contexts in Memory

Laura Convertino

Institute of Cognitive Neuroscience

Wellcome Centre for Human Neuroimaging

University College London (UCL)

Submitted to UCL in fulfilment to the requirements for the degree of

Doctor of Philosophy

March 2024

I, Laura Convertino confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Singed

Date: 02/03/2024

Abstract

Context is fundamental to create a coherent understanding of experience, index memories and control generalisation from them. Although context has been defined in a variety of ways, clear understanding of how different contextual cues interact and interfere with each other is still missing. Here, I investigate possible behavioural, computational, and neuronal mechanisms of temporal, semantic and spatial context.

In Chapter 2, I first implemented a modified version of the Deese-Roediger-McDermott (DRM) paradigm for false memory. This allowed me to investigate the interaction and interference of semantic and temporal context in recognition and source memory. Moreover, an auto-associative (Hopfield) network reproduced the behavioural results and helped me investigate the role of semantic context and pattern completion. I then implemented a successor representation temporal context model, to test how the temporal context interacts with semantic context in free recall. Finally, I developed two generative models of false memory within the active inference framework as Bayesian optimal behaviour.

Using magnetoencephalography (MEG), I investigated whether neural populations involved in representing spatial context (grid cells) could be impaired in patients with schizophrenia [Chapter 3]. We were able to find 6-fold modulated theta activity in the entorhinal cortex during virtual navigation in healthy controls, but not in schizophrenic patients. This suggests that impairments in knowledge structuring and relational inference associated with schizophrenia may arise from disrupted grid firing patterns.

Finally [Chapter 4], I developed an fMRI task to investigate whether retrieving a list of words corresponds to navigation of a 2-dimensional abstract space, whose axes are organised over the temporal and semantic distance between words, and whether a grid-like code is used for this.

In summary, my work suggests that semantic and episodic memory are deeply interconnected, and that different forms of context - spatial, temporal and semantic - interact and interfere in memory retrieval.

Impact Statement

Modern neuroscience has made incredible progress in understanding the mechanisms and brain structures involved in the building blocks of cognition, exploring the biological bases and computational processes of human thinking. Only very recently, however, this foundational knowledge and new methodological developments allowed us to start the exploration of more complex, integrated, and multifunctional mechanisms to shed light on the real-world use of artificially taxonomically separated cognitive processes. Memory, navigation, and abstract thinking offered the perfect example of this paradigm shift. This work contributes to the field of cognitive and computational neuroscience by proposing alternative integrated computational and experimental approaches to investigate and interpret the role of space, time, and semantics in declarative memory. With rigorous testable hypotheses, my computational work guides future computational and experimental work towards a mechanistic understanding of declarative memory processes. Moreover, this pioneering research bridges the gap between cutting edge discoveries in neuroscience of memory and neuropsychiatry research. This work opens the way for an exciting new field of study in clinical research, where newly discovered neuronal mechanisms of computation can be translationally studied in their potential pathophysiological role in a variety of neuropsychiatric conditions, starting with Schizophrenia. It also validates computational data analysis approaches to study grid cell activity in non-invasive human neuroscience across a variety of methodologies, not only in fMRI but also in MEG studies. Finally, my fMRI study provides preliminary evidence and exciting opportunities for further investigations on the tri-dimensional integration of temporal, semantic and spatial information in the brain, naturally occurring in untrained simple memory tasks. If validated, this final contribution will have a significant impact on supporting the perspective of higher cognitive functions as highly integrated and flexibly utilised by the brain beyond single domains.

Overall, this thesis provides original insights into the most fascinating aspects of human cognition, at the crossroad between memory and abstract thinking. It contributes to latest advances in computational, methodological, and experimental approaches. Moreover, it is one of the very first works to provide scientific evidence and new theoretical hypotheses for a pathophysiological role of an impairment in grid cell population activity in neuropsychiatric conditions, and for the mechanistic biological bases of the multi-dimensional integration of temporal and semantics information via spatial mapping in the human brain.

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis Chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

Reduced grid-like theta modulation in schizophrenia

b) **Please include a link to or doi for the work**

https://doi.org/10.1093/brain/awac416

c) **Where was the work published?**

Brain, Volume 146, Issue 5, May 2023, Pages 2191–2198

d) **Who published the work?** (e.g. OUP)

Oxford University Press.

e) **When was the work published?**

10 November 2022

f) **List the manuscript's authors in the order they appear on the publication**

Laura Convertino, Daniel Bush, Fanfan Zheng, Rick A. Adams, Neil Burgess

g) **Was the work peer reviewed?**

Yes

h) **Have you retained the copyright?**

Yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

Yes: Convertino, L., Bush, D., Zheng, F., Adams, R., & Burgess, N. (2022). Impaired Grid-Like Representations at Theta Frequency in Schizophrenia. BJPsych Open, 8(S1), S48-S48. doi:10.1192/bjo.2022.185

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   a) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   c) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   d) **List the manuscript's authors in the intended authorship order**

   Click or tap here to enter text.

   e) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   Laura Convertino and Daniel Bush contributed equally to this work**.** LC and DB conceived and designed the analysis, performed the analysis, and wrote the paper. DB, FZ and RA collected the data and contributed to planning the experiment. NB supervised the work. LC, DB, RAA and NB developed the theory. DB and NB designed the analysis tools.

4. **In which Chapter(s) of your thesis can this material be found?**

   Chapter 3

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

   Candidate

   Laura Convertino

   Date:

   02/03/2024

   Supervisor/ Senior Author (where appropriate)

   Neil Burgess

   Date:

   11/03/2024

Acknowledgements

With the passage of time, while chapters of life open and close without courtesy of notice, it is difficult to look back and imagine what wasn't and now is, loudly, your life. The people who made it, the experiences that moulded it, the unexpected circumstances and one too many changes of plan that are now either faded memories or embedded realities. Whether they belong to the former or the latter, without these people my life over the last five years, and hence my scientific adventure, would have had a very different flavour, if any at all. In no way this space is quantitively or qualitatively fit for purpose, and while I will delegate meaningful gratitude to better suited means, I hope that with mentioning them here I will give an idea of how little this work is the result of a lonely independent intellectual, and much more the communal effort of a supporting community. It truly takes a village to complete a Doctorate.

To Menno Witter, Alan Warren, Ole Paulsen and Susanna Mierau, Annarita Patrizi, Clifford Woolf, Inge and Cedric, for having put faith in an unripe medical student, opening the door towards a future of scientific exploration, at times when my faith didn't lie on me. To Mona Garvert and Rick Adams, for the gift of knowledge; for UCL. To Gabriella Vigliocco, Mirco Musolesi and Hugo Spiers for believing in me when it mattered the most. To my EcoBrain fellows, Sebastijan, Marie, Sara and Lara, for making it, together. To Sebastijan, once again, not only for being a lovely housemate, but more than anything for the intellectual honesty and constant clarity of mind over any uncertainty. Most importantly, to Neil Burgess and Karl Friston, for making me a scientist, and for the gift of always making me the least smart person in the room; for my future. To Neil, for actively supporting my explorative ideas over years of growth and iterative mistakes. To the wonderful scientists of the Space and Memory lab. To Ingrid, for sharing her beautiful self, on top of the struggle of completing a PhD during a global pandemic. To Dan and Alexandra, for their patience and knowledge. To Rick, once more, for showing me the way forward by example. To Oliver, a caring force in crumbling times. To Jesse, because without him I would have never found the strength to write this thesis. To Berk,

inspiring example of intellectual skills, perseverance, and kindness, for allowing me to understand beyond my own capability. To Michael Moutoussis, travel companion in unexplored lands. To Umesh Vivekananda and Matthew Walker, to Beate Diehl, Fahmida Chowdhury, and Emma Torzillo, for keeping the clinician in me and its meaning alive in the most caring way, proving that medicine is, first and foremost, science.

To the inspiring humans of Science for Democracy, for their constant presence and powerful light. To my TNB companions, Viky, Noor, Lance, Seb and Thomas, because looking at them growing has given hope for my own strength. To the loving and forgiving presence of Shwetha, Maddi, Yan and Anna K. To Jenny, for being the most inspiring superhuman. To the kindness of Clive and to the gift of time (for the lemon). To the bravery of Dot, the strength of Xiao (and to Kev for Xiao); for constantly bringing me back to myself, when the certainty of failure overtook the joy of trying. To Anjali, for sharing the complete Chaos of her Cosmos with us. To Elena, for having been my Pattina over the last 26 years. To my long-lived world-wide family; where a book wouldn't suffice, I will stop at the archetypical introduction of my heroes: to Oreste the honest, Fede the fabulous, Tinelli the bestia, Jean the balanced, Anna M. the warrior, Filippo the charming, Marta C. the timeless, Martina the warmest, and Elisa the sister; for saving me, constantly, in so many ways. To Silvia S, for having made London home. To Rolando, for Rolando. To Silvia and Rolando, for the best bubble a global pandemic could offer, through rain and sunshine, Oronzo and Patrizio (to Oronzo, just to avoid any risk of divine anger). To princess Valia, for appearing in my life as a spirit guide, and never leaving my heart. To Nadine and Peter, wonderful humans, for having found ourselves in each other, and for teaching me love. To my (also)blood-related family, to ~~Nancy Brutammerda ('ugly shit' for the English readers)~~ Edo, for being a bit of Bruno too, and for making me the proudest sister of others' disappointments; to Nanna, for taking care of her future better than anyone could (I can't wait to see it happen!). To Jared, for our adventure together, for the patience…for loving, always. To Mombo too, dumbest animal and best psychotherapist. To Silvia K., for having domesticated my T.rexes into kittens, for the quarter biscuits under my nose, for the gift of unconditional love.

To Pina, Nonna. Always and forever.

Finally, to those who had been and are no more, but always will stay. To Andrea Belvedere, Belse, a mentor, a friend, an ally and a fearless challenger, a teacher; for the chance of the life I wished for.

To Bruno, Nonno, for seeing me before I could, for showing it to me, for everything.

Despite the patriarchy.

Table of Contents

Chapter 1. General Introduction

This work aims to investigate some of the complex and multifaceted mechanisms involved in declarative memory. I will focus on how spatial, episodic and semantic memory entail overlapping and interacting processes across different experimental and computational studies. The unifying key that guided the development of the work included in this thesis is the concept of context and contextualisation of memory. The definition of context varies wildly across neuroscientific sub-fields of research, depending on the question of interest and on the methodology of choice. In this work, the idea of context refers to the realm within which remembered experiences are embedded, and hence reconstructed, whether this entails temporal, semantic, spatial contexts, or a combination of them to create new multi-domain contextual cues.

In the first part of this thesis (Chapter 2), I introduce the concept of temporal and semantic contexts with the aim to explore the underlining computational mechanisms responsible for the integration of newly acquired knowledge and experiences with the pre-existing contextualised and generalised associations between concepts. I approach the problem using a well-known behavioural paradigm of induced false memory, the DRM paradigm (Deese, 1959; Rodiger and McDermott, 1995), driven by the interference of pre-acquired semantic knowledge in newly formed memories, with a new experimental version of the DRM and three different computational models. This allows me to investigate the issue with mathematical rigour, biologically plausible model structures and a well-established and validated memory paradigm.

First, in Chapter 2.1, I develop a modified version of the DRM paradigm and test the false memory effect as pattern completion. In this implemented DRM paradigm, I structured the experiment to allow assessment of pattern completion of the contents of each list on the basis of its temporal (time distance between words and list grouping) or semantic context (words belonging to the same semantic group); this implementation preserved the original DRM paradigm effect (which was validated across participants), while providing more insight in the underline involved mechanism. I then (Chapter

2.2) investigate whether a Hopfield network model of auto-associative memory encoding and retrieval can provide insight into the neuronal structures likely involved in misremembering. This model mirrors the neuronal structure of the CA3 area of the hippocampus, which supports the process of pattern completion involved episodic memory retrieval (Marr 1971; McClelland, McNaughton, O'Reilly, 1995; Horner & Burgess 2013, 2014). I then approach modelling these memory 'mistakes' with a successor representation version (Gershman 2012) of the temporal context model (Howard & Kahana, 2002) (Chapter 2.3), by integrating elements of pre-learnt semantic memory. With this model I aim to contribute to the successor representation implementation of the temporal context model by introducing new hypothesis and computational elements towards a future unification of multiple aspects of declarative memory within the same computational framework. In Chapter 2 Appendix 2, I finally explore how the same mechanism can be understood and reproduced by hierarchical Bayesian modelling within the active inference framework, where I developed two different model approaches to test how misremembering might result from Bayesian optimal processes. These models' ability to misremember derives from an adaptive mechanism of memory integration between newly learnt temporally contextualised information and previously acquired generalised semantic associative knowledge, in a similar fashion to real-world experiences. In Chapter 2, I finally compared these models' behaviours and mechanisms to bring new insights into the possible cognitive processes involved in semantic-induced false memories.

In the second part of this thesis, I focus on the role of spatial representations in memory, including their direct use in encoding spatial context and spatial location (in Chapter 3), and their indirect use in representing non-spatial semantic context (in Chapter 4). Space is the second fundamental aspect of episodic memory, along with time, and understanding of the underlying neuronal representations and computations involved in spatial navigation is pivotal to further investigate how different contextual domains might be integrated into memory processes.

Space is possibly the most intuitive and most studied example of contextual information in neuroscience. In recent years, this field of study has produced an incredibly rich body of evidence and theoretical work that succeeded in providing groundbreaking understanding in how the brain navigates space, whether this is physical of abstract. Thanks to recent methodological advances, it is now possible to explore human navigation with non-invasive neuroimaging techniques. Traditionally, the activity of specialised cells involved in spatial coding, such as grid cells, has mainly been explored in humans in healthy volunteers using fMRI tasks. With a newly developed approach, in Chapter 3 I aim to use magnetoencephalography (MEG) data in a spatial navigation task in a virtual reality environment to bring new insight into the possible pathophysiological role of impaired grid cell activity in schizophrenia. I first extracted spectral power as a function of movementdirection in MEG data from a spatial navigation task and looked for grid-like activity (hexadirectional modulation of power) in theta frequencies; this was compared with navigation performance for the first time in MEG. When comparing the grid-like signal with navigation performance I found a significant correlation between stability of the grid activity and performance in the navigation task in control group. Moreover, when comparing controls to patients with diagnosis of schizophrenia, I found that the grid-like signal in theta frequency was disrupted in the patient group. This work brings new insight into one of the most likely causative mechanisms of impaired inferential reasoning and thought disturbances in affected individuals.

Finally, in Chapter 4 I integrate and unify the insights from previous chapters to explore whether the brain integrates semantic and temporal dimensions to create an abstract 2D space, and whether the same brain mechanisms that are responsible for computing spatial information are also involved in this process. To test this hypothesis, I created a behaviourally simple task of word list memory, carefully built to implicitly create a 2D structure with dimensions of temporal and semantic similarity. The prediction that retrieval from these lists uses the grid-like representations used for navigation could then be tested by looking for grid-like representations in this temporal-semantic space. Although preliminary, this final chapter opens the door to future analysis and

hypotheses testing towards a more wholistic interpretation of aspects of semantic and episodic memory within an integrated declarative memory system.

Each chapter will provide the reader with an independent introduction, to facilitate the integration of a diverse set of theoretical backgrounds and methodologies, while maintaining a coherent narrative throughout the body of work.

Chapter 2. Effects of temporal and semantic context on false memory

Part of the work presented in Chapter 2.1 and 2.2 has been previously presented as a conference poster at the ASSC (Association for the Scientific Study of Consciousness) 2023, and at the conference ‘New Perspectives in Declarative Memory’ University of East Anglia 2022.

Introduction

In this Chapter, I investigate the influence of semantic context (derived from pre-learnt semantic associations) on time-embedded memory retrieval. To do so, I first implement the DRM paradigm and ran a behavioural experiment to compare the semantic and temporal contextual influences on the DRM effect, and how this might relate to pattern completion mechanisms (Chapter 2.1). I then look at how computational models of memory can explain the DRM paradigm, including your new experimental results.

In the following sub-Chapters (2.2, 2.3, 2.4), I will describe and discuss several computational models that can explain the occurrence of false memories driven by semantic interference. Computational modelling has proven to be a valuable tool in neuroscience (e.g., Wand et al., 2020) to better understand human behaviour (e.g., Wilson & Collins, 2019), neuronal processes (e.g., Shine et al. 2021) and brain connectivity (e.g., Friston & Dolan, 2010). The use of computational models in neuroscience support description, explanation, and prediction of biologically plausible mechanisms behind observed evidence; it allows us to fit experimental data, and to simulate and test hypotheses before collecting new experimental evidence.

An intuitive way to think about how different models contribute to understanding was offered by Marr (1982), who promoted the idea that complex systems such as the brain should be understood at different *levels* of analysis. He proposed to call these the computational, algorithmic and implementation levels. The top or computational level describes the problem that the brain might be solving. For example, the problem of maximizing reward in a task. The algorithmic level describes *how* this problem is solved. For example, reward maximization might be achieved using a specific reinforcement learning strategy. Finally, the implementation level describes the physical mechanisms that achieves this, for example neurons and synapses in the brain.

In what follows, I develop three separate models of sequential memory and false memory as captured by the DRM paradigm (Deese, 1959; Roediger and McDermott, 1995). Each model can be placed at a different level of Marr's framework. First, I approached the problem with an auto associative Hopfield network model (Hopfield 1982). This model, which would represent the implementation level, captures the biologically plausible behaviour of the neurons in the CA3 area of the hippocampus. This area is responsible for the mechanism of pattern completion (Horner et al., 2015, Marr 1971, McClelland et al. 1995), which plays a fundamental role in the holistic retrieval of multiple elements that constitute an episode in declarative memory. A core strength of this model is its relatively simple structure, which facilitates understanding of the underlying computations and of the effect of different parameters on the model's behaviour. Moreover, the model has a direct biologically plausible parallelism with a well-defined biological structure, i.e. the CA3 area of the hippocampus. This model is a valuable tool to develop directly testable hypothesis and predictions on the hippocampal role in memory via pattern completion. However, simplicity and well-defined biological mechanisms come at the price of lack of flexibility and understanding of the broader aim of the observed behaviour. These aspects can be better grasped by a higher level of analysis.

Encompassing elements of both the algorithmic and computational level, I approached the process of misremembering in free recall and recognition tasks within the successor representation (SR) and the active inference frameworks, respectively. Both models express algorithmic implementations of biologically plausible mechanisms, as well as higher order dynamics that guide the behavioural goal at the computational level.

I developed a successor representation (SR) model (Dayan, 1993, Gershman, 2018) of sequential memory and free recall (Gershman, 2012, Zhou et al., 2023). This approach reframes the temporal context model (TCM, Howard & Kahana, 2002), which hypotheses a distributed temporal context as a fundamental component of episodic recall, within the computational framework of SR models. The SR framework is rooted in reinforcement learning, in which the aim is to maximize reward. That reward-maximization approach offers a computational level rationale, while using the SR points to a specific algorithmic approach (Stachenfeld et al. 2017, Gardner et al. 2018, Geerts et al., 2020, 2023). There has also been a lot of recent work designing implementation level models of SR computation in the brain (e.g. Fang et al. 2023; Georg et al. 2023), but this will not be further discussed in this thesis. In conclusion, the SR model presented here fits mostly in the computational and algorithmic level of analysis, allowing for interesting predictions at the behavioural level.

In the Appendix of Chapter 2, I approached the DRM task of false recognition within the active inference framework, developing two different hierarchical models of semi-Markovian decision processes (Smith et al. 2021). Within this realm, the higher order goal of the model – of any living organism - is to minimise expected free energy (or in other words, to minimise surprise or uncertainty) and to maximise the model's evidence. The model behaviour is then explained as being Bayesian optimal to minimise expected free energy and improve evidence for the implicit model of the world (Friston, FitzGerald, et al. 2017; Friston, Rosch, et al. 2017). At this level of analysis, we can conceptualise why an optimal behaviour could still result in false memory

based on inference and learning processes, while constraining this cognitive process to broader computationally efficient physics principles of any organised living thing.

In conclusion, the different models discussed in this thesis each shed a different light on the false memory effect, allowing for predictions at the neural and behavioural levels. Future work should build towards an integrative approach where false memory formation is understood from the computational to implementation level.

2.1 A pattern completion account of contextual and semantic influences on source (false) memory: A modified DRM task.

Introduction

Misremembering (i.e. the recollection of an event that never occurred) is a common phenomenon with several real-world implications. In particular, the study of false episodic memories can help to elucidate the constructive nature of normal memory function (Schacter et al., 2011).

Traditionally, episodic and semantic memory have been described as well differentiated phenomena (Tulving and Gazzaniga 1995, Tulving 1972, 1983, 2002). Episodic memory refers to the autobiographical context-dependent re-experiencing of a multimodal event (where temporal-spatial relationships between chronologic events are preserved), while semantic memory provides the knowledge of concepts that are independent from subjective experience (and from context) and is often associated with the use of language.

More recently, an increasing body of literature has provided conceptual and experimental evidence challenging this view and promoting the idea of interdependency between semantic and episodic memory (Greenberg and Verfaellie, 2010)

In neuroscience, the process of memory retrieval has been studied as the result of a neural mechanism of pattern completion (e.g., Marr 1971, Hopfield 1982, Horner et al., 2015), where the retrieval of different elements of the same memory reinstates each other via multiple learned associations. In this work, I investigate how episodic or context-dependent memory reflects both semantics and temporal contributions to source memory. To do so, I validate

a modified version of the DRM paradigm (Deese, 1959; Rodiger and McDermott, 1995). The DRM paradigm consists of presenting lists of semantically related words during encoding, inducing the false memory of a not-presented word that is semantically related to the others (a 'lure word' or 'false memory'). The off-list word (target word, or lure) shows a high degree of both recollection and recognition, with no explicit way to distinguish between false and true memories (there being a high confidence rate for both true and false memories). Although the DRM paradigm has been validated repeatedly, the neural mechanisms that generate false memories are not fully understood (Gallo, 2010; Boggio et al., 2009; Warren et al., 2014).

The DRM paradigm has been used to study the construction of false episodic memory, where the remembered episode is the presentation of the list during the encoding phase. However, the most likely cause of misremembering is the semantic relationship between words in each list, which relates to semantic memory (pre-acquired associations). This enables us to investigate the overlapping boundaries of semantic memory and episodic memory. These two forms of memory may be more flexibly interconnected than previously thought (Tulving 1972, Greenberg and Verfaellie 2010). By implementing the DRM paradigm in a new version that includes context or source memory, I aim to bring more understanding to episodic and semantic memory mechanisms.

In this experiment, I aim to further test whether the misremembering effect can be the result of a mechanism of pattern completion driven by the hippocampus, where semantically related words may be incidentally retrieved. This incidental retrieval of semantically related items subsequently leads to the inception of false memories for those items during retrieval – i.e. recently ‘active’ items are believed to have been part of the original word list, even though some of those items were activated by memory retrieval (i.e. pattern completion) rather than sensory input. Hence, I hypothesise that the mechanisms of memory inference and of false memory are both supported by hippocampal pattern completion processes. The CA3 area of the hippocampus is described as an attractor (recurrent) network, where different patterns of neuronal activity are learnt in a Hebbian-like manner (Marr 1971; McClelland, McNaughton, O’Reilly, 1995). The strengthening of neuronal associations allows pattern completion of the

learnt patterns of activity. If we think about the associative aspect of semantic structure as a pre-learned pattern of recurrent connectivity, then we can imagine how this would interfere with the encoding and retrieval of new episodic associations, to allow pattern completion to activate semantically similar items as well as those in the original list.

In the first part of the study (Chapter 2.1), I aimed to validate the results of the classic DRM paradigm with shorter lists of words. The DRM literature has failed to show any explicit way to recognise the occurrence of false memories (e.g. the confidence ratings of the participants for false memory is the same as for true recognition). However, there is evidence that the recollection of inferred associations in episodic memory tasks is related to an increased reaction time (RT) (Coane et al. 2007). I tested the RTs for both lure recognition (in the first part of the test) and in paired-associate recollection (second part of the test). Previous work showed that the increase of RTs for false vs true memory in the DRM task, occurs without an explicit change in metacognitive measures (Agosta and Srtori, 2013; Marini et al. 2012). In the second part of the testing phase, I aimed to investigate the effects of semantic similarities (measured as cosine distances between word vectors in the word embedding space Word2Vec), membership of semantic 'groups', lists (whether two words were part of the same list or not) and temporal distance (distance in time at encoding between words) on the source recollection of pairs of words, where the list in which they were presented at encoding is the intended source.

A parallel line of work on episodic memory helps us understand the DRM as an inferential task, where the structure for the inferential process is provided by the pre-acquired semantic associative memory. In conventional memory inference tasks, participants are presented with overlapping pairs of items (i.e. A-B, B-C Zeithamova et al., 2012). Later, they are tested on their memory both for those direct associations, and for the inferred association that has not been directly presented (i.e. A-C). The hippocampus is strongly implicated both in associative memory function and inference, which could arise via a process of hippocampal pattern completion (Horner et al., 2015). Importantly, the process of hippocampal pattern completion has been shown to drive cortical reinstatement of non-target items after presentation of a partial cue (i.e.

reinstatement of B when presenting A and retrieving C). Moreover, it has been demonstrated that functional connectivity between hippocampus and medial prefrontal cortex (mPFC) supports memory inference – possibly by aiding retrieval of the first association (i.e. A-B) during presentation of the second association (i.e. B-C; Backus et al., 2016). The pattern completion mechanism can be tested in behavioural experiments with dependency measures (Horner & Burgess 2013, 2014), which show whether the memory of an association in an event is related to the memory for the other associations within the same event (i.e. whether remembering the association A-B relates to the memory of the associations B-C and C-A). In this modified version of the DRM paradigm, I was able to use the same dependency measures to investigate whether the recollection of a pair of words in the same list and of the potential lure word associated to them relates to the recollection of the other pairs in the same list.

Methods

Participants

Thirty-two English native-speakers with age range 18-35 years old were recruited from the general population. Participants were recruited using the Prolific platform (Palan & Schitter, 2018; https://app.prolific.co/), while the experiment was run online using the experiment building platform Gorilla (Anwyl-Irvine et al., 2020; https://gorilla.sc). The same participants' results were included in all the performed analyses. The study was approved by the UCL Research Ethics Committee and all participants gave informed written consent before taking part.

Materials

I used the DRM pool of 40 lists of 15 word each with coherent semantics, adapted from Sadler et al. (1990). The words within each of the 40 groups were semantically related to each other, and all related to one additional lure word for each list. From the original forty 15-word lists, I created sixteen 4-word lists of coherent (eight lists) or mixed (eight lists) semantics (Fig.2.1.1).

The eight lists with words coming from the same semantic group (related), were formed from four semantic categories, so that from each group I derived two lists, with different words but paired semantics. The eight lists with mixed semantics were created using words from different semantic groups, picked randomly from the remaining thirty-six of the original 40.

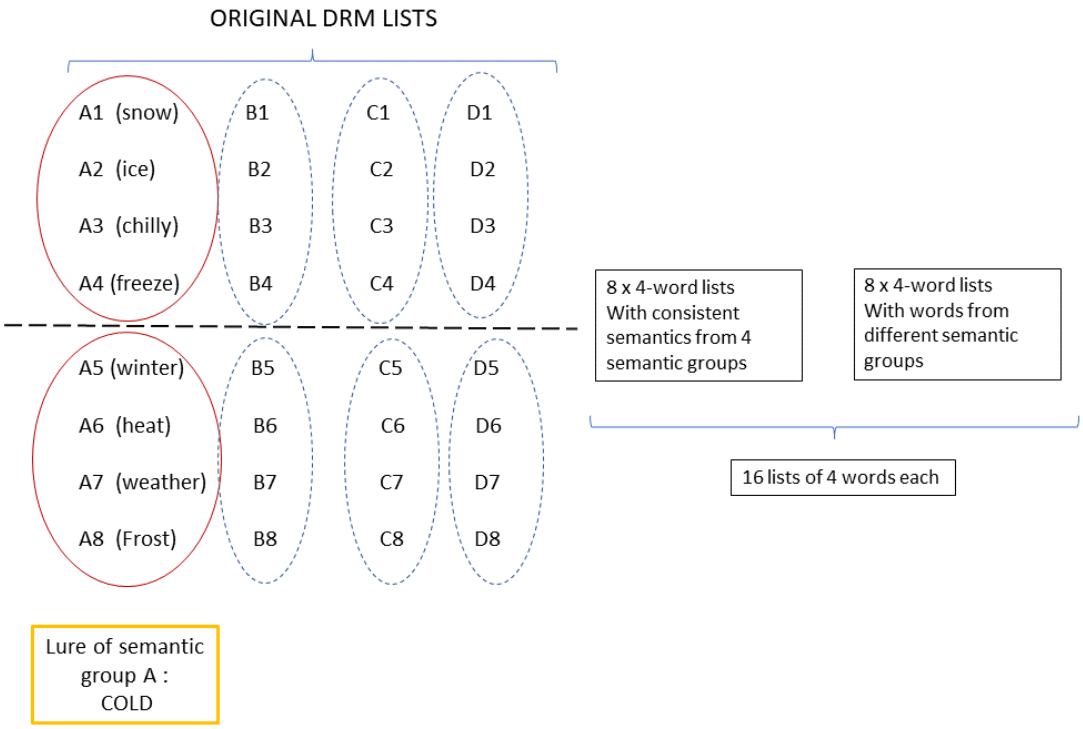


Figure 2.1.1. Creating the 4-word lists. Each column represents one list of the original DRM task (semantic group), identified by a capital letter that relates to a specific semantic category. The number that follows the letter indicates a different word within the list. I picked eight words from four of the original lists (of 15 words each), which resulted in eight lists of four words each, with two lists from each semantic group. Each of the four semantic groups is associated to a lure word, which provides the false memory effect in the DRM task. From the remaining semantic groups, I created an additional eight lists with mixed semantics, i.e. with words taken from different semantic groups.

Procedure

During encoding (Fig. 2.1.2), lists of words were presented in a randomised order. Within each list, the words were also randomised in their presentation order. The participants were informed that they were going to be presented with lists of 4 words each, and they should try to remember them. After having read the 16 lists once, the participants performed a recognition task and a recollection task. The time gap between words within the list was 250ms, while the time gap between lists was 1500ms. Only one word at a time was presented on the screen (see Fig. 2.1.2) for 1500ms. The participants were asked to read each word out loud. The trial ended when all the sixteen lists were presented.

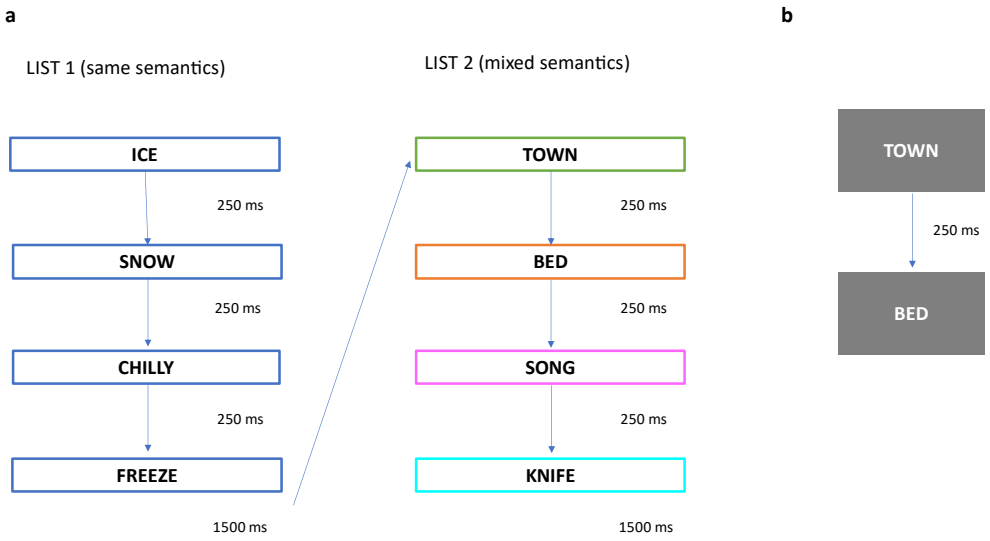


Figure 2.1.2. Design for Encoding. a) Example of randomised presentation of two different lists at encoding. In the figure the colour of the outline of each box indicates the semantics of the word. **b)** Example of what the participants saw during encoding. During the experiment the words were all presented in white on a dark grey screen; there was not any difference in colour to suggest a difference in semantics.

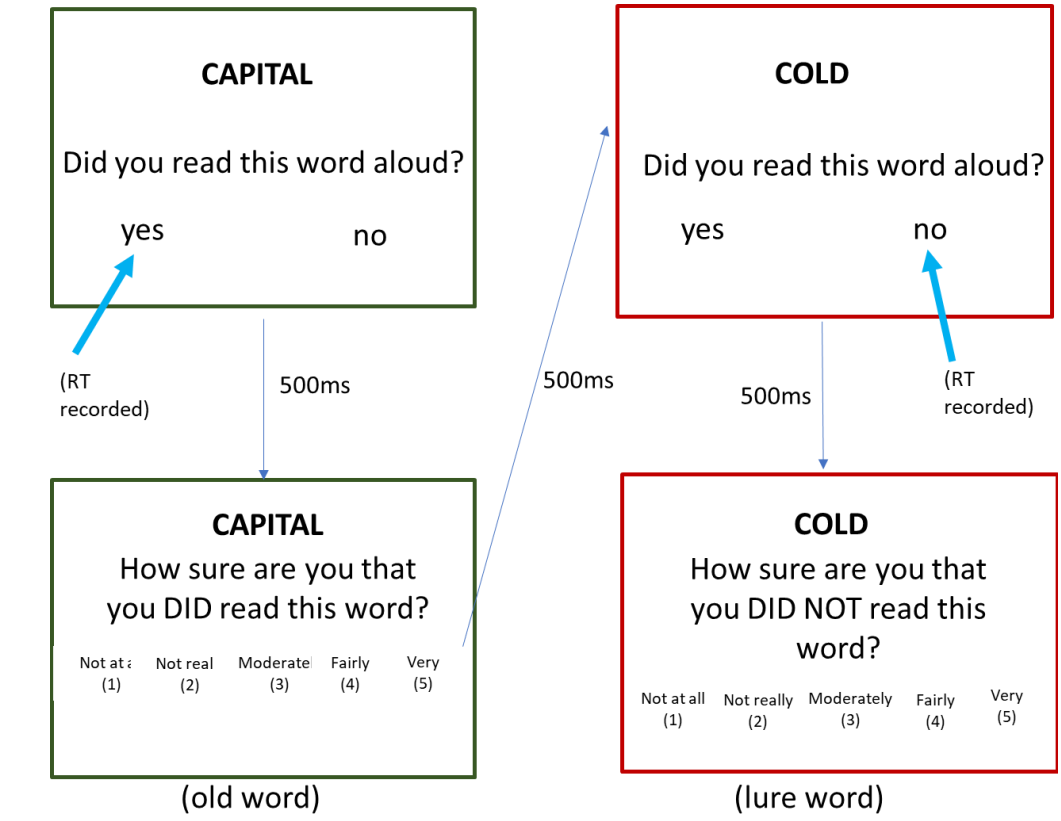
The testing phase was divided in two different sections.

In the first part of testing, I tested the recognition for words that were either presented in the lists at encoding or not. In the recognition task (Fig. 2.1.3 a),

the participants were presented with an equal number of old words, presented at encoding (64) and new words (64), and with 4 lure words (specific for the 4 selected semantic groups). For each word, the participants were asked to decide whether they had read the words in any of the presented lists (‘old’) or not (‘new’). The time gap between each word, as the one between questions, was 500ms. The order of presentation of all the words was randomised. After each choice, no feedback was provided, and the participants were asked to rate on a scale from 1 (not sure at all) to 5 (fully sure) the confidence of their answer. Then, the task moved to the following word automatically. I recorded answers, reaction times and confidence ratings for each answer of each participant.

The second part of the testing was a recollection task. The participants were presented with two words on the screen at a time and asked whether the 2 words were part of the same list at encoding (Fig. 2.1.3 b). The pairs of words were matched as summarised in Figure 2.1.4: three pairs of words in the same list for each of the sixteen lists (48 pairs); two pairs with one word from the corresponding list and the lure word for each semantically consistent list (two pairs from each of the eight lists with coherent semantics. 16 pairs in total); six pairs from eight different couples of lists (48 pairs), made of words from two different lists (each list was paired with another list, so that the list with the same semantic group were paired with each other, and the list with mixed semantics were associated with another list with mixed semantics). After having prepared the pairs of words (112), I randomised the order of presentation. The participants were asked to remember whether the two presented words were part of the same list of four words at the encoding, and they had to answer by using 2 different keys on the keyboard. Each pair was presented on the screen until the participant answered the question (either remembering the two words in the same list or not). After each answer, the participants were asked to rate their confidence between 1 to 5, from the lowest to the highest certainty (Fig. 2.1.3). The time gap between each question was 500ms. Confidence ratings and reaction times for each question were recorded.

a


CAPITAL
Did you read this word aloud?
yes
no
(RT recorded)
500ms
CAPITAL
How sure are you that you DID read this word?
Not at a (1)
Not real (2)
Moderate (3)
Fairly (4)
Very (5)
(old word)
500ms
COLD
Did you read this word aloud?
yes
no
500ms
(RT recorded)
COLD
How sure are you that you DID NOT read this word?
Not at all (1)
Not really (2)
Moderately (3)
Fairly (4)
Very (5)
(lure word)

b

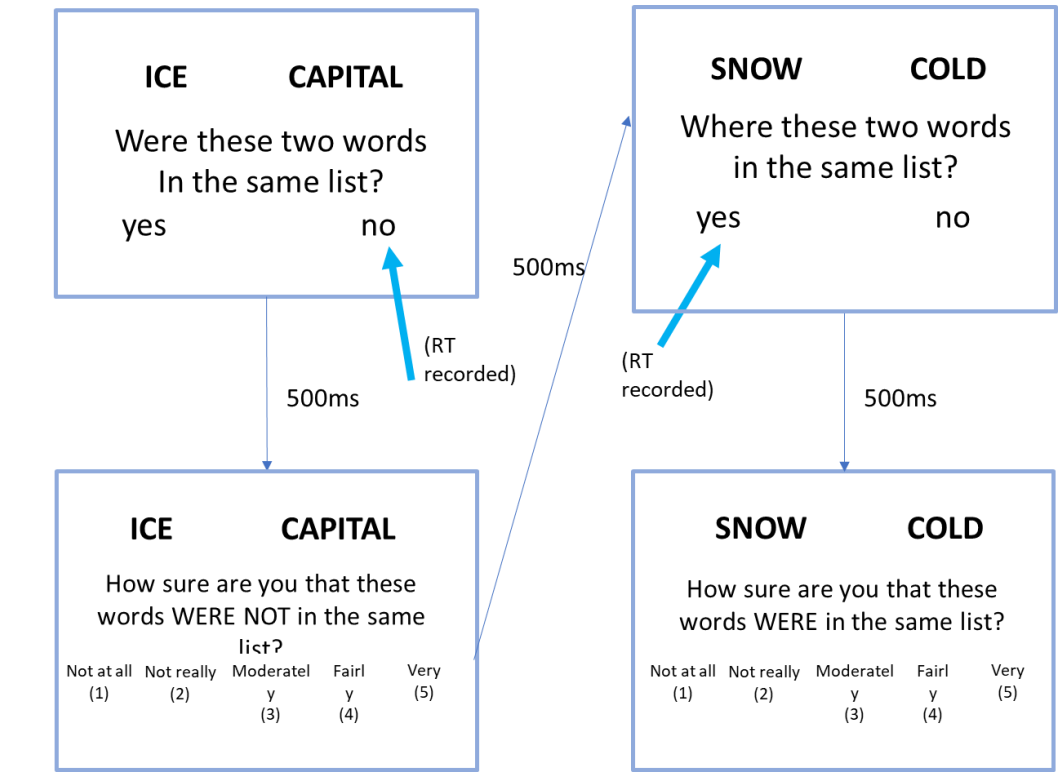

ICE CAPITAL
Were these two words In the same list?
yes
no
(RT recorded)
500ms
ICE CAPITAL
How sure are you that these words WERE NOT in the same list?
Not at all (1)
Not really (2)
Moderatel y (3)
Fairl y (4)
Very (5)
500ms
SNOW COLD
Where these two words in the same list?
yes
no
(RT recorded)
500ms
SNOW COLD
How sure are you that these words WERE in the same list?
Not at all (1)
Not really (2)
Moderatel y (3)
Fairl y (4)
Very (5)

Figure 2.1.3. Design of testing phase. a. Recognition task. Each tested word was presented individually on the screen, and the participants were asked to recognise whether a word was previously presented in a list or not. **b. Recollection task.** Two words were presented on the screen at the same time, and the participants were asked to remember whether the two words were part of the same list. For both tests, both choice and reaction times (RTs) were recorded. Eventually, the participants were asked to rate the confidence of their choice from 1 to 5. The blue arrow indicates the choices in the example.

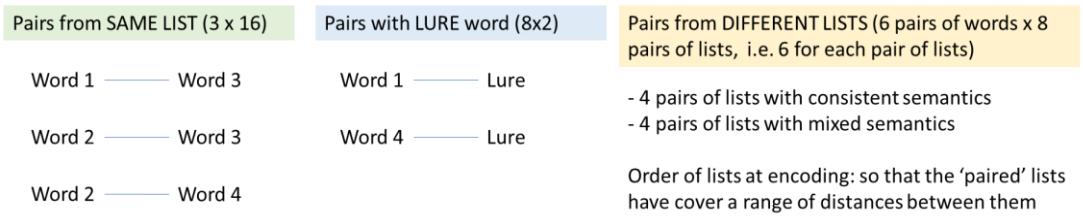


Figure 2.1.4. Structure of the pairs of words. Summary of how pairs of words were created from the presented lists. All words in the pairs were presented at encoding, but the four lure words. The number next to each word indicates the position of each word in a list of four items. Only eight of the sixteen lists had an associated lure, with the same lure word being associated with two lists from the same semantic group.

Results

Recognition

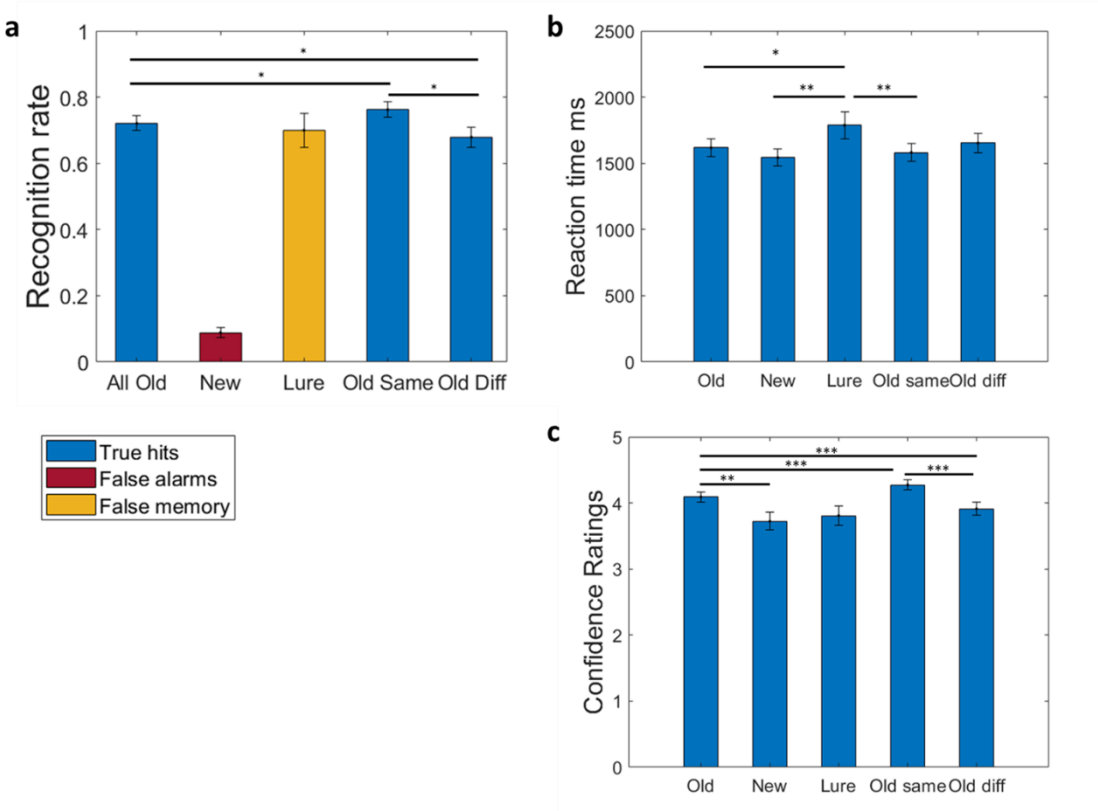


Figure 2.1.5. Behavioural results for DRM recognition task. a. Performance for recognition of old, new and lure words. Bar 'Old Same' and 'Old Diff' show the recognition rates for old words (first bar) divided into old words from lists with coherent semantics ('Same') and mixed semantics ('Diff'). The recognition rate for lure word and old words is not significantly different. However, there is a significant difference in recognition performance for old word from lists with coherent semantics and mixed semantics. **b.** Reaction times in milliseconds for recognition task. The average reaction time for lure words is significantly longer than the ones for all old words, new words and for old words from lists with coherent semantics. **c.** Confidence ratings for difference response types. Participants were more confident for old vs new words answers. However, answers for lure words (false memory) didn't show any significant difference in confidence ratings from either or new word. The error bars represent the standard error. Legend: * = *p-value* < *0.05;* ** = *p-value* < *0.01;* *** = *p-value* < *0.001.*

Thirty participants out of thirty-two were included in the analyses. Two participants were excluded since their performance in the recognition task for true hits and false alarms (excluding the lure word) were below mean – 2.5 standard deviations. For the recognition task, I computed the performance for the recognition of true memory (old words: *M* = 0.72 , sd=0.12), false memory (lure words: *M* = 0.7, *sd*= 0.28) and new words (*M*= 0.09, *sd*= 0.08). For old words, I analysed the performance for all old words, and for old words from lists with coherent semantics and from lists with mixed semantics separately (Fig. 2.1.5 a).

These high false recognition rates are consistent with the literature: the DRM paradigm resulted in inducing the misremembering effect in 70% of cases, i.e. the lure words were recognised as 'old'. These result replicates the original version of the DRM paradigm, where longer lists of words were used at encoding. Recognition rate for lure words and old words was not statistically different (*t(29)*= -0.41, *p*= 0.69, *sd*= 0.28). However, the standard deviation for lure word recognition was significantly higher than for old words (*f(29)* = 0.1868, *p*<0.001). Recognition for old words from lists with coherent semantics was also significantly higher than recognition rate for words from lists with mixed semantics (*t(29)*= 2.7, *p*= 0.0115 *sd*= 0.17).

The reaction times (RTs) (see Fig. 2.1.5 b) were longer for lure words – both recognised as 'old' or reported as 'new', compared to answers for all old words (*t(29)*= *-2.37*, *p*= 0.02, *sd*= 393.82) and new words (*t(29)*= -2.94, *p*= 0.0064, *sd*= 455.12). When comparing RTs for lure words vs old words from different list types, the RTs for lure words were still higher than the ones for words from list with coherent semantics (*t(29)*= 2.83, *p*= 0.0083, *sd*= 397.31), but not compared to the ones from mixed semantics (*t(29)*= 1.74, *p*= 0.09, *sd*= 424.56). No other difference between answer types was significant.

The confidence ratings (reported on a scale from 1 to 5) show differences between old and new word answers, and between old words from lists with coherent and mixed semantics (with more confident answers for words from coherent semantics). However, the confidence for lure word answers was not significantly different from any of the other answer types. This confirms the

lack of an explicit metacognitive way to differentiate between false and true memories.

We would expect longer reaction times to be associated with lower confidence ratings for that response; in other words, the more confident we are in our response, the faster the answer. Here, while the RT associated with (false) recognition of the lure word are significantly longer than the RTs for other responses, this difference was not captured by confidence rating results. This suggests that the participants do not have conscious access to the false memory effect, while the recognition of lure words might activate longer neuronal pathways.

Recollection

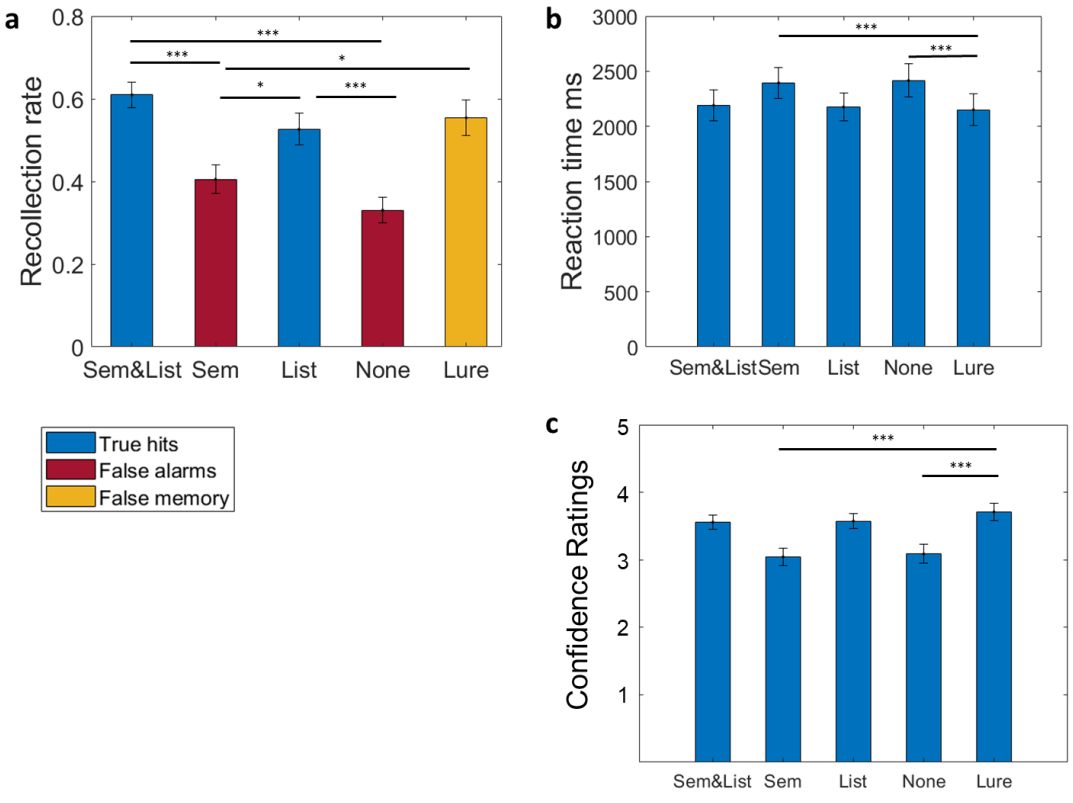


Figure 2.1.6. Behavioural results for word pair recollection as a function of their relationship. 'Sem&List' indicates pairs of words from the same list at encoding and coherent semantics; 'Sem' represents words from different lists but with similar

semantics; ‘List’ indicates words from the same list with mixed semantics; ‘None’ are pair of words from different lists with mixed semantics; ‘Lure’ are pairs of words where one word was the lure word and one word was from a semantically related list. **a.** Recollection for pairs of words. Rate of recollection for pairs of words coming from the same list. ‘Sem&List’: *M* = 0.61, *sd* = 0.17; ‘Sem’: *M*= 0.41, *sd* = 0.19; ‘List’: *M*= 0.53*, sd* = 0.21; ‘None’: *M* = 0.33, *sd* = 0.17; ‘Lure’*: M*= 0.55 *sd* = 0.24. **b.** Reaction times for all answers in recollection tasks, in milliseconds. **c.** Confidence ratings for all answers. Legend: * *p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001.*

To analyse the source memory performance, I averaged the rate of recollection for each kind of pair (5 categories) across all the participants (Fig. 2.1.6 a). There was not a significant difference in recollection rates between pairs with the lure word and pairs of words from the same list, both with coherent semantics (*t(29)*= 1.30, *p*= 0.20, *sd*= 0.23) and mixed semantics (*t(29)*= -1.13, *p*= 0.27, *sd*= 0.14). However, recollection for lure pairs was significantly different from pair of words from different lists, either with coherent semantics (*t(29)*= -2.36, *p*= 0.03, *sd*= 0.35) or different semantics (*t(29)*= -4.51, *p<0.001*, *sd*= 0.27). There was not significant difference between RTs (Fig. 2.1.6 b) of pairs from the same lists and pairs with the lure word. However, there was a significant difference between ‘Sem&List’, ‘List’ and ‘Lure’ pairs’ RTs and the RTs for pairs of words from different lists. In particular, the RTs of pairs with the lure words were higher than the RTs for pairs of words from different lists with coherent semantics (*t(29)*= -4.10, *p<0.001*, *sd*= 327.20) and pairs of words from different lists with different semantics (*t(29)*= -3.90, *p<0.001*, *sd*= 374.54). In the recollection test, the lure word was as a strongly represented in the participants memory as words in the semantically related lists and no difference could be found in either RTs or confidence ratings. The confidence ratings, similarly to the RTs, did not show any significant difference between pairs in the same lists and pairs with the lure word; however, these were significantly higher than confidence ratings for words from different lists. There was no significant difference between confidence (Fig. 2.1.6 c) in responses between pairs with the lure word and pair of words from the same list, but there was between pairs with lure words and pairs of words from different lists, both with coherent semantics (*t(29)*= 5.1, *p<0.001*, *sd*= 0.71) and with mixed semantics (*t(29)*= 4.51, *p<0.001 sd*= 0.76).

To investigate the effect of temporal and semantic distance between words in a pair at encoding (i.e. how many words were in between the two words in each pair during encoding) on recollection, I ran a generalised estimating equation (Liang & Zeger, 1986). The measure for semantic distance between words in each pair was computed as the cosine distance in the word2vec embedding model, while the temporal distance between words at encoding was approximated as the number of words presented between them.

There was a significant effect of temporal distance on pairs recalled (*X2 (df =1)* = 26.55*, p < 0.001*, b=-0.019, 95%CI [-0.26 -0.12]), with a reduced probability of recalling words as coming from the same list that were further apart in time at encoding.

The semantic distance measured as cosine similarity showed a significant effect on pair recollection (*X2 (df =1)* = 5.89*, p =0.015*, b=1.34, 95%CI [-0.26 2.42]).

Moreover, I tested for effect of list (same vs different list at encoding) and semantics (same vs different semantic categories between words in the same pair) and found a significant main effect of both list (*X2 (df =1)* = 107.69*, p < 0.001*, b=0.8, 95%CI [0.65 0.95]) and semantics (*X2 (df =1)* = 20.24*, p < 0.001*, b=0.35, 95%CI [0.2 0.5]).

Dependency

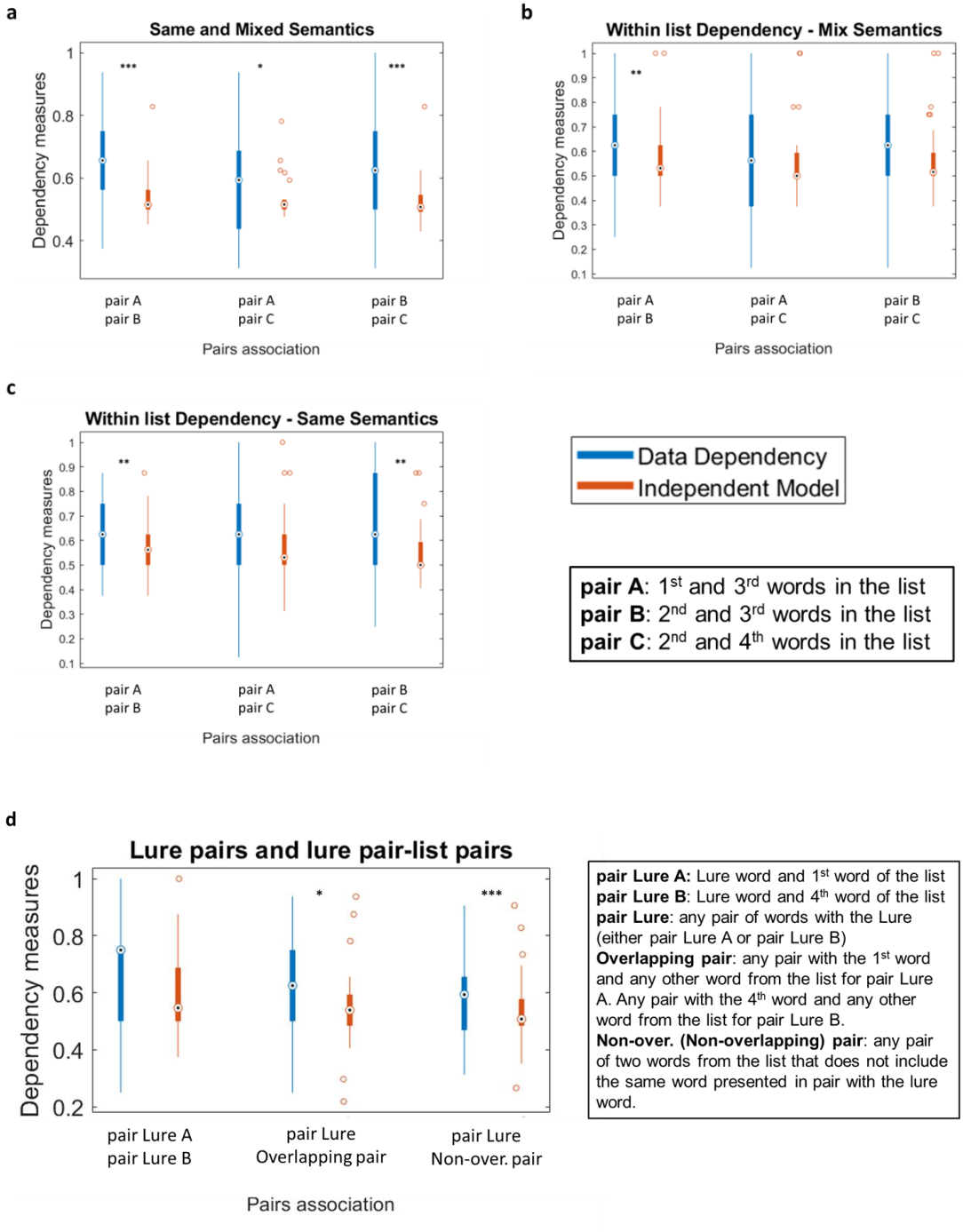


Figure 2.1.7. Dependency measures. Different words are represented by different numbers. The dependency measures were ran to test whether recollection of

different pairs of words were mutually dependent or independent from each other **a**. Comparison between independent model and dependency in pair recollection of pairs from any list. **b**. Comparison between independent model and dependency in pair recollection of pairs from lists with mixed semantics. **c**. Comparison between independent model and dependency in pair recollection of pairs from lists with consistent semantics. **d.** Comparison between independent model and dependency in pair recollection of pairs with lure words. The lure word was either presented in association with the 1st (pair Lure A) or with the 4th (pair Lure B) word in the associated list. 'Overlapping' pairs represent pairs of words from the list containing the same word presented with the lure pair (1st or 4th word), associated with any other word from the same list. Non-overlapping pairs are pairs of words from the semantically associated list to the lure word, but that do not include the word presented with the lure during testing. . Legend: * = *p-value < 0.05;* ** = *p-value < 0.01;* *** = *p-value < 0.001.*

Finally, I investigate whether the performances in retrieval of pairs of words from the same list (and lure word) depend upon each other; in other words, I test if the recollection of one pair in one list depends on the recollection of other pairs in the same list, including pairs with the lure word (false memory) where all the words in a list are semantically related. A similar approach has been used in previous research on associative memory (Horner and Burgess, 2013; Horner and Burgess, 2014; Horner et al. 2015), where dependency measures were developed to investigate whether the memory for one element of an episode (as overlapping paired associations of object, location and subject), cued by another element from the same episode, depended on the cued memory performance of other elements in the same episodes. This analysis is usually built with a contingency table, where the retrieval of each element of a 3-elements event is cued by the other two elements, and where an element retrieves the other two (see table 1, from Horner et al. 2015). In this case, I did not retrieve one word by cuing it with another word from the same list, but I presented a pair of words on the screen at the same time. What I looked for was not the dependency of the memory of one word on the memory of the others, but the retrieval of one pair of words as dependent on the memory of the other pairs from the same list. Suppose that list *A*, with all the

words from the same semantics category, is made of words *A1, A2, A3, A4* at encoding, and it is associated with lure *LA*. The pairs that are presented are: *A1-A3, A2-A3, A2-A4, A1- LA, LA-A4*.

I use dependency measures to investigate whether the recollection of a pair of words within a list was dependent on the rate of recollection of other word pairs in the same list and I compared them to measures of dependency in an independent model of the data, using an adapted version of the measures from Horner et al. (2015). The dependency measure (refer to Horner and Burgess, 2013 for details) represents the rate of trials in which two pairs of words were both successfully retrieved, or both not retrieved. The scale (0-1 interval) gives value 1 for full dependence and 0.5 for full independence (below 0.5 represents an inverse effect between the recollection of the two pairs). Across lists (with mixed and same semantics), the recollection of a pair in the list was dependent on the recollection of other pairs in the list (Fig. 2.1.7 a), with the rate of recollection between overlapping pairs of words more strongly mutually dependent ($t(29)$= 4.37, $p<0.001$, sd= 0.1187). Analysing the dependency measures in pair recollections for lists with mixed semantics (Fig. 2.1.7 b) and consistent semantics (Fig. 2.1.7 c) separately, the recollection of pairs with overlapping words was not independent, while the recollection of non-overlapping pairs was not significantly different from the independent model.

The recollection of pairs with lure words was not independent from the recollection of other pairs with overlapping non-lure words ($t(29)$= 3.43, $p<0.01$, sd= 0.11) and non-overlapping non-lure words ($t(29)$= 3.87, $p<0.001$, sd= 0.07), while was independent from the recollection of other pairs with the same lure word (Fig. 2.1.7 d). This means that if a participant recalled more pairs of words from a list with coherent semantics, they were more likely to recall the lure word in association with another word from the same list. However, recalling a lure word paired with a presented word from the coherent list did not depend on recalling the lure word paired with another word from the same list. This suggests that the event (i.e. list) pattern completion effect drove the recollection of the lure word paired with another word from the list;

however, the pair with lure per se did not affect the recollection of the lure word paired with another word in the list.

Discussion

I investigated the role of episodic and semantic memory in a modified version of the DRM paradigm for false episodic memory. These findings open a broader discussion on the role and mechanism of ‘context’ - where the contributions of both semantic structure and temporal context play an important role at retrieval - but keep open the debate on the common or separate nature of these two forms of context.

In the first part of the experiment, the performance results show how the DRM experiment can be validated also by using lists of only 4 words each. The lure word was recognised as seen before in a high proportion of cases (70%), with no statistically significant different from recognition of old words (true hits).To better investigate whether we could find a behavioural mark to disentangle between true hits and false memory, we then focused our attention on reaction times of the responses. The comparison across average reaction times (RTs) showed a significant difference between presented words recognised as old and false memory. This result was obtained without the use of an implicit association test (used by Marini et al. 2012), but just automatically collecting the timings of the participants’ answers on the keyboard. This underlines the presence of an implicit measure to detect false memory in the DRM task, although the confidence ratings did not differentiate between false and true memory. The increase in RTs for false memory might result from a pattern completion process based on semantic associations. A possible interpretation is that the retrieval of lure words reflects pattern completion via semantic associations alone, whereas the retrieval of words from the list is faster because of the additional associations learned during list presentation (see also Staresina and Wimber, 2019). Similarly, a possible role of pattern completion mechanism for false memory can be related to conceptual

associations (semantic structure) across perceived and not perceived items. Alternatively, longer RTs might be suggestive of engagement of processes related to reality monitoring or source attribution mechanisms (Ranjan et al., 2024; Johnson et al., 2024). Further investigation is necessary to investigate a potential role of these mechanism during encoding and retrieval.

The pair recollection test gave me the possibility to better investigate the role of temporal distance and semantic context in source memory retrieval. The performance analysis showed how the lure words, once recognised, are incorporated in the episodic memory, and become indistinguishable from old words with the same semantic category. RTs confirmed that no implicit or explicit measures can help to identify the lure words in the associative recollection task. The lack of either an implicit or explicit way to differentiate pairs of old words from the same list with similar semantics and pairs of old words with a related lure is most likely due to the interposition of a recognition task between encoding and the second source recollection task, creating a misinformation effect for the lure words (Loftus and Hoffman, 1989). Interpolating the recognition task could indeed provoke the recollection of pairs of words with a lure, then indiscernible from true memory.

It is important to underline that the choice of intercalating the recognition task before the pair recollection task might have influenced our results. In this experiment, the participants were previously exposed to the lure word in the testing phase of the first (recognition) task, and this might have biased them towards recognising the lure words in the pair recollection task in a second moment. While our results support the hypothesis that recognised lure words are embedded in memory for word associations (the RTs difference for lure words were indeed lost in the second task), it also prevents us from having comparative performance for pair recollection immediately post encoding. Future research might be needed to repeat the pair recollection task after list presentation, without the intercalated word recollection task, to investigate how the lure word is incorporated into source memory without being ever presented to the participants.

The manipulation of the DRM task to include testing whether paired associates were from the same list allows us to examine the role of semantics and temporal elements in the retrieval of associated items. The time-related variables, time distance and list, can be interpreted as the continuous flow of experience and its fragmentation into separate episodes. The subdivision of time points in the lists is driven both by the instructions given to the participants at encoding (their goal is to remember lists of 4 words each) and by the progressive change in temporal context, informed by a longer time gap between lists and by the change in semantic context between lists (Howard and Kahana, 2001, 2002). I included, for semantic analysis, both the semantic distance between word vectors (as cosine of the angle between word vectors in Word2Vec) and the semantic category of the different groups of words (discrete measure). If, on one hand, the continuous semantic distance between words can be interpreted as the tendency of words with similar semantics to appear in similar context, on the other the categorical classification provides a more interpretable correspondence with the hypothesis of a semantic category (namely, as a pattern of activity linked to recurrent associations across words in the same category; the category itself can indeed result from the array of overlapping associations).

The dependency analysis adds another element of support for the presence of a semantically related pattern completion mechanism. This, in association with the main effect of both semantics and time (i.e. the words being presented in the same list) for recollection of words coming from the same list with the same semantics, suggests that remembering the lure word as part of the list facilitates the memory for the lure word in association to both semantically related words. The reverse is also true, if the lure word is not remembered. The behavioural results suggest the presence of a semantic-based pattern completion mechanism, which supports the presence of false episodic memory.

Interestingly, these findings relate to previous work in memory research (Howard and Kahana,2002; Naim et al. 2020). In particular, the presence of two aspects involved in memory retrieval – semantic and temporal distance

(or context) – had been studied in experiments of free recall of lists of random words. In free recall, both conditional response probability (the probability of two words being recalled in close proximity) and conditional response latency (the distance in seconds between two recalled words) could be expressed as a function of the semantic similarity between words (measured as LSA cos θ) and as a function of the distance between words at the encoding, with the two components playing a role both at retrieval.

The core question that arises from these results is the role and the neural correlates of different contextual cues. Semantic and temporal context might be integrated in the temporal context theory, with one resulting context that can include both systems of 'coordinates'. Alternatively, the presence of semantic factors in episodic memory could result from two different mechanisms of contextual knowledge, which both facilitate memorisation and result in false memory induced by the DRM task.

In summary, this work provides evidence for a pattern completion mechanism responsible for semantically driven false memory. Indeed, the results of the dependency analysis support the hypothesis of the involvement of a pattern completion mechanism. The hypothesis that the hippocampus plays a role in semantic associations and conceptual context, poses further questions, which could be tested with further neuroimaging and behavioural studies. Further investigations are also needed to better explore the role and meaning of 'episode' and 'context' in memory encoding and retrieval, and whether the role of the hippocampus in memory is not purely episodic (Ekstrom and Ranganath, 2017; Mok and Love, 2019), but more broadly responsible for patterns of associations in different dimensions (time, space, semantics).

2.2 An Auto-associative Hopfield Network Model of Pattern Completion for False Memory.

The results of pair recollection and the dependency analysis in Chapter 2.1 provide new insight on a potential mechanism of pattern completion in false memory for an event. The dependency measures suggest that pattern completion based on semantic associations could be responsible for the presence of false memory in a source memory task. To investigate whether the dependency findings could result from the associative neural network of the hippocampus, I built a Hopfield network model of associative memory (developed from Horner et al., 2015). In the pre-experimental learning phase, the fully recurrent neurons learned associations between word-related neurons coming from the same semantic group. During encoding, the network learned associations based on the presented lists (one word after the other). When recalling pairs of words, two neurons were partially activated (simulating the two presented words at encoding); the recollection of the pair of words together was achieved by the full activation of both neurons over a set threshold, driven by the auto-associative network of learnt associations.

Methods

The simple network was made of rate-coded neurons (Equation 2.2.1). The N neurons formed a fully recurrent network (Horner et al. 2015). I set a time constant $\tau_r = 25ms$ and used a sigmoidal function (Equation 2.2.2) with external current and recurrent synaptic current to obtain the firing rate r_i of the neurons. I used a threshold for input firing rate $r_t = 10$ and a peak firing rate $r_{max} = 10Hz.$ The firing rates were initially set to zero. Each word was associated with a unique neuron, and I included all the words presented during encoding plus one lure word for each semantic group.

$$\tau_r \frac{dr_i}{dt} = -r_i + f(I_{i,ext} + I_{i,syn}) \quad [2.2.1]$$

$$f(x) = \frac{r_{max}}{1+\exp(r_t - x)} \quad [2.2.2]$$

With my simulation, I tested 4 semantic groups and 8 lists of 4 words each for 30 participants. To test all the kinds of pairs tested, I created 4 lists of words from the same semantic group, an additional 2 lists from the first 2 semantic groups but with different words and 2 lists with words from the last 2 semantic groups (mixed semantics). The synaptic connections were set to zero for all synapses before the simulation of the encoding phase. To simulate pre-leant semantic similarities, we provided the model with pre-encoded connections.

The pre-encoding semantic learning was introduced using a probability-based system, where the probability of potentiating a synapsis was changed randomly for each trial from a standard uniform distribution between 0 and 1. The probability of having a pre-encoded synaptic connection between semantically related words was 0.3. If the Hebbian condition was met, there was a random chance of increasing the connection weight to 1.1. The semantic group was recreated by setting a pre-task knowledge for association across words in the same semantic group and between lure word and related words. These pre-encoding associations simulate the presence of semantic memory. To assign these pre-encoded associations, I used different probabilities for learnt associations between words from the same semantic group (0.4) and for lure word and related words (0.3). All learnt associations for pre-task knowledge of associations had strength =1.1. At encoding, I kept the same structure of the behavioural experiment by presenting one word at the time, with 1500msec of presentation for each word and 1500msec gap without words between different lists. I ran the model with two different methods for encoding: a) a learning dynamic based on the Hebbian learning rule for association between presented word and previous one; b) probability-based for each trial, with probability of strengthening a connection that changes across trials.

For the Hebbian encoding method, I used an Hebbian learning function (Equation 2.2.3). I set the time constant for the firing rate decay to 25msec

(time gap between words in the same list), and the learning constant k to 0.005. The external stimulus for each word was set as a constant current of 7 milliAmps for a period of 1500msec. The associations were learnt thanks to the overlaps in activation between the previous and following words (modulated by the decay constant). The same dynamics were run again between different lists for a time gap of 1500msec, without any external stimulus.

$$w_{ij} = w_{ij} + k * r_i * r_j \qquad [2.2.3]$$

The second encoding method used a probability-based system, similarly to the methods used for pre-learnt semantic associations, where the probability of potentiating a synapsis was changed randomly for each trial from a standard uniform distribution between 0 and 1. The probability threshold to learn associations between words from the same list was 0.55. Over time, the probability of learning associations between words was decayed by decay constant= 0.3. When the model moved from one list to the next, the probability of encoding association was further decreased by decay constant, so that the probability of learning associations between words in the same list was higher than the probability of learning associations between the last word of a list and the first word of the following list (taking into account the longer time gap between lists). Once learnt, each association was assigned strength of 1.1.

The retrieval phase was prepared with randomised presentation of different pairs of words: pairs of words from same list with the same semantic group, from the same list with different semantic groups, from different lists with the same semantic group, from different lists with different semantic groups, and pairs with a lure word and a related word from same semantic group. For each pair, I induced a constant current for both word-related neurons of 9 for a period of 1500msec. The time constant for firing rate decay at retrieval was set to 25msec. The dynamics included an additional recurrent synaptic current

(see Equation 2.2.1). The recurrent current is the product of the firing rates of the connected neurons and the associated synaptic weights (Equation 2.2.4).

$$I_{i,syn} = \sum_j w_{ij} r_j \qquad [2.2.4]$$

For each trial, I look at the final firing rates of the two neurons associated with the words in the pair as a measure to identify successful retrieval. I used a threshold of 8Hz; if both neurons had a firing rate higher than the threshold, the answer for the trial was recorded as 'yes, the two words were presented together', otherwise as 'not in the same list at encoding'.

Results

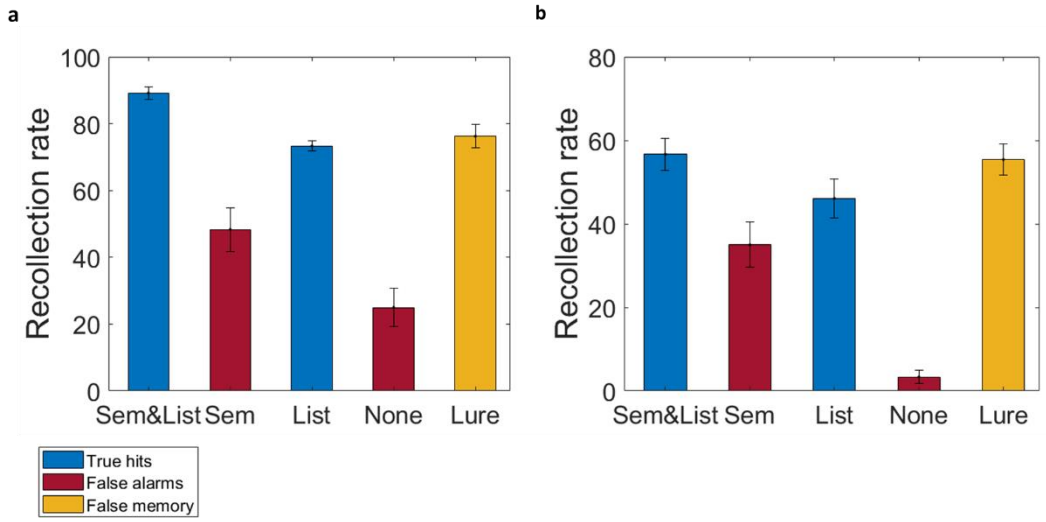


Figure 2.2.1. Simulated pair recollection as a function of their relationship, using methods a and b at encoding. The graphs show the average rate (in percentage) of recollection for different pairs of words for method a (a, Hebbian learning) and b (b, probability-based): from same list with same semantics ('Sem&List'), words with same semantics from different lists ('Semantic), words from same list with different semantics ('List'), pairs with no relationship ('None'), pairs with lure words ('Lure').

Using both method a (Fig. 2.2.1 a) and b (Fig. 2.2.1 b), the model was able to reproduce a similar pattern of pair recall performance to the one of the participants, with highest rates of recollection for words from the same list with same semantics, followed by pairs with lure words, words from same list with different semantics, words with same semantics from different lists, and finally words with no relationship. Overall, the model shows a better performance than the behavioural experiment for retrieval, with higher rates of true hits and lower rates of false recollection for pairs from different lists. This result is explained by the lack of any response bias.

I repeated the dependency measures analysis for the model simulation, both with method a and b at encoding. Both model versions found a significant difference between dependency measures of the data and predicted independent measure only for recall of pairs with lure word ($p<0.05$), while all the other pairs recalled did have a significant difference in dependency measure from the predicted independent measure.

Discussion

The Hopfield associative network simulations, in which semantic associations aid pattern completion, succeeded in capturing the results. These results support the significant role of semantics for episodic memory retrieval, both as continuous semantic distance between words and as abstract categorical semantic context. The temporal context model (TCM, Howard and Kahana, 2001), can be accommodated with the evidence of episodic retrieval as a pattern completion mechanism. The mechanism of pattern completion might potentially be performed by the auto-associative connections between neurons of the CA3 area of the hippocampus (Horner et al. 2015). The temporal context model is based on the progressively changing mapping between temporal context and presented items (whose identity is embedded in the semantic memory). However, the theory lacks an atemporal structure

referring to the presence of a semantic mapping (or context) between related items. A later development of the theory (Howard et al. 2009) proposes a predictive temporal context model (pTCM) for the progressive learning of semantic representations through episodes. While this model uses the idea of a gradually changing temporal context to explain the progressive 'semanticisation' of experience, further studies are needed to explore a model for the role of pre-acquired semantic structure in episodic encoding, consolidation, and retrieval.

The results of the model simulation reproduce the behavioural performance found in the pair recollection test. The auto-associative structure of the model makes the retrieval of the two pairs of words containing the lure dependent upon the presence of an 'episodic' effect. Interestingly, in this case the 'episodic' effect – as defined by Horner et al. 2015 – is not related to the overlapping paired associations between elements presented in the same episode, but on pre-learnt semantic associations between them (confirmed by the fact that the lure word was not presented at encoding). The selection of the model parameters was based on biologically plausible values, as per previous work in the field (Horner et al., 2015; Binte et al., 2020). The results of this model bring additional support to the hypothesis that a mechanism of pattern completion could be involved in the retrieval of semantic memory as associations between concepts, and that this mechanism might interfere with episodic memory. the Hopfield associative network simulations, in which semantic associations aid pattern completion, succeeded in capturing the results.

2.3 A Successor Representation approach to Semantic Interference in Temporal Context Model.

Introduction

In the two preceding sections, I have shown that the false memory effect observed in the DRM paradigm can be reproduced by a Hopfield network model. A second way of looking at the false memory effect is that it might arise because of the way that semantic memory interacts with memory of recent temporal context. This motivates the following section, in which I investigate the temporal context model (TCM), which proposes that episodic recall is guided by a distributed representation of "temporal context" (Howard & Kahana, 2002). TCM has successfully reproduced many behavioural phenomena observed in free recall tasks, as well as providing a biologically plausible explanation for why these phenomena arise. I hypothesised that the false memory effect seen in the DRM paradigm can also be explained within the framework of TCM.

TCM was initially proposed as a descriptive model to capture the rich body of experimental evidence that temporal context plays an important role in memory encoding and retrieval. This rich body of evidence, both in animal models (Mankin et al. 2015, Cai et al. 2016, Rubin et al. 2015, Eichenbaum 2014) and human research (El-Kalliny et al. 2019, Chien et al. 2020, Hsieh et al. 2014), showed how different brain structures and neuronal populations encode temporal relationships between stimuli and process information on the temporal structure of experience. In human cognitive neuroscience, the influence of temporal context becomes apparent in free recall and serial learning tasks (Murdock 1962, Howard and Kahana 1999, Kahana 1996), where recency, contiguity and asymmetry effects provide direct insight into

how the temporal context impacts human cognition. As better defined in the next paragraph, these typical characteristics, resulting from the effect of temporal encoding in human cognition, are referred to as temporal context effects. Moreover, beyond free recall and serial learning tasks, multiple studies have proposed evidence for temporal context effects in a variety of other behavioural findings (Howard et al. 2005, Hamid et al. 2010, Smith D.E. et al. 2022, Howard 2017, Wang et al. 2017).

One of the most used tests for memory performance in human participants is free recall of a list of items or words. This kind of task allows us to detect some key aspects of human memory. Experimental findings and theoretical models of memory alike have been seeking explanations for some of these key principles: primacy, recency, contiguity, and asymmetry. Primacy refers to the phenomenon whereby the first item in the list has a higher probability of being recalled, likely due to rehearsal (Howard and Kahana, 2002). Recency refers to better recall of recently presented items. Memory for an item in the list is helped by the correct recollection of other items that were presented close in time during encoding; this effect is called contiguity. Traditionally, contiguity is measured by the conditional response probability (CRP) as function of time (temporal distance between items at encoding). CRP is a measure of the probability of recalling an item after another item is recalled. When calculated as a function of the relative time distance (lag) between items at encoding, CRP provides a useful measure of lag-recency, or contiguity. This function peaks at 1 lag, and progressively decreases over incremental lag between items. Moreover, this effect is asymmetrical, with forward lags having higher CRPs than backward ones. Asymmetry is another key feature of free recall and serial recall (Kahana & Caplan 2002, Li & Lewandowsky, 1993, 1995; McGeoch, 1936; Raskin & Cook, 1937).

The temporal context model (Howard & Kahana, 2002) accounts for the memory effects described above by hypothesizing that recollection of an item is accompanied by recall of the contextual cues associated with the item at encoding. During encoding, the presented items inform the contextual information, which is used by the model to update a drifting contextual

representation as well as learning a context-to-item association matrix. During retrieval, the combination of this drifting contextual representation and the learned association mean that previously recalled memories work as cues for progressive recalls of items associated with similar context states, i.e. that were presented close in time during encoding.

The original TCM works as a descriptive model. More recently, the temporal context model of episodic memory has been reconciled with the broader reinforcement learning literature, providing an alternative interpretation in terms of reward-maximization and planning. Making use of the temporal difference learning algorithm (Sutton & Barto, 1998), Gershman (2012) and Zhou et al. (2023) propose an adaptation of the temporal context model by showing that under some conditions, TCM is equivalent to use of the "successor representation" (SR) (Dayan, 1993). This line of work successfully provides a different description of the TCM, while providing new experimental predictions.

Here, I adapt the SR version of the TCM by Gershman (2012) by equipping the model with pre-learnt semantic knowledge. This allows us to reproduce the false memory effect seen in the DRM paradigm, and to develop further new experimental predictions, which we aim to test in the future. In the remainder of this section, I will first formally describe the canonical TCM framework, as well as the SR-based variant. I will then use this model to simulate several experiments, including the DRM paradigm.

Methods: Model description

According to TCM, the brain maintains and updates a slowly drifting representation of temporal context, which it uses to facilitate retrieval of relevant memories. For example, during learning in a free recall paradigm, the theory states that each item in the list is added to a contextual representation

of past items, weighted by their recency. During learning (encoding of new memories), each stimulus contributes to the context update, and associations between the context and each item are learned. At retrieval, the likelihood of retrieving a particular item depends on the current temporal context, as well as the learned associations between items and this context. Together, the context and learned associations serves as a temporal cue to trigger retrieval of specific items. Once an item is retrieved, the context is updated, taking into account the previous contextual cues associated with the item, which embeds information on other nearby stimuli learnt at encoding. The recall pattern thus recreates the sequential retrieval of words presented closer in time in the list.

More formally, the model is initiated with a pre-experimental (pre-encoding) temporal context vector, c_t^{IN} , at time *t.* Each stimulus *x* is represented at encoding by a one-hot vector x_t , and by the vector x_i during retrieval. Time steps during encoding are represented by $t \in \{1, \ldots, T\}$, while retrieval time steps are expressed as $i \in \{1, \ldots, N\}$. In TCM, two matrices drive contextual encoding and retrieval; M^{CS} represents learnt context-to-stimulus associations, while M^{SC} denotes stimulus-to-context associations. The function of M^{CS} in the model is to connect a contextual cue to an item, so that the context vector can trigger retrieval of an associated item. On the other hand, M^{SC} triggers retrieval of associated contexts by the presentation of an item.

M^{CS} is reset to zero at the beginning of each encoding trial. For clarity, this matrix is the result of the sum over encoding time steps {1, …, T} of the outer products of all the presented items' vectors at encoding and their associated context vectors:

$$M^{CS} = \sum_t x_t \, c_t^T \qquad [2.3.1]$$

Which is equivalent to updating the matrix at each time step with the:

$$M_{t+1}^{CS} \leftarrow M_t^{CS} + \alpha c_{t+1} x^T{}_{t+1} \quad [2.3.2]$$

Intuitively,M^{CS} represents how each item has been associated with each dimension of the temporal context vector in the recent past. Equation 2.3.2 represents the Hebbian learning rule for M^{CS}.

When a stimulus is presented, the stimulus-to-context matrix M^{SC} drives the retrieval of contextual cues associated with the stimulus. The input to the temporal context, c^{IN}, may therefore include weighted components of previously presented stimuli if the item has already been presented at earlier time points:

$$c_t^{IN} = M^{SC} x_t \qquad [2.3.3]$$

In practice, in many cases M^{SC} is set to the identity matrix I, in which case $c^{IN} = x_t$. When $M^{SC}$is not set to the identity matrix, it has been defined as the transpose of M^{CS}, which encodes backwards stimulus-to-context transitions (Zhou et al. 2023).

When a stimulus is presented, the temporal context is updated by the stimulus, weighted by a constant beta and possibly by M^{SC}, and decaying the previous context with a time constant ρ :

$$c_t = \rho c_{t-1} + \beta c_t^{IN} \qquad [2.3.4]$$

Where c_t is set to be a unit vector via setting of parameters ρ and β. When $M^{SC} = I$, c_t^{IN} takes the form of a vector, with value 1 for the item presented at time t . The context vector at any given time is obtained by the context vector at the previous time step and the input to the temporal context.

During retrieval, the context-to-item association matrix M^{CS} is multiplied by the current context vector to get a predicted stimulus x_i:

$$x_i = M^{CS} c_t \qquad [2.3.5]$$

Where the context-to-item matrix drives the reinstatement of learnt items using the context vector as cue.

A key aspect of TCM is that the context gets updated not only during encoding but also during retrieval. This means that a key prediction of the theory is that retrieval of items should increase the probability of retrieving items that were experienced close in time.

TCM succeeds in reproducing recency and contiguity effects in free recall (Howard & Kahana 2002). Moreover, it describes the asymmetry observed in free recall tasks, since the retrieved context associated with an item contributes to weighting the drifting temporal context at subsequent time steps, but not in preceding ones.

Methods: Model Implementation

Recently, theorists have reframed TCM in terms of reinforcement learning (RL; Sutton & Barto 1998) and successor representations (SR; Dayan 1993, White 1995; Gershman 2012, Zhou et al. 2023). In the Chapter 2 Appendix 1, I introduce some key concepts of RL and SR, to then progress towards a description of the work of Gershman (2012) and Zhou (2023), which serves as skeleton to the new proposed implementation.

The SR provides an efficient and biologically plausible account of episodic memory and free recall tasks. Although the model can reproduce sequence learning and free recall behaviour, previous versions of it do not account for semantic learning or previously learnt associations, which plays an important role in free recall and memory tasks when the learnt items are words (Morton & Polyn 2016, Howard & Kahana 2002b, Polyn et al. 2005, Romney et al. 1993, Bousfield 1953, Glanzer 1969). In humans, investigating how words are

encoded, learnt and retrieved enables to directly test hypotheses on how episodic and semantic memory interact and interfere with each other, and how systems typically associated with only one of the two mechanisms are more broadly involved in an interconnected complex system, where episodic, semantic memory and planning share mechanisms and anatomical structures.

In further work, Howard et al. (2011) proposed a version of the TCM, the predictive TCM (pTCM) that makes use of the TCM machinery to create new learnt semantic representations from repeated associations over time. This proposition underlines how other aspects of declarative memory can be implemented into the TCM, and at the same time how the TCM alone does not suffice, in its current formulation, to provide a holistic model of memory functions. Here, I propose a further implementation of the SR TCM, where the model is equipped with previously learnt semantic knowledge. This implementation succeeds in reproducing the DRM paradigm, while providing new testable hypotheses and predictions that we aim to investigate in future work.

Differently from pTCM, my work focuses on how pre-learnt semantic associations interfere with encoding and retrieval of sequential memory rather than on how new semantic knowledge is consolidated over time. To do this, M^{CS} and M^{SC} are paired with an additional matrix S, which encodes the pre-learned semantic relationship between words. In our model implementation, M^{SC} is set to the identity matrix I (Equation 2.3.3 of the main text); however, the model is equipped with both matrices to allow for simulations beyond this special case. The values assigned to S are not random values, but they are based on cosine semantic distances between selected words calculated using word2vec (Mikolov et al. 2013); this makes the model directly testable with real-word behavioural tasks and easy to adapt to any set of words. Each entry of the S matrix represents the cosine semantic distance between words in the semantic space for word embedding. The values are then transformed into probabilities via the softmax function, so that each column (word to word transitions) sums to 1.

$$S = \frac{e^{(\beta S)}}{\sum e^{(\beta S)}} \quad [2.3.6]$$

Where β is an "inverse temperature parameter" which regulates how peaky this distribution is.

Before retrieval, the M^{CS} matrix in the model is replaced by a weighted matrix, Q, obtained by a linear transformation of the context to item matrix M^{CS} and of the semantic association matrix S as follows:

$$Q = \mu M^{CS} + (1 - \mu) S \quad [2.3.7]$$

Where $\mu \in [0\ 1]$. If the value of μ is 1, the model behaviour is equivalent to Gershman (2012) since the semantic association component would bring no contribution to the SR matrix Q. When, however, $\mu = 0$, the model retrieves items in free recall without considering any contribution from the encoded order of words in the presented list, but only based on pre-learned semantic associations. Values between 0 and 1 thus modulate the contribution of temporal and semantic associations with hybrid simulated behavioural results.

The encoding (Equations 2.3.1 and 2.3.4) and retrieval process (Equations 2.3.5 and Equation 2.6A1) follow the same dynamic of previously published studies (Gershman 2012, Zhou et al. 2023), replacing M^{CS} with Q.
To avoid that the model recalled the same word more than once, we set the activation of each recalled word (x_i in Equation 2.3.5) to 0 before moving to the next recalled word.
The model was tested to reproduce the typical free recall effects described previously, as well as additional properties of the integration of semantic knowledge into the SR matrix.

Results

Each simulation was performed, and the results averaged over 500 trials. Each trial simulated the encoding and free recall of a single list of 30 words. At the beginning of each trail, the SR matrix and context vector were reset to baseline, so that learning and retrieval were independent across trials. At the beginning of each trial, the only pre-learnt contribution to SR came from the matrix of semantic associations S.

The model was presented with a list of 30 words. The words were selected from a pool of high frequency nouns (https://memory.psych.upenn.edu/Word_Pools), and controlled to have a balanced distribution of semantic distances between words (as cosine distance in word2vec, Mikolov et al. 2013) at different serial position distances, so that the distance in serial position did not covary with the degree of semantic distance in word2vec.

First, simulations were run to reproduce the SR TCM behaviour described by Gershman (2012) and Howard and Kahana (2002). To do so, the Q matrix was reduced to the M^{CS} matrix by setting $\mu = 1$ (Fig.2.3.1).

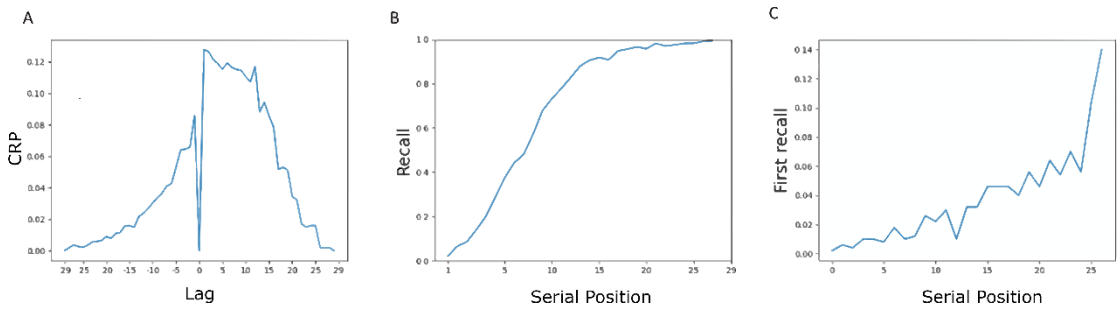


Figure 2.3.1. The results of our first simulations reproduced the effect of asymmetry and contiguity (A) and recency (B and C). A) shows the simulated lag-CRP of the model averaged over 500 trials. CRP (conditional response probability) is presented as function of the distance in position within the presented list (Lag) between words recollected in sequence; B) shows the absolute probability of recalling individual words in each serial position. Words in later serial positions are more likely to be

recalled, a recency effect (because their associated context is closer to the temporal context of the retrieval testing time). Similarly, C visualises the probability of each word being the first one to be recalled during simulated free recall. Words that were presented later in the list are more likely to be recalled first.

The same simulations were ran modulating the value of $\mu \in [0\ \ 0.25\ \ 0.5\ \ 0.75\ \ 1]$, progressively moving from a semantic-only weighted Q to a temporal context-only SR matrix M^{CS} (Fig 2.1A1).

As our simulations show, the linear weighting of temporal and semantic components in the model result in a behaviour that is progressively more or less driven by M^{CS} , which encodes learnt temporal associations, or by S, which expresses the contribution of previously learnt semantic relationships between words.

We then tested our model to simulate the DRM effect. To do so, we provided the model with a list of 15 semantically related words (Roediger & McDermott 1995). The S matrix was prepared using the same word2vec cosine distance value between words, as described above. To reproduce the false memory effect, while the presented list had 15 items, the S matrix encoded pre-learnt semantic knowledge for 16 items: the 15 included in the list and one additional word (lure word), semantically related to all the words in the list but not presented at encoding. We ran our simulations as above, with $\mu \in [0\ 0.25\ 0.5\ 0.75\ 1]$.

Our model was able to reproduce the false memory effect of the DRM paradigm (Fig.2.3.3), with the lure word (false memory), not presented at encoding, never being recalled in the trials for $\mu = 1$. These simulations show how the model not only recall the 15 words presented at encoding, but also the lure word (false memory) semantically related to them via the S matrix when $\mu < 1$. In case of $\mu < 1$, the model includes the weighted effect of the

semantic associations, and recalls the lure word with higher probability the smaller the value of μ (shown as the 16th word on the x axes, 'lure', in Fig.2.3.3).

The effect of contiguity, asymmetry and recency where preserved. Moreover, the false memory effect (i.e. the probability of recalling the lure word) tended to happen at the end of the recalled list of words, which is consistent with behavioural evidence (Roediger & McDermott 1995). In our model, this is explained by the fact that after being recalled, each word activation (x_i in Equation 2.3.5, main text) is set to 0 after a word was recalled. By the end of the recalled sequence, the lure word is one of the few choices left for the model to recall. When the value of μ is set to lower values, the lure word is progressively recalled in a more distributed manner over the recalled order. For simplicity, I here reported only 5 examples of the simulations with values of $\mu \in [0\ 0.25\ 0.5\ 0.75\ 1]$. While the model behaviour is quite similar for $\mu \in [0.25, 0.5, 0.75]$, with a hybrid behaviour between temporal context-driven and semantic association effect, it is clear how extreme values of $\mu = 0$ and $\mu = 1$ deviates from the rest of the simulations.

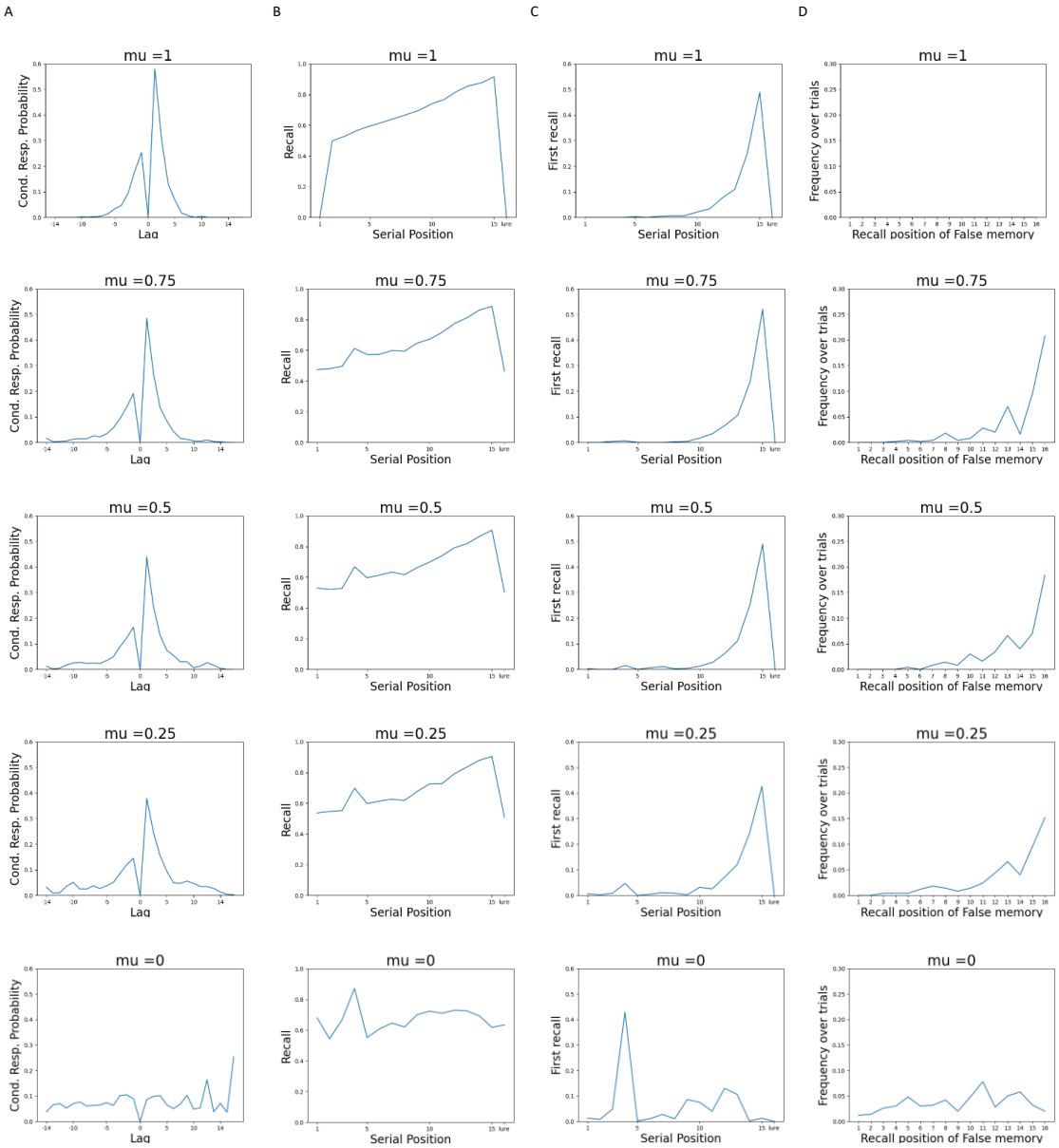


Figure 2.3.3. Simulation results for DRM effect. The model was presented with 15 words at encoding. The plots columns B and C, respectively probability of recall and probability of first recall for all words, represent the lure word (false memory) as the 16th word on the x axis, since this word was not presented at encoding. In this case, the lure word is the 16th words on the x axis of the second and third column, where recall probability and probability of a word being recalled in first position are shown. Row 1 shows the results for $\mu = 1$ (only temporal context effect, without any contribution from semantic associations), while row 5 includes results for $\mu = 0$, when the effect of temporal context is fully removed. The lag-CRP graph (A) shows how the effect of contiguity and asymmetry are preserved in the model. Similarly, B and C show the recency effect, with the overall higher chance of recalling words presented

later in the list at encoding (B) and of recalling later words in first position (C). Interestingly, down-weighting the effect of M^{CS} with smaller values of μ quickly removed the temporal context-related behavioural effects of recency, contiguity, and asymmetry. Column D shows where the false memory (recall of the lure word) happens in the recalled list, if the lure word is recalled.

Discussion

In Chapter 2.3 I tested how the SR formulation of the TCM can capture effects of underlying semantic knowledge on the encoding and free recall of lists of words. Previous work (Gershman 2012, Zhou et al. 2023) already showed how TCM and its SR formulation can reproduce typical effects, such as contiguity, recency, and asymmetry. With this work, I modulated the temporal context encoding in the M^{CS} with pre-existing semantic associations between stimuli, found in S. This means that temporal associations between stimuli are not learnt in a naïve manner, but they are encoded overlaying the substrate of pre-acquired mapping between stimuli from previous experience.

My contribution to this existing body of work explores another fundamental aspect of declarative memory, which often interact and interfere with episodic memory, i.e., pre-acquired semantic knowledge. Although not conclusive yet, this is a first step towards the unification of aspects of semantic and episodic memory in one flexible and coherent model, which can be directly behaviourally tested. Moreover, the use of SR allows to place the investigation of episodic memory dynamics within the realm of future planning and prediction, as well as context-dependent behaviour (Geerts et al. 2023, Gershman 2018).

This model's implementation preserves the ability to simulate the effects of contiguity, asymmetry, and recency. It does not, however, reproduce primacy. The choice of not implementing the model with a primacy effect was justified by the will to keep the model as relevant as possible and as simple as possible for encoding and retrieval processes. The effect of primacy can be quite easily

induced in the model by adding a rehearsal (Tan & Ward 2000) or attention-like element to it. The temporal context model per se does not entails the primacy effect. However, in previous versions of free recall models (Sederberg et al. 2008), primacy was induced by additionally boosting the learning rate at encoding of the first words in the list (which can be interpreted as a proxy for attention), which is then decayed over later serial positions. Since inducing primacy in our simulations would have not contributed to the current purpose of the model, we decided not to do so at this time. However, if inducing primacy effect will be needed in future applications of the model, this can be done following Sederberg et al. (2008) as follows:

$$\alpha_i = \alpha_{t_0} e^{-\alpha_d (i-1)} + 1 \qquad [2.3.7]$$

Where α is the learning rate, α_d is the decay of learning rate over time and $\alpha_{t_0} + 1$ determines the strength of boost in learning rate for the word in serial position one (adapted from Sederberg et al. 2008).

This work shows how a relatively simple implementation of the model with pre-acquired semantic knowledge (embedded as pre-learnt associations between items) can provide a descriptive account of false memory in the DRM paradigm. However, these first simulations open to new theoretical questions and testable hypotheses. Firstly, this first proof of principle used a linear combination of the SR matrix and the pre-learnt semantic association matrix. However, this assumption can be challenged by exploring other mathematical non-linear combinations of the two matrices, where the modulatory effect of pre-learnt knowledge on the temporal context can follow more complex dynamics.

In future work, we aim to develop different implementations of the model, with non-linear alternatives, and to test the predicted behaviours fitting real-world data. This can be done with standard behavioural tasks, such as the DRM paradigm in free recall, as well as with more complex experiments (see below), where an event boundary element can help disentangle different predictions. It is interesting how the S matrix was not created by providing a set of random values to the learnt semantic associations, but these were based on word2vec,

a commonly used model of word embedding trained on real-world data. This makes the model ready to be used for real world testing and applications, where the matrix can be easily adapted to the specific stimuli presented during encoding.

In the current work, our interest focused on pre-learnt semantic knowledge. However, given that the model is a learning algorithm, it would be interesting to test how the SR implementation of TCM would perform if used as an algorithm to learn new semantic associations. In a similar way to the pTCM (Howard et al. 2012), Gershman (2012) version, basing learning on predictions rather than Hebbian plasticity, would be a natural candidate to reproduce the long-term effects of repeated presentation of items (such as repeated presentation of related words, which induce over time semantic knowledge) on the pre-encoding SR algorithm. This longer time scale acquisition of associative knowledge would replace in the current implementation the externally integrated word2vec distances with an internal SR-generated structure. Not only SR is a promising approach to model the contribution of pre-learnt knowledge to episodic (as in temporally embedded) learning, but it already proved itself relevant to language structure acquisition (Stoewer et al. 2022), towards a more unifying approach to memory, navigation, planning, and language and thought organisation.

Previous work had implemented the TCM to account for the semantic component of presented items. In the Context Maintenance and Retrieval (CMR) model by Polyn et al. (2009), the classical TCM model is equipped with pre-experimental (pre-learnt semantic associations) and experimental (based on temporal sequence learning) associations between items and context. In the CMR, an additional feature layers play a critical role in representing the stimuli (items) and associating them with temporal context during both encoding and retrieval. These feature layers encode the properties of the items presented during an experimental task (e.g., a word list in a free recall experiment), and they interact with the context layer to facilitate memory encoding and retrieval. The feature layer serves as a representation of stimuli at a perceptual or semantic level, while the context layer captures temporal information. Together, they form an associative memory system that binds

features (the content of the memories) to the context (the temporal aspect). During encoding, the feature layer (representing the current stimulus) helps to shape the current context by modifying the context vector, linking the item's features with the temporal context. During retrieval, the context layer guides the reactivation of feature representations, with the system retrieving items based on the similarity of the current context to the context stored in memory traces. In CMR, the context is a distinct representation from the feature layer, and the key interaction in CMR is the binding of items to their temporal context, facilitating the retrieval of the original memory trace through context reinstatement. Similarly to the CMR for the TCM model, conceptually, our model aims to provide a similar implementation of the pre-learnt semantic knowledge to the SR TCM of Gersham. However, unlike CMR, our model does not explicitly introduce a further feature layer but delegates the semantic information to the S matrix. Via the SR matrix, our model learns state transitions that predicts the future states (or items) likely to follow from the current state, informed by pre-acquired semantic knowledge (S matrix); there is no explicit binding between individual features and context. Instead, the model predicts future states based on the current state. Both CMR and our model successfully reproduce the effect of recency, asymmetry and contiguity. However, this foundational work did not account for investigating predictions and model fitting of our model compared to the CMR. However, we aim to continue this work with further model comparison between CMR and our model, and model fitting on new experimental data.

To investigate what kind of computation better explains the degree and nature of the interaction between SR leaning and pre-learnt associations, we plan to test our model starting with behavioural tasks (Fig.2.3.4). Informed by previously published experimental work (Smith et al. 2013, Pu et al. 2022), I am first planning to test the basic model dynamics to inform further implementations. In their study, Pu et al. (2022) investigated the mechanism of event boundaries in temporal order memory with a series of behavioural tests and a computational model of temporal context. While their work defined event boundaries as changes in the frame colour of the presented items, the definition of context can entail a variety of features, depending on the focus.

In the case of language, context can also be interpreted as semantic context of words related by meaning. Our model is well suited to answer whether semantic context interacts with the drifting temporal context in a similar way as environmental cues do. We aim to adapt the behavioural tasks utilised by Pu et al. (2022) by organising the ordered items using abstract event boundaries, with the change of context defined by the change from one semantic realm to another (such as 'cold'-related words followed by 'music'-related words). This implicit semantic context would replace the visual (environmental) change in the colour of the frame surrounding the items in the original experimental conditions. This test will provide insight on what the nature of context, the role of abstract contexts in learning and memory, and possibly the possible computations guiding the integration of complementary systems of knowledge, or complementary aspects of the same system.

When focusing on the role of context in memory and learning, the context repetition effect (Smith et al. 2013) - i.e. the fact that the recollection of stimuli experienced in the past is enhanced by the repetition of the context in which they were experienced without the repetition of the stimulus per se - figures as a testable mechanism with theory-driven experimental predictions. Based on the temporal-different SR model (TD-SR, Dayan 1993), Smith et al. (2013) use the idea of context as repeated presentation of temporally contiguous items to show how the repetition of the same context enhanced memory for items previously associated to them (and hence predicted when the context is presented), when the items themselves are not represented. The context repetition effect, was tested using both images and words as presented items. Our implementation of the SR model with pre-learnt semantic structures allows to test further characteristics of the repetition context effect when this interacts with underlying semantic structures and abstract contexts. While the repetition of the same two words can provide the contextual element to boost memory for a third associated item (Smith et al. 2013), we wonder if at a higher level of conceptual hierarchy, the repetition of different words from the same semantic context would boost memory for items originally presented in association to those words, only thanks to the reinstatement of closely related concepts. This

would suggest that higher level of abstraction can contribute to lower-level predictions and directly contribute to how memory is formed and retrieved.

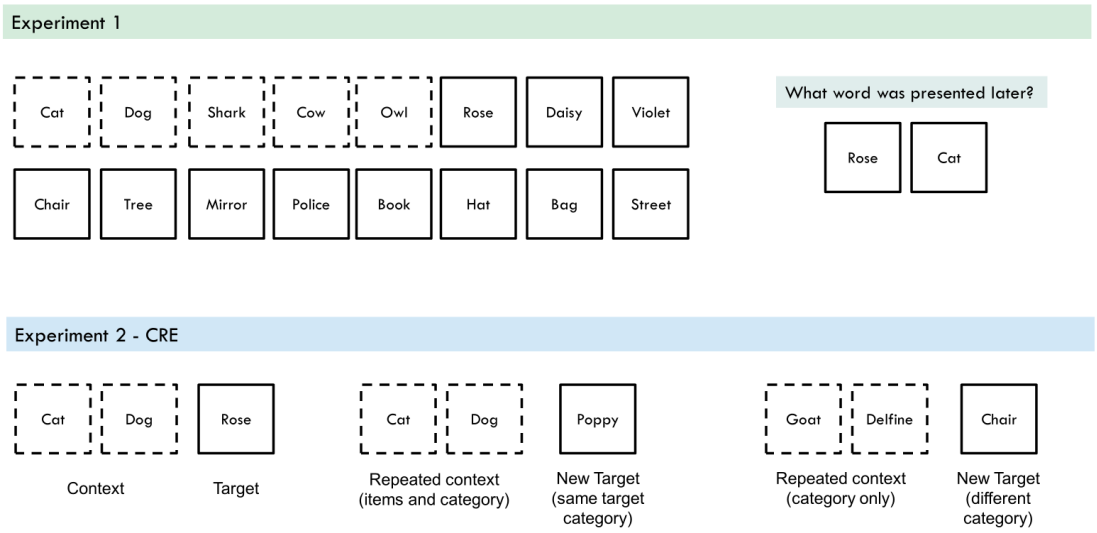


Figure 2.3.4. Future experiments. Experiment 1 (adapted from Pu et al., 2022) will test performance for order recollection of words presented in a list. The experiment will have two conditions. In the semantic context boundary condition, the list is organised so that words are grouped based on semantic categories. In the control conditions, words will be presented in random order without semantic grouping. Participants will then be asked to remember which word between a tested pair was presented later. The pairs of words tested will be picked as per Pu et al (2022). Experiment 2 will test the CRE effect, as in Smith et al. (2013) experiment, where the first two items of a 3-item series represent the contextual cue, and the third item is the target. We will also test whether a similar CRE effect will be elicited not only by the repeated context provided by the same two words, but also by the presentation of two different individual words from the same semantic group (category) of previously presented contextual cues (animal category in the example above). Refer to Pu et al. (2022) and Smith et al. (2013) for further details on experimental design for experiment 1 and 2, respectively.

These future experimental tests will help to better understand the nature of the interaction between semantic and episodic memory, as well as the mutual contributions of the temporal context in memory processes and language

processing. This further part of the work will also contribute to develop additional model implementation. In particular, it will bring new understanding on the nature of the computational nature of the combined contribution of SR learning (M^{CS}) and pre-acquired semantic knowledge (S), where different linear and non-linear combination of the matrices can be tested via model fitting.

Although purely theoretical, the SR TCM and its implementations provide biologically plausible hypotheses on the neurobiological structures and mechanisms contributing to these computations in humans (Geerts et al. 2023, Gershman 2018). Human neuroimaging methods have been developed to investigate SR-like representations in the human brain, which are providing a growing body of evidence to support the hypothesis of SR models being better suited to explain neuronal activities and human cognitive processes (Garvert et al. 2017, Ekman et al. 2023, Russek et al. 2017, Russek et al. 2021) In future work, informed by the above mentioned and further hypothesis-driven tests, the biological plausibility of the SR TCM and of its interaction with semantic memory and language could be investigated via ad hoc neuroimaging work to provide understanding on the anatomical structures and the temporal dynamics contributing to contextual and ordered memory in humans.

2.4 – General Discussion of Chapter 2

Overall, Chapter 2 proposed new behavioural and computational approaches to investigate the role of semantic and temporal context in false memory. It also developed new testable hypotheses for the possible involved mechanisms.

In Chapter 2.1, I first studied the interplay between temporal and semantic contexts using the Deese-Roediger-McDermott (DRM) paradigm. These behavioural experimental findings suggest that pattern completion mechanisms, like the ones usually studied in episodic memory in the hippocampal formation, might be responsible for the semantic-driven false memory effect. This inspired the development of a Hopfield associative network model (Chapter 2.2), which was able to reproduce the findings of Chapter 2.1 and proposed a realistic mechanistic understanding of a possible role of the CA3 area of the hippocampus in inducing false memories in healthy individuals. The way in which the sequence of words is learnt in the Hopfield model is based on a Hebbian-like learning process. This is based on the temporal correlation between word-associated simulated neuronal activity. This mechanism mirrors a form of temporal context at the neuronal level, not dissimilar to the one proposed by the TCM theory. However, in the case of the Hopfield network, the temporal context is not explicitly represented in the model as an independent factor, but it is implicitly resulting as the effect of concatenated neuronal activities via learnt associations of sequential neuronal firing.

This insight motivated the development of a more sophisticated model of temporal context using the SR computational framework in Chapter 2.3. In this case, I looked at the problem of false memory via semantic interference from the perspective of SR models from reinforcement learning (see Chapter 2 Appendix 1) and studied the ability of the SR version of TCM to encapsulate the influence of pre-existing semantic knowledge on word list encoding and free recall. In this case, the temporal context was explicitly represented in the model and used for memory retrieval via eligibility traces. This work explores

how pre-learnt semantic associations modulate temporal context encoding through a linear combination approach between contextual cues. As mentioned, future implementation of the model will explore non-linear combinations, aiming to uncover more complex interactions between semantic knowledge and temporal context. Although the SR aims to reproduce the false memory mechanism at a higher level of description, it accomplishes so while encapsulating biologically plausible neuronal mechanisms that are easily conceptually related to the same medial temporal cortex structures responsible for both memory encoding and retrieval, as well as future planning.

This approach to the computational study of the interaction between semantic and temporal contexts has proven fruitful to develop and test different hypotheses via in silico simulations and real-world experiments. The validity of this approach was further investigated using another computational approach, within the active inference framework (see Chapter 2 Appendix 2), and will be additionally put under test in the future with behavioural human experiments and model fitting (see also Chapter 2.3), to better understand the complex dynamics of contexts integration in human cognition.

Chapter 2 Appendix 1. TCM in Reinforcement Learning

Here, I provide the theoretical background to describe how the TCM has been integrated computationally in the reinforcement learning, and in the successor representation framework. This computational approach to TCM served as the foundation to my model implementation.

The goal of RL models is to maximise cumulative reward or value, which guides learning and behaviour of the RL agent. Broadly, RL models can be classified as model-based and model-free (Sutton & Barto, 1998). Model-based RL agents are equipped with an internal model of the world, which informs policies based on predictions about which actions lead to which states. The internal model of the world can be updated with new evidence on associations between states, and between states and reward in the environment. Model-free agents, on the other hand, do not implement a flexible model of the surrounding environment, but rely on decisions based on simple associations between states and future expected rewards. In many model-free methods, this involves estimating the “value function”, which maps from states to expected future rewards. This learning is often mediated by a reward prediction error (RPE), which has been linked to dopamine neurons in the mid-brain (Schultz, Dayan & Montague; 1997).

In addition to the dichotomy between model-free and model-based methods, there are hybrid models, which incorporate elements of both model-based and model-free learning. One strategy is to learn a “successor representation” (SR; Dayan 1993), a representation of states in terms of the future states they predict. RL methods based on the SR represent a middle ground between traditional model-based and model-free methods, because they incorporate some information about an environment’s transition structure, while not requiring offline simulation for planning. Interestingly, some theorists have drawn parallels between aspects of the SR and neural firing patterns in the

hippocampal formation (Stachenfeld et al. 2017), the dopamine system (Gardner et al. 2018), as well as aspects of human and animal behaviour (Momennejad, Russek, et al. 2017, Russek et al. 2021, Piray & Daw 2021, Geerts et al. 2020, Geerts et al. 2023).

The key idea of SR models (Dayan 1993, Gershman 2018) is to learn an SR matrix of expected future occupancy of each state from each other state. The SR matrix (M) works as a predictive map between each state and the other states over time. More formally, M expresses the discounted cumulative number of time steps for which a future state s' is expected to be occupied by the agent, under a policy π starting from state s (Russek et al. 2017, Gershman 2012):

$$M^{\pi}(s,s') = \mathbb{E}[\ \sum_{t=0}^{\infty} \gamma^{t}\, \delta(s_t, s') | S_0 = s\ |] \qquad [2.1A1]$$

$$= \mathbb{E}\left[\delta(t,j) + \gamma M_{tj} | s_n = i|\right] \qquad [2.2A1]$$

$$= \ \sum_t T_{it}\left[\delta(\mathrm{t},j) + \gamma M_{tj}\right] \qquad [2.3A1]$$

Here, the Kronecker delta function $\delta(n,m)$ assumes value of 1 if $n = m$ and the value of 0 otherwise. Where i represents the current state, j the occupancy of future states, t indexes all possible future states.

One of the strengths of the SR model is that the value function can be easily (linearly) computed, given the SR and an estimate R of the rewards found in each state.:

$$V(i)\ = \sum_j M(i,j)R(j) \qquad [2.4A1]$$

In the SR reward function, $V(i)$, is decomposed as the dot product of the rewards R and the predictive state occupancy matrix M . The $M(i,\ j)$ vector represents the cumulative (averaged over possible trajectories starting in state

i) discounted occupancy of state j under a policy π. The factorisation of the value function into M and R provides the SR model with a higher degree of flexibility: a change in reward can be learnt independently from an update of the transition dynamics over state occupancy, which are slower to re-estimate. These features of the SR algorithms are particularly beneficial when reward changes over time, while state transitions are stable.

The SR cumulates state occupancies, rather than rewards. Similarly to TD learning of value, SR updates via a TD prediction error (PE) resulting from differences between expected and observed state occupancy (Russek et al. 2017, Gershman 2018):

$$PE_t(j) = \delta(s_t, j) + \gamma M(s_{t+1}, j) - M(s_t, j) \qquad [2.5A1]$$

The first argument on the right-hand side can assume value 1 if $s_t = j$, and 0 if $s_t \neq j$; γ is a discount factor with values $\in [0,1]$. M is updated via TD learning rule (equation 2.7A1), as a result of PE_t, modulated by learning rate and eligibility trace, which informs the model on recent state occupancy (equation 2.6A1). The model updates its prediction of the time spent at each state via a vector that represents expectation for all future states, increasing the expected occupancy following positive prediction error, and reducing it in case of negative prediction error.

In both model-free RL and when estimating the SR, learning can be sped up using "eligibility traces". These traces allow for credit assignment to recent past states, weighted by recency. Once a state is occupied at trial t, the eligibility trace $e_t(i)$ for that state i is reinforced and the traces of other states decay exponentially. As noted by Gershman (2012), the eligibility trace $e_t(i)$ corresponds, mathematically and conceptually, to the temporal context vector in TCM (equation 2.3.4). Formally, $e_t(i)$ is defined as:

$$e_t(i) = \begin{cases} \gamma\lambda e_{t-1}(i) & if\ i \neq s_t \\ \gamma\lambda e_{t-1}(i) + 1 & if\ i = s_t \end{cases} \qquad [2.6A1]$$

Where the decay parameter $\lambda \in [0,1]$ modulates the states that will be updated, with higher values of λ allowing for more ancient state to be eligible. This informs the model on recent state occupancy, with a corresponding role to c_t.

When using eligibility traces, the SR matrix is updated in each component as (Gershman 2012):

$$M_{ij} \leftarrow M_{ij} + \alpha \left[\delta(s_{t+1}, j) + \gamma M_{s_{t+1}, j} - M_{s_t j} \right] e_t(i) \qquad [2.7A1]$$

Where the learning rate α takes values between 0 and 1. The eligibility trace acts as the context vector. By multiplying it by M, $e_t(i)$ predicts the probability of each item of being retrieved.

At the same time, the model learns not only state occupancy via M , but also rewards R at each state to obtain the value function $V(i)$ using equation 2.4A1.

The SR affords more flexibility in decision-making than simple model-free learning because value is generalised over all the states that predict similar futures. Since representations of reward and transitions are factorised, changes in the reward function can be flexibly adapted to. However, it is not as flexible as model-based algorithms, because changes in the transition functions will have to be slowly relearned.

These computational characteristics matched with its biological plausibility makes SR a promising approach to test existing theories and models of neuronal and behavioural mechanisms.

Gershman (2012) successfully pointed out how TCM can be expressed in equivalent terms by the TD learning rule of SR. In particular, the SR learning

rule corresponds to TCM learning of sequences of items in the case in which each item is presented only once, without repetitions.
In its simplified form, item-to-context associations in TCM update via a Hebbian rule as follows:

$$M_{ij} \leftarrow M_{ij} + \alpha x_{t+1,j}\, c_{t,i} \qquad [2.8A1]$$

So that each item is bound to the context vector at the time when it was presented. This is equivalent to equation 2.3.2 for M^{CS} Hebbian learning.
In case items are presented only once, the SR learning rule can be reduced to:

$$M_{ij} \leftarrow M_{ij} + \alpha\, \delta(s_{t+1}, j)\, c_{t,i} \qquad [2.9A1]$$

Which is equivalent to Equation 2.8A1 in the special case where items are never repeated. The difference between the traditional TCM vs SR learning rule becomes apparent only when items are presented more than once, with TCM being based on Hebbian associative learning, while SR learning uses prediction error on state occupancy to update M. This means that, while in TCM the strength of learnt associations increases every time two items are presented close to each other in time, in SR the update is proportional to the prediction error computed at each transition. This results in experimental predictions that would disentangle the different behaviour between TCM in its original form and the SR TD adaptation proposed by Gershman: if two items are presented in sequences during encoding multiple times, the Hebbian learning in TCM would consistently strengthen the association between two items, while SR predicts a limit for increase in association strength between repeated items. This is due to the SR learning via prediction error (Equation 2.8A1). Once the repetition of two items is learnt and highly predictable (i.e., the prediction error tends to 0), the agent's learning of the association would plateau.

Gershman's proposition of the TCM using SR TD learning can be summarised as follows, in encoding and retrieval. The model simulations of free recall reproduce the same effects of recency, asymmetry, and contiguity in free recall that the TCM of Howard and Kahana succeeded in reproducing. During encoding, a list of items is presented in order. At the beginning of each trial, the SR matrix M and the eligibility trace (context vector) c_t are initiated to all 0 values. When an item is presented, the eligibility trace is updated as per Equation 2.3.4, main text, while the SR matrix is updated following Equation 2.8A1. At the end of encoding, the SR matrix is learnt. At retrieval, the current temporal context (eligibility trace c_t) is used to prompt encoded items associated with similar temporal context. Mathematically, this is obtained by the product between the eligibility trace c_t and the SR matrix M. The result of this operation is a vector with a strength distribution across all items. This vector of stimulus activations then needs to be turned into an actually retrieved item. In Gershman (2012), this is achieved by feeding the vector into a linear ballistic accumulator applied to each item (LBA; Brown & Heathcote 2008) with a set threshold. In Gershman's model, the LBA's noise and slope are set via random numbers for each item, which results in different times to cross the set threshold. The first item that passes the accumulator threshold is recalled by the model. In our implementation of the model, following Zhou et al. (2023), the model simply normalises the activation vector to obtain a probability distribution of future items, and samples from this distribution.

Once an item is retrieved, the eligibility trace is updated as per Equation 2.6A1. In this case, the context update is driven by both the features of the presented item, and by the reinstated context associated with it. New items are then retrieved by the updated context. Intuitively, this retrieval mechanism already brings understanding on how the model will reproduce the asymmetry and contiguity effects. At encoding, the item-to-context learning enables each new items to retain information on previously presented items, since the context associated with each new presentation has been updated by the previously seen elements. At retrieval, once an item is retrieved with and the context is updated, this serves as cue to retrieve items with similar contexts at encoding.

Since only following items but not preceding ones in the encoding phase are embedded in contextual information about preceding items, it is more likely for the model to retrieve items presented in later positions rather than earlier positions in the list (asymmetry), with the ones closer in time more likely to be retrieved (contiguity).

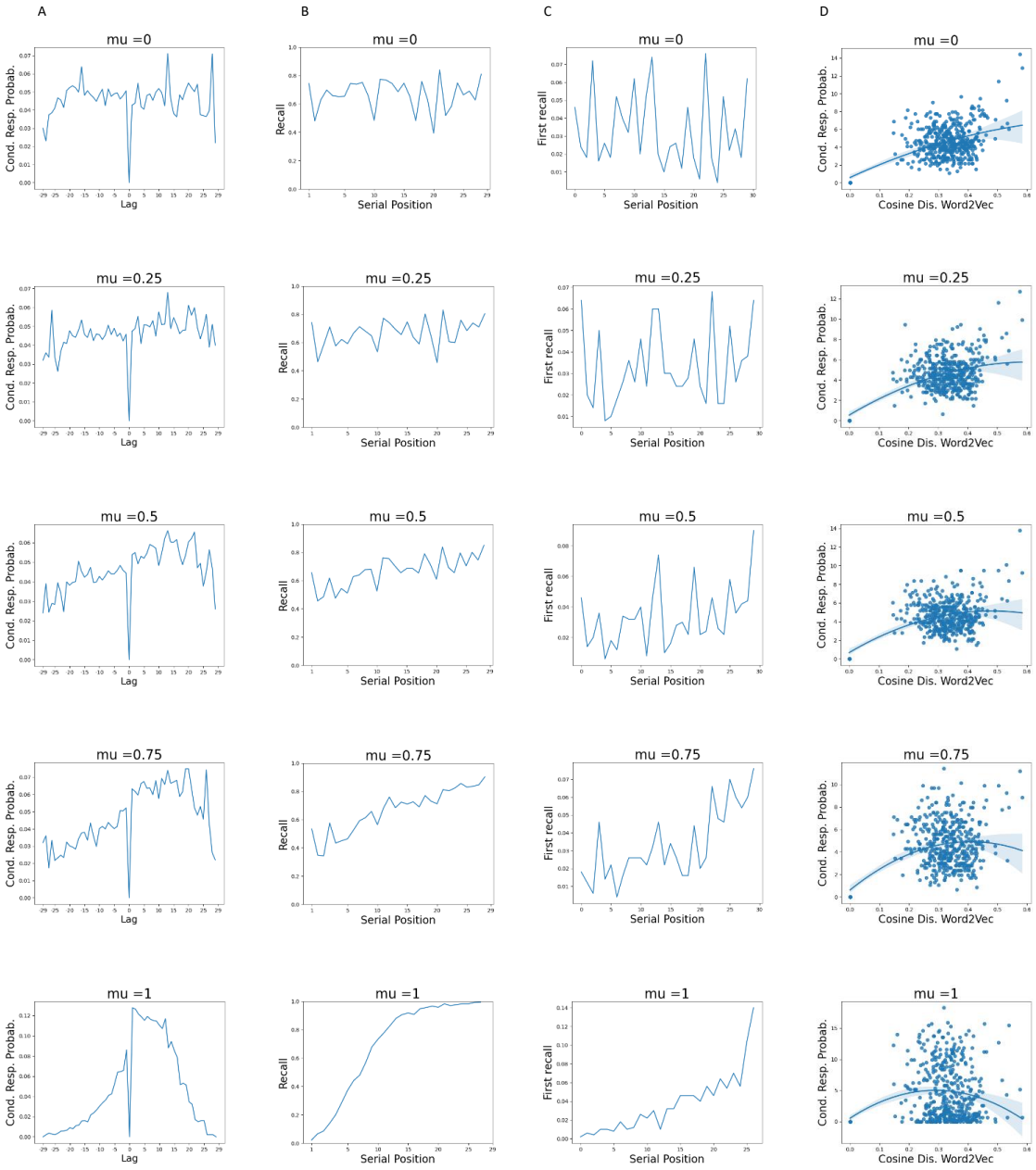


Figure 2.1A1. Each row reports the result of one set of simulations with a specific value of μ: in order, from top to bottom row: $\mu = 0$; $\mu = 0.25$; $\mu = 0.5$; $\mu = 0.75$; $\mu = 1$. .Column A reports lag-CRP, column B shows probability of recall, column C represents probability of first recall, while column D shows the effect of semantic cosine distance between two words on the CRP. It is clear how the closer μ is to 0, the smaller the influence of the temporal context over recall, but the higher the effect of the semantic associations between words. When the value of μ progressively moves closer to 1, the model behaves more as temporal context driven.

Chapter 2 Appendix 2. A Bayesian account of Episodic memory and misremembering.

Introduction

This appendix concerns the computational anatomy of episodic memory. It introduces a neuronally plausible belief updating scheme — based on active inference — to reproduce a canonical false memory (DRM) paradigm *in silico*. Crucially, this model is based upon a first principles account of what it means to remember, and the requisite generative model of successive episodes. The ensuing model generates neuronal and behavioural responses, in the form of local field potentials, reaction times and choice behaviour. These synthetic responses can, in principle, be used to disambiguate distinct generative models or architectures that underwrite episodic (and false) memories. In this foundational chapter, we consider two mechanisms for episodic memory: first, an implicit mechanism that rests upon perceptual learning; where recently encountered stimuli are inferred with greater precision or confidence. Second, equipping the generative model with an explicit representation of time, to enable inferences about when a particular stimulus or episode was encountered. Both models reproduce the basic phenomenology of the false memory paradigm, with some subtle nuances in belief updating and accompanying neuronal (electrophysiological) correlates.

Here, I consider episodic memory under the active inference framework — and describe a first attempt to address some key characteristics of declarative memory and pre-learned associations (semantic memory). This approach is both integrative and simplifying, in that it integrates elements of different models of episodic memory, such as models based upon attractor networks (Burgess et al. 2002), scene construction (Hassabis and Maguire, 2007, 2009), and simulation theory (Zeidman et al. 2015). At the same time, it uses a simple generative model to account for core aspects of episodic memory.

These core aspects emerge when making sense of the world under certain kinds of generative models.

We focus on a core aspect of episodic memory, namely, its sensitivity to context as revealed by the Deese-Roediger-McDermott (DRM) paradigm — a paradigm designed to reveal the effect of (semantic) context through a kind of false memory effect. Using the DRM paradigm, we tested two plausible generative models that entail memory. The former builds upon a typical (Markov decision processes) model used to illustrate active inference in working memory tasks (Parr & Friston, 2017) based upon the *implicit* learning of recent associations. The second model includes learning to encode temporality; i.e., the *explicit* order of events over time.

Both models reformulate episodic memory as an inference process: the agent infers which past episode offers the best account for the current experience. Our work builds upon four stipulative aspects of declarative memory. First, the process of retrieval is a process of inference: the agent infers that the experience happened in the past. Second, episodes presuppose temporal (ordinal) organisation. Third, retrieval is necessary for declarative memory. Finally, remembering, as inference, is essentially a (re)constructive process.

The reconstructive nature of memory means that the inference in the present not only depends on the immediate precedents, but also on a more remote past. This (semi-Markovian) aspect differentiates our work from most models based on the Markovian assumption. Under an active inference perspective, memory could be embedded solely by updating (Bayesian) *beliefs about hidden states*, or by *learning the parameters* of the mapping between states and observations. The key difference between the two is that the former represents the inference about states of the world, while the latter entails learning the parameters of the generative model along with inference about the states. Retrieving an episode under the latter necessitates both forms of belief updating (about the hidden states and model parameters) with an inferential component that rests on working memory. In neurobiological terms, to simulate episodic memory, the retrieval must therefore depend upon both perceptual inference and experience dependent plasticity.

In what follows, I first briefly summarise the core principles of active inference, focusing on the difference between inference and learning under this framework. We then describe two distinct generative models, emphasising their similarities and differences. Equipped with these generative models, the DRM task can then be reproduced *in silico*, by (i) specifying a particular generative model — in terms of how hidden states generate observable outcomes, and (ii) using standard marginal message passing schemes to simulate belief updating (and implicit decisions). Finally, we use the results of the simulations — based on implicit and explicit generative models — to illustrate the behavioural correlates of episodic retrieval. Using two generative models can be read as evaluating the face validity of two foundational hypotheses for the computational anatomy of episodic memory in the human brain; namely, with and without an explicit representation of time.

Active Inference

Active inference provides a unifying account of action, perception and learning, based on the notion that the brain performs approximate Bayesian inference under a generative model of the world, under which it tries to maximise model evidence (or minimise variational free energy, or surprise). The basic assumption is that any living organism maximises the evidence for its implicit model of the world or, equivalently, minimises variational free energy (Friston, FitzGerald, et al. 2017; Friston, Rosch, et al. 2017), building on the predictive brain hypothesis (Hinton, Dayan, Frey, Neal 1995; Rao and Ballard 1999). In short, the goal of an active inference agent is to behave in a Bayes optimal fashion by selecting actions that minimise free energy over time, and implicitly garner evidence for its own existence. This is sometimes referred to as self-evidencing (Hohwy, 2016).

Active inference reframes the problem of action and perception under the assumption that actions aim to minimise surprise. The agent is equipped with a generative model of observed outcomes that serves to assimilate available sensory evidence to form optimal predictions, via Bayesian belief updating. In other words, the agent is presented with sensory information (observations or

outcomes) and the generative model furnishes expectations — and ensuing predictions — about the unobserved causes (i.e., hidden states) of observed outcomes. This belief updating minimises variational free energy. These states generating outcomes are 'hidden' in that they are not observable and can only be inferred based on (usually limited and sparse) observations.

Variational free energy represents an upper bound on the negative logarithm of Bayesian model evidence (a.k.a., self information or surprise), so that minimising free energy, maximises model evidence (a.k.a., marginal likelihood) and implicitly reduces surprise. Learning, planning, and perception all minimise free energy. This minimisation can be read as updating model parameters, beliefs about policies and beliefs about hidden states, respectively. This results in a progressive improvement in the ability to infer hidden states given sensory information (i.e., perception), and the selection of the optimal behaviour (i.e., action), to minimise predicted surprise (Kaplan et al., 2018; Botvinick et al., 2012; Attias et al., 2003). Operationally, this means that the agent evaluates different policies (i.e., sequences of actions) in terms of their expected free energy when planning forward, and the policy with the least expected free energy is selected.. In other words, much of our behaviour is driven to resolve uncertainty about hidden states – e.g., "what caused that?" – or model parameters – e.g., "what would happen if I did that?".

Variational free energy is a functional of a probabilistic generative model and approximate posterior distribution over the hidden causes (e.g., states, policies, parameters, etc.). Free energy can be derived from Jensen's inequality (Parr and Friston, 2019), and expressed in different complementary ways (Friston et al., 2016):

$$\begin{aligned} Q(x) &= \arg min_{Q(x)} F \\ &\approx P(x|o) \end{aligned}$$

$$\begin{aligned} F &= E_Q[lnQ(x) - lnP(x,o)] \\ &= E_Q[lnQ(x) - lnP(x|o) - lnP(o)] \\ &= E_Q[lnQ(x) - lnP(o|x) - lnP(x)] \end{aligned}$$

$$= D[Q(x)\|P(x|o)] - lnP(o) \quad \textit{(upper bound on neg. log evidence)}$$

$$= D[Q(x)\|P(x)] - E_Q[lnP(o|x)] \text{ (complexity – accuracy)}$$

Where $Q(x)$ is the approximate posterior over the hidden causes x given observations o. Here, we have lumped together different sorts of hidden causes in $x = (s, \pi, \gamma)$, where s, π and γ correspond to states, policies and model parameters, respectively.

Free energy is an upper bound on surprise or negative log evidence. This can be seen clearly in the penultimate equation above, where the free energy is expressed in terms of negative model evidence and a KL-divergence term (which cannot be less than zero). The KL divergence expresses the dissimilarity between the approximate and the true posterior distributions over the hidden causes. Free energy minimisation suppresses the divergence between the approximate and true posterior distributions, which is why the former becomes an approximation to the true posterior. Free energy can also be expressed in terms of complexity and accuracy (see the last equation above). This means that minimising free energy is also equivalent to minimising the complexity of an accurate explanation for observed data. Note that negative free energy in physics is exactly the same as the free energy used in machine learning, where it is known as evidence lower bound or ELBO (Winn et al., 2005.; MacKay et al., 1995; Hinton et al., 1993)

Policies (trajectories of actions over time) are evaluated in terms of their expected free energies, where an action is sampled from the policy with the least expected free energy (Friston et al. 2016, Parr and Friston, 2018). The expected free energy can be decomposed into two main components, namely epistemic and extrinsic value. Epistemic value allows the agent to evaluate different policies in terms of how much information they can solicit from the world. This means that the agent will try to avoid the states that would produce uninformative or ambiguous outcomes (Schwartenbeck et al., 2013). Extrinsic value depends upon prior beliefs about future outcomes (i.e., prior preferences). These prior beliefs express how much one outcome is preferred relative to another, given the kind of agent in question. The more likely policies are those that generate preferred outcomes, such as positive primary

reinforcement. Expected free energy is derived from the variational free energy and can be expressed as follows (Mirza et al., 2018):

$$G(\pi) = \sum_{\tau} G(\pi, \tau)$$

$$G(\pi, \tau) \approx -\underbrace{\mathbb{E}_Q[\log Q(s_\tau|o_\tau, \pi) - \log Q(s_\tau|\pi)]}_{Epistemic\ value} - \underbrace{\mathbb{E}_Q[\log P(o_\tau)]}_{Extrinsic\ value}$$

Where $Q = P(o_\tau|s_\tau)Q(s_\tau|\pi)$ and $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$. Note that expected free energy is effectively the variational free energy expected under the predicted outcomes given a policy. Because these outcomes have yet to be observed, they become random variables. This has the important — if curious — consequence that minimising expected free energy entails *maximising* the expected KL divergence. This expected KL divergence corresponds to the information gained or epistemic value, sometimes known as intrinsic value in robotics (Barto et al., 2013; Schmidhuber et al., 2010; Oudeyer et al., 2007).

The Generative Model

Generative models express how observed outcomes are generated. In Markov decision process models of discrete states, this involves specifying some probability distributions that express the dynamics of the hidden causes (e.g., hidden state transitions) and the mapping between the hidden states and outcomes. Given some observed outcomes, perception corresponds to the inversion of the generative model to infer the hidden states that generated the outcomes. Similarly, the optimisation of beliefs about policies (from which actions are sampled) rests upon Bayesian beliefs about hidden states. Active inference can thus be interpreted as generalisation of planning as inference (Kaplan et al., 2018; Botvinick et al., 2012; Attias et al., 2003; Mirza et al., 2016). In summary, a generative model is a probabilistic description of how (unobservable) causes generate (observable) consequences, while model inversion recovers the causes from the consequences in a Bayes optimal fashion.

The generative model can be described in terms of the following factorisation:

$$P(\tilde{o}, \tilde{s}, \pi, \gamma) = P(\tilde{o}^1|\tilde{s}^1, \ldots, \tilde{s}^N) \ldots P(\tilde{o}^M|\tilde{s}^1, \ldots, \tilde{s}^N)\, P(\tilde{s}^1|\pi) \ldots P(\tilde{s}^N|\pi) P(\pi|\gamma) P(\gamma)$$

where $\tilde{o} = (o_1, \ldots, o_t)$ and $\tilde{s} = (s_1, \ldots, s_t)$ are observations and hidden states over time.

The generic structure of the ensuing generative model is based upon two main sets of matrices, which specify action-dependent transitions among hidden states and the mapping from hidden states to outcomes. In practice, the agent's beliefs about the state of the world and policies (as trajectories of actions) are updated or optimised at each unit time as new outcomes are sampled.

What follows is a brief explanation of the main components that specify the generative model. A^m is the likelihood matrix for the *m*-th outcome modality. This matrix expresses a probabilistic mapping from hidden states to the outcomes of the *m*-th modality. Different outcome modalities correspond to different kinds of observations, for instance different perceptual senses. The second set of matrices $B^n(\pi)$ expresses the transition probabilities among the hidden states of the *n*-th hidden state factor. Crucially, the agent has control over some of its states, and the transitions among those states depend upon the policy π (or action u = π(t)). Different kinds of hidden states represent different attributes of the world, which together generate the outcomes (for instance, *what* and *where* hidden states generate what would be observed when sampling a visual scene in which a particular object was in a particular location). Moreover, the model has prior beliefs over the initial state of the world D^n vector, and prior preferences over outcomes in the matrix C^m (Mirza et al., 2016). D^n encodes the prior beliefs about the initial states (n represents the dimension of the hidden states), while C^m represents the preferences over outcomes (m ∈ {1..number of outcome modalities}) .

As noted above, an agent is equipped with prior beliefs that it will pursue policies that minimise expected free energy. This can be seen in the final

equation below, which expresses the prior beliefs over the policies as a SoftMax function of (negative) expected free energy.

$$P(o_\tau|s_\tau) = Cat(\boldsymbol{A})$$
$$P(s_{\tau+1}|s_\tau, \pi) = Cat\left(\boldsymbol{B}\big(u = \pi(\tau)\big)\right)$$
$$P(o_\tau) = Cat(\boldsymbol{C}_\tau)$$
$$P(s_1) = Cat(\boldsymbol{D})$$
$$P(\pi|\gamma) = \sigma(-\gamma.\boldsymbol{G})$$

In this formulation, Cat represents the categorical distribution, γ is the inverse temperature (precision of beliefs about policy) and G is the expected free energy for each policy and sigma is a softmax function.

Learning

Both inference and learning can be cast as belief updating (about hidden states and parameters, respectively). The distinction between inference and learning is of key importance for the purpose of this work. While inference is the optimisation of Bayesian beliefs about hidden states, policies and precision, learning is expressed by the optimisation of model parameters (see Friston et al. 2016 for a full discussion). The generative model contains prior beliefs about its parameters. These are usually parametrised in terms of Dirichlet parameters (FitzGerald et al., 2015), namely, parameters that represent the number of co-occurrences of hidden states and related outcomes in the past (Friston et al. 2016, Parr 2019). The model includes prior beliefs about likelihood (a) and transition concentration parameter matrices (b). In this work, learning applies to the Dirichlet parameters of the likelihood concentration matrix (a). This can be formalised as follows:

(Prior beliefs)

$$P(\boldsymbol{A}|a) = Dir(a) = \begin{cases} E_{P(\boldsymbol{A}|a)}[\boldsymbol{A}_{ij}] = \dfrac{a_{ij}}{\sum_k a_{kj}} \\ E_{P(\boldsymbol{A}|a)}[ln\boldsymbol{A}_{ij}] = \psi(a_{ij}) - \psi(\sum_k a_{kj}) \end{cases}$$

(Learning)

$$\mathbf{a}^m = a^m + \sum_\tau o_t^m \otimes s_\tau^1 \otimes s_\tau^2 \ldots$$

Where ψ is a digamma (logarithmic derivative of a gamma) function. Here, *i* and *j* indicate the mapping between hidden states and outcomes where each entry of *a* is normalised (see top equation). Here, $\otimes$ indicates the Kronecker tensor product. The Dirichlet parameters are accumulated over time to assimilate new observations (increasing the parameter relating a particular outcome to an inferred state). This increases the parameters that encode the mapping between states and the sampled outcome. Strengthening the mapping between two variables if they manifest at the same time relates to Hebbian plasticity (Hebb, 1949). In short, this kind of learning corresponds to experience-dependent — or associative — synaptic plasticity.

Generative Model Specification

In this work, I use two versions of a generative model, which simulate memory for lists of words. This approach offers a different interpretation of attractor and recurrent neural network models (Botvinick and Plaut, 2006; Huh and Todorov, 2009; Maass et al., 2002; Martinet et al., 2011; Whittington et al., 2019): the attracting points represent the states that encode posterior beliefs about latent causes that minimise variational free energy (i.e., maximise marginal likelihood) (Friston, Lin et al., 2017a). The first version (*implicit* model one) performs the task using inference over hidden states in a working memory-like manner. The second version (*explicit* model two) includes learning via update of the Dirichlet parameters associated to the likelihood matrix A at the higher level of the model, to furnish an explicit encoding of stimuli in time.

The generative models used in this work are both hierarchical. Hierarchical (i.e., deep) models allow for deep temporal structure, where different levels of the hierarchy model state transitions over different time scales (slower for the higher levels) (George et al., 2009; Friston et al., 2017b). The specification of

hierarchical models is generally in terms of the A, B, C and D matrices at each level, where higher level states provide a *context* for the transitions among lower level states.

In our case, the connection between hierarchical levels means that the initial states at the lower level are caused by the hidden states of the level above. In other words, the priors over the initial hidden states $P(s_1^i)$ at lower level *i* is replaced by $P(s_1^i|s^{i+1})$. The ascending messages from the lower level correspond to the posterior beliefs about the initial hidden states. These enter the higher level as evidence for the contextual hidden state at the level above. In other words, the states at the higher level contextualise the initial states of lower levels. This means that a context specifies the beginning of a trajectory of state transitions at the lower level, thereby enabling a separation of temporal scales.

This also means that the agent has an explicit representation of the state of affairs at the beginning of a sequence of state transitions and at the end of that sequence (memory). Specifically, the setup of our models makes use of the hierarchical structure to reproduce the effect of learnt semantic associations between words on the DRM recognition memory task. As explained in the discussion, if we were not interested in the semantic effect (lure recognition), the same task could have been reproduced by using only one level of the models (in this case the higher level). Figure 2.1A2 shows the form of the hierarchical MDP used in this work.

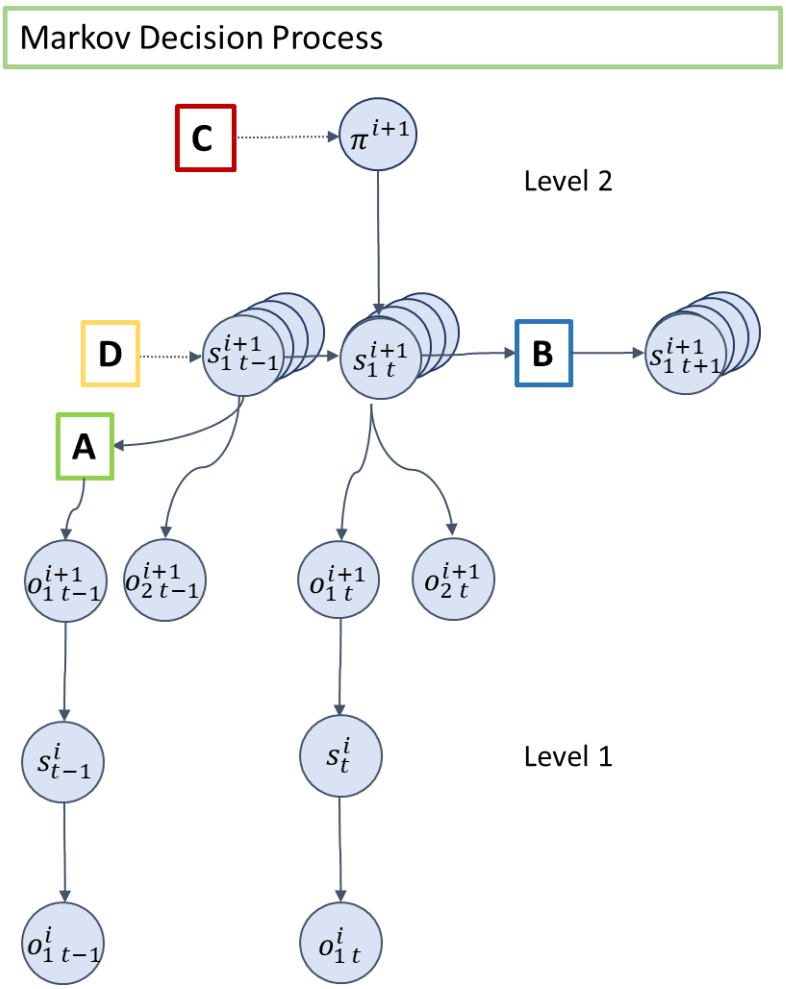


Figure 2.1A2. Scheme of Hierarchical Markov decision process. A represents the likelihood matrix, B the transition matrix, D the prior believes over initial states, and C the preference over outcomes.

The DRM paradigm involves two successive phases, namely encoding and retrieval. In the encoding phase, the agent is presented with a list of four words. In this phase, the agent can only wait and observe the words. In the retrieval phase, one word is presented, and the agent has to report whether she saw the word before or not ('yes' or 'no' choice). Crucially, I presented three lists, two of them contain semantically related words (cold-related words for the first list, and sleep-related words for the second); the third list contained words that do not have any semantic relationship. The agent is tested with three kinds of words: words that were presented before (old), words that were not presented before (new), and words that were not presented before but that belong to the same semantic group of words as in the list ('lure' words, to test

for false memory effect). The two models share the same structure of the lower level (level 1) but are structurally different at the higher level (level 2).

The two levels are linked via the ‘word’ outcome modality at the second level and ‘word’ hidden state factor at the first level (see Figure 2.1A2). The word outcome on level 2 determines the initial word (hidden) state on level 1. The expectations about the outcomes (word) at the higher level is passed down to the lower level as empirical priors over the hidden state ‘word’. Posterior beliefs about the hidden state ‘word’ at the lower level are passed upwards to the higher level as evidence for the (semantic) context in play.

Level 1

The lower level of the model allows us to simulate the semantic effect responsible for the false memory effect, or intrusions. The lower level involves a single outcome modality and hidden state factor, namely word outcomes and hidden states. The word state factor has nineteen levels: null (for fixation cross), four words for each of the three list, two lure words (one for each semantically related list), four words not presented in any list, used in testing (see Figure 2.2A2 left panel). The likelihood matrix A maps the word states to the word outcomes. In the generative process, there is an identity mapping between word states and outcomes (see Figure 2.2A2 top right panel). In the generative model, the prior likelihood concentration parameters entail a form of semantic knowledge, i.e., pre-learnt associations between words (Figure 2.2A2 bottom right panel).

Semantic knowledge is implemented by assigning non-zero Dirichlet parameters to semantically related words and assigning relatively higher parameters to the diagonal entries. This is conceptually equivalent to the semantic associations matrices in the Hopfield network (Chapter 2.2) and TCM models (Chapter 2.3). For instance, in the generative model, the most likely outcome under the state ‘rest’ is again the word ‘rest’, but the semantically related words ‘sleep’, ‘dream’, ‘nap’ and ‘bed’ also have non-zero concentration parameters. While the identity mapping applies to any word,

only the words in lists *cold* and *sleep* (i.e., lists containing semantically related words), including their respective lure words (false memory) had non-zero concentration parameters (see Figure 2.2A2 bottom right panel). The concentration parameters are converted to probabilities by normalising each column of the likelihood matrix. At this level, the same word is observed over successive time points (at most five) until a certain threshold is achieved in terms of confidence in inferences about the word states. The successive time points at the lower-level maps onto a single time step at the higher level. We will later use the number of time points — before reaching the confidence threshold at the lower level — as reaction time.

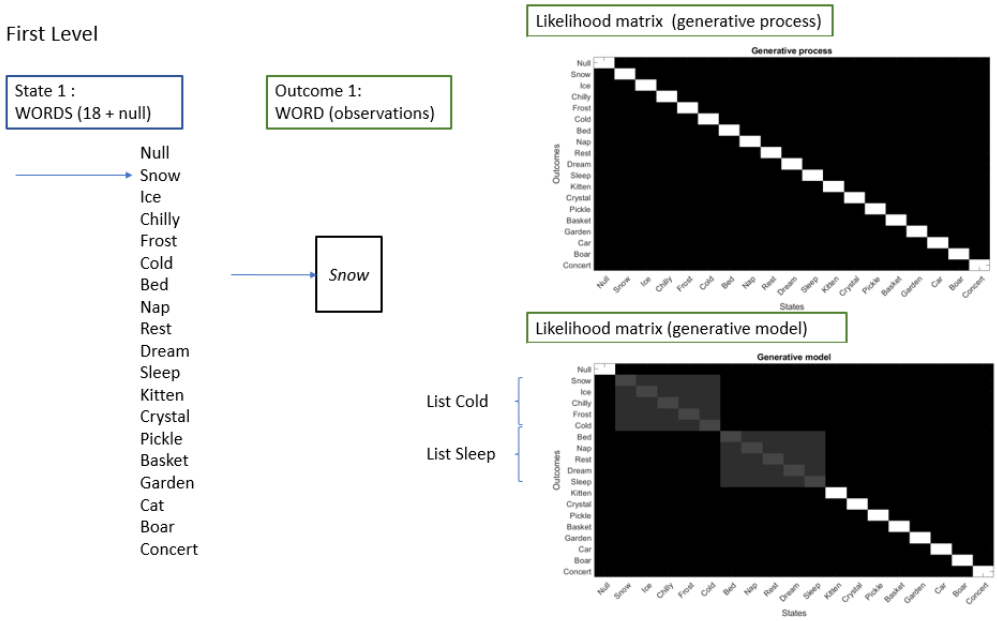


Figure 2.2A2. Level 1 of Model 1 and Model 2. The panels on the right illustrate normalised likelihood Dirichlet parameters. Each column of the matrix sums up to 1.

Level 2

The higher level of the hierarchy is responsible for the heavy lifting of the implicit mnemonics. Here, we describe the higher level of model 1 (working memory-like model) and model 2 (with explicit encoding of temporal context) separately. The outcomes of the second level specify the initial states of the first level. In brief, the agent observes four words (encoding), is then presented

with a test word (probe) and asked whether she saw the word during the encoding or not (‘wait’, ‘yes’ or ‘no’). After the answer, the agent receives feedback.

Implicit Model 1

The second level contains four hidden state factors and two outcome modalities (see Figure 2.3A2). The task starts with a fixation cross, represented as a ‘null word’ in the model. The first hidden state factor ‘list’ represents the list presented at the encoding phase and it has three levels (i.e., three possible lists). The second hidden state ‘When’ represents time and it has eight levels, each corresponding to one time step in the task (null – word – word – word – word – null – probe word – null). The third hidden state ‘Testing Words’ indicates the word (probe) that is presented for testing after encoding (eighteen levels, namely all the words but the null word). The fourth hidden state ‘Decision’ maps the choice of the agent. This has three levels, ‘wait’, ‘yes’, ‘no’, where the agent is encouraged to wait during encoding but can make a choice — either yes (i.e. ’the word was in the list’) or no (i.e. ‘the word was not in the list’) — after being presented with the probe word (time step seven).

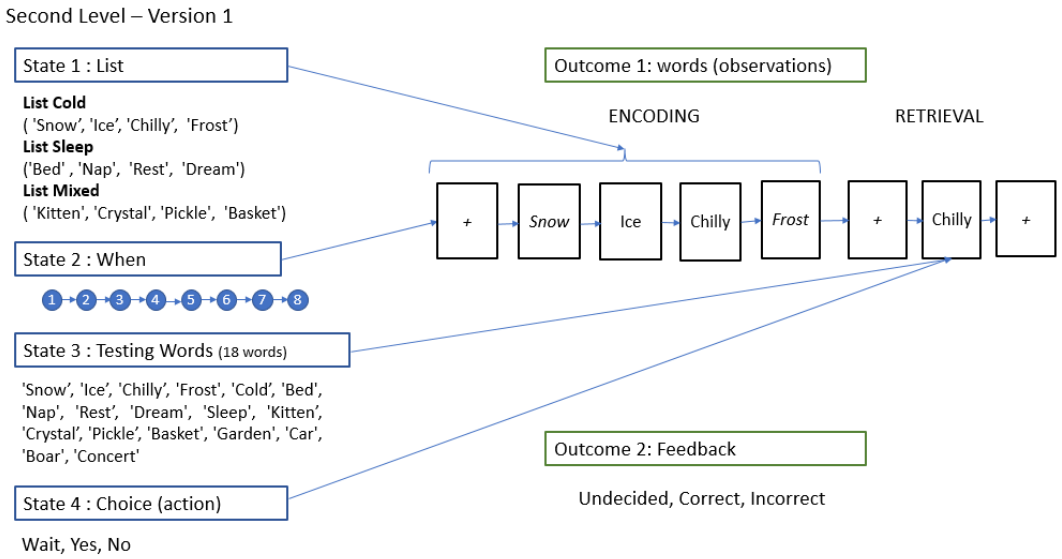


Figure 2.3A2. Level 2 of model 1.

There were two outcome modalities: word (o^1) and feedback (o^2). The first outcome modality (word) has nineteen levels (4 words for each of the three lists + 2 lure words + 4 new unrelated words + null word), while the second outcome modality has three levels ('undecided', 'correct', 'incorrect'). The generative model likelihood matrix for the *word* modality was such that the words associated with each list state (e.g., snow, ice, chilly, frost, cold for the first list) and the new unrelated words were assigned equal probabilities while the other words had a probability of zero for the time points associated with the encoding phase (except the first time point). This allowed the model to infer the list state upon observing a word in the encoding phase. There was an identity mapping from the *testing word* states to the probe words at the seventh time point in the retrieval phase. Finally, there was a precise mapping to the null words when a 'null' word was presented.

We assigned positive and negative log probabilities for the correct and incorrect outcomes under the feedback modality to define the model's prior preferences over the outcomes. The agent received 'correct' feedback if it answered 'yes' when presented with a previously seen (i.e., from the encoding

phase) probe word, ‘no’ for words that were not previously seen. The feedback was ‘incorrect’ if the agent answered ‘no’ when tested with a previously seen probe word, ‘yes’ when presented with new words (including lure words). If the agent chose to ‘wait’ after seeing the probe, it received ‘undecided’ feedback. We also encouraged the agent to choose ‘wait’ in the encoding phase by giving ‘incorrect’ feedback to any action but ‘wait’ in the first five time points.

The only factor under active control is the fourth factor (‘Decision’). There are three policies over the progression of the task: ‘wait’ at the first six-time steps (encoding), followed by either the answer ‘yes’ or ‘no’. The transitions among the list and testing word states are precluded (i.e. identity matrices), and the progression of time steps is pre-determined — from one time step to the next. The word outcomes are generated in the following way: the list states generated the words in the first five time points (i.e., encoding phase). With reference to the onset of the retrieval phase, the list states can only generate the previously seen words (e.g., snow, chilly, etc.). On the sixth time point, a new fixation cross (i.e., null word) is presented to cue the onset of the retrieval phase. At the seventh time point the *testing word* states determined the probe word. This meant that any word from any list can be tested (except the null word, i.e. fixation cross) and the choice of action depends upon the combination of list (which defines the previously seen words) and testing word (probe) states. If the agent believed that the probe word was part of the list, it would choose ‘yes’, to realise its preferred outcomes, otherwise it would choose ‘no’. Note that the hidden states at the higher level (i.e., list, and testing words) determined the word states at the lower level. The implicit separation of temporal scales means that evidence for states of a contextual sort at the second level can be accumulated over time, given the evidence from the lower level.

The behaviour under this generative model can be summarised as follows: in the encoding phase, the agent accumulates evidence for words drawn from a given list at the lower level of the hierarchy and infers the list state at the higher level. At the lower level, the concentration parameters associated with the observed words (diagonal entries) increase to a larger extent compared to the

semantically related words as each word is correctly inferred. When tested with the probe word in the retrieval phase, the agent infers whether the word was part of the list presented at the encoding or not. If she infers that the probe was part of the same list she saw during the encoding, she will answer ‘yes’, otherwise she would answer ‘no’. In other words, this memory retrieval is simply an action that reports the inference that her experiences are consistent with her memory of the recent episode (list).

In short, this account of episodic memory rests upon short term plasticity modelled in terms of accumulating Dirichlet parameters during recent perceptual processing. The *implicit* increase in the precision or confidence — when inferring the cause of a subsequently presented word — enables the agent to infer she has seen that word recently.

When presented with a probe word, the inference over the *previously seen* words is more precise than the *lure* words. This is because the concentration parameters associated with the *previously seen* words are updated in the encoding phase at the lower level of the model, leading to greater confidence in the inferred word. In contrast, the inference over the *lure* word is less precise because the evidence for the word states is distributed among the semantically related words to a larger extent. The model makes precise inferences over the *new unrelated words* as they are not semantically related to any other words. We therefore expected the model to be least confident about its responses, in terms of inferred policies, when a lure word was presented.

Explicit Model 2

This version of the model equips the agent with working memory by learning at *both* levels. Here, the representation of the past plays a different role from the first model. The ‘when’ hidden state can be thought of as generating outcomes in some mnemonic space — in the same way in which representations of spatial location underwrite scene construction. In other words, we model episodic memory as a form of scene construction over time, to generate outcomes from latent (i.e., hidden) *what* and *when* causes (Friston

et al., 2016).

We imagine that a short temporal scene (narrative or episode) can be constructed by associating unique content with a specific point in time. If the association of ‘what’ and ‘when’ is unique, we can infer when something occurred just by knowing what occurred. This simple observation is the basis of episodic memory in this explicit model. Focusing only on the recent past, we can encode all the combinations of ‘what’ and ‘when’ presented during the encoding phase. The co-occurrences of what happened when can be ‘remembered’ by accumulating co-occurrences. Computationally, this corresponds to the accumulation of Dirichlet parameters in the model, which models experience-dependent learning or associative plasticity. This means that the model includes an explicit representation of where in time an event occurred.

We initialised the likelihood concentration matrices with uniform and small concentration parameters (except for when a null word is presented). This means the agent expected any word at any time point. In this setup, the first eight time points generated the words presented in the encoding phase, whereas the last three time points generated the words presented in the retrieval trials, which could be old, new or lure words.

The problem of deciding whether something has been seen in the past can be reframed as inferring when something was observed. In this setting, the agent is presented with a sequence of stimuli. In the encoding phase, the learning happens via the accumulation of Dirichlet parameters over the likelihood mapping, based on the posterior beliefs about the ‘time’ states and the observed word outcome. For example, if the word ‘chill’ was presented at the second time point, the concentration parameters associated with the second level of the time factor and the word ‘chill’ in the first modality would increase in the likelihood mapping. In the retrieval phase, the role of the time factor changes such that the agent infers when a word was presented to discern whether it saw the word before, based on the learned likelihood mapping in the encoding phase. In other words, in the encoding phase, the agent infers the time and learns when a word was presented whereas in the retrieval

phase, the agent only infers when a word was presented. The agent will refuse the probe word if the inferred time is not associated with any of the precedent time steps. See Figure 2.4A2 for the structure of this MDP model.

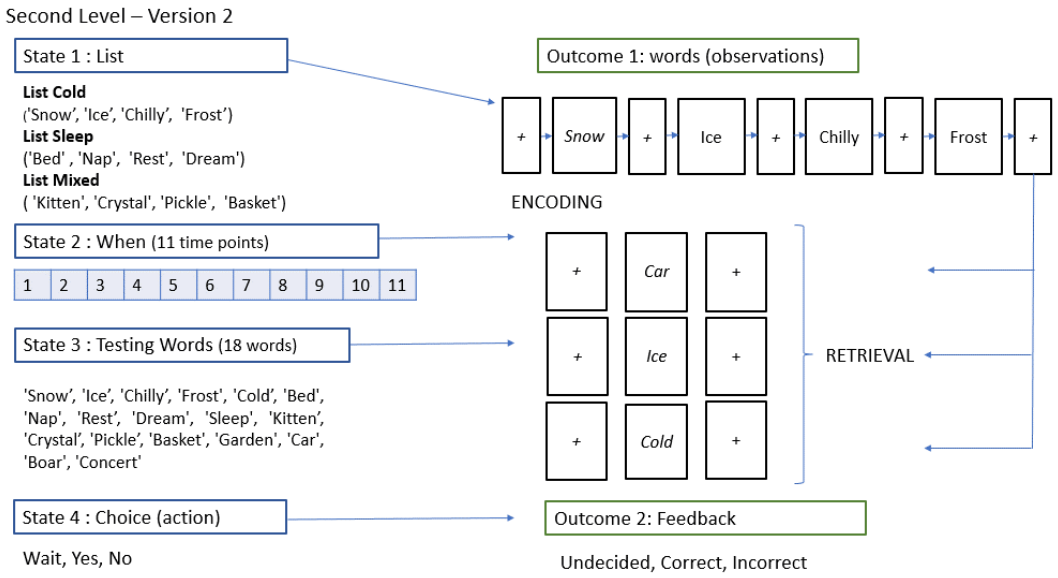


Figure 2.4A2. Level 2 of model 2. Although the choice of hidden factors at the higher level of this model might seem equivalent to the previous version, the mechanics of the task differs in the role of the second hidden state through the update of the model's Dirichlet's parameters. The higher level of this model has the same two outcome modalities ('word' and 'feedback'), and four hidden states as the first model. The first, third and the fourth hidden state factors, 'list', 'testing words' (probes) and 'choice' respectively, play the same role as in model 1. However, this model learns when a word was presented by updating its Dirichlet parameters in the likelihood mapping from the time or temporal encoding factor to the word outcomes. In this model, the task was split into encoding and retrieval trials. The encoding trial comprised eight time steps (null –word – null – word – null – word – null – word). After the encoding trial, the agent is presented with a retrieval trial (null – word – null). Each retrieval trial is made of three time steps (null-testing word - null). Crucially, in this model the agent was allowed to update its likelihood (over A matrix) concentration parameters at the end of the encoding phase; however, these updates were precluded at the end of the retrieval trials.

There is a key difference between this model and the previous model. In the explicit model 2, accumulation of Dirichlet parameters was allowed at the end of the encoding trial at the higher level of the model. This changes the way the 'time' hidden states are interpreted, such that model 2 allows (simulated) experience-dependent synaptic plasticity in the form of parameter updates. This means that the Dirichlet parameters — mapping from 'time' states to 'word' outcome — are updated, and the model learns the temporal order of the presented words. The agent starts the retrieval trials with the likelihood concentration parameters that are updated at the end of the encoding trial, but not after each retrieval trial. Therefore, the order of the retrieval trials does not affect the answer of the agent.

In summary, this simple generative model simulates retrieval in a recognition memory task via accumulation of Dirichlet parameters in the likelihood mapping (i.e., experience-dependent learning). For every combination of hidden states and outcome, a Dirichlet count is added to the appropriate entry in the likelihood mapping. By starting each trial with a small number of (*a priori*) Dirichlet parameters, any conjunction of outcomes and combinations of hidden states are accumulated, leading to a form of declarative memory (Parr et al., 2020)

An interesting question now arises: namely, what would happen if we presented a lure from the same category as the recent episode, which was not a previously encountered exemplar? (i.e., the word 'cold' for list cold and the word 'sleep' for list sleep). Having specified the generative models, we are now able to use standard message passing schemes to simulate perceptual inference, learning and choice behaviour to dissect the computational anatomy of the DRM paradigm and answer this question, with and without an explicit encoding of time.

Results

We first tested whether both models could successfully complete the recognition task, i.e. if they were both able to recognise the previously seen words in the retrieval phase and reject new words. We then tested for an effect of false memory (induced by the DRM task). Note that misremembering — although not correct in terms of behavioural performance — is still Bayes optimal: the use of pre-learned associations (i.e., semantic memory) might simplify the brain's model of the world (i.e., maximise model evidence by minimising the complexity in the above equations). In this sense, learning patterns of stimuli can result in a coarse-grained explanation for stimuli, so that the number of hidden states required by the system for successful encoding might be smaller than the number of observable outcomes. This means that more than one observation could be sufficiently explained by the same cause. If a stimulus is similar enough (in this case, semantically related) to the stimuli presented in the list at the encoding, the agent would then recognise the stimulus as already seen.

The false memory effect associated with the lure rests upon prior beliefs about the way words are generated (in relation to other words). As such, a parallel can be drawn between false memories and illusions: both are the most plausible *a posteriori* inferences based on prior beliefs, although not veridical causal explanations for observations. This leads to the interesting notion that a false memory is a mnemonic illusion, and, in principle, all our memories are illusory to a greater or lesser extent. In other words, using this kind of generative model means that retrieval corresponds to selecting an episode that has the greatest posterior probability of accounting for current experience. This places both prior beliefs of a semantic nature and time sensitive (episodic) priors at the heart of retrieval. In what follows, we review the results of numerical simulations to verify these hypotheses and illustrate the different behaviours that unfold under different prior beliefs.

When comparing the synthetic electrophysiological and psychophysical responses for the two models, some subtle differences were apparent.

Although both formulations were able to reproduce the basic phenomenology of the DRM paradigm, the explicit model was able to reproduce empirical reaction time results more faithfully, with some clear, disambiguating, electrophysiological correlates. In more detail, while the implicit model bases its electrophysiological activity on coded list recognition, the explicit model's electrophysiological responses mirror biologically plausible time-to-item learnt associations. In other words, the explicit model synthetic electrophysiological activity reflects lower confidence on the episodic embedding of the lure word at the time when the learnt list was presented.

Implicit Model 1

The first model simulates the DRM task using a working memory-like process, which does not entail learning about timing at the higher level. The model recognises whether a word was presented in the list at the encoding trial, or not, based on inference about the list at the higher level and learning of the presented words at the lower level. We also tested for an effect of lure words. The recognition of lure words 'cold' and 'sleep' recapitulates a false memory effect with a mean recognition rate of about 60% (Figure 2.5A2).

In our simulations, misremembering (false memory effect) is expressed by the model answering 'yes' to words that were not presented at encoding but were semantically related to the words in the list (lure word). Correct recognition indicates that the agent answered 'yes' to probe words that were part of the list, while correct rejection indicates answer 'no' to new words, not present in the list at encoding. Misremembering is licensed by the growing expectation that the semantically related words in the same list cause the lure word. The list with semantically unrelated words' did not elicit any false memory effect, as expected. We ran 32 trials at recognition for each list and type of probe word (i.e., old, new and lure). We ran the same simulation and tested for false memory at retrieval.

We also tested whether our model could reproduce the longer reaction times for false memory, found in experimental work (Figure 2.5A2 C). Model 1 was

not able to reproduce the longer reaction time for false memory, showing equivalent reaction times for all three types of recognition trials.

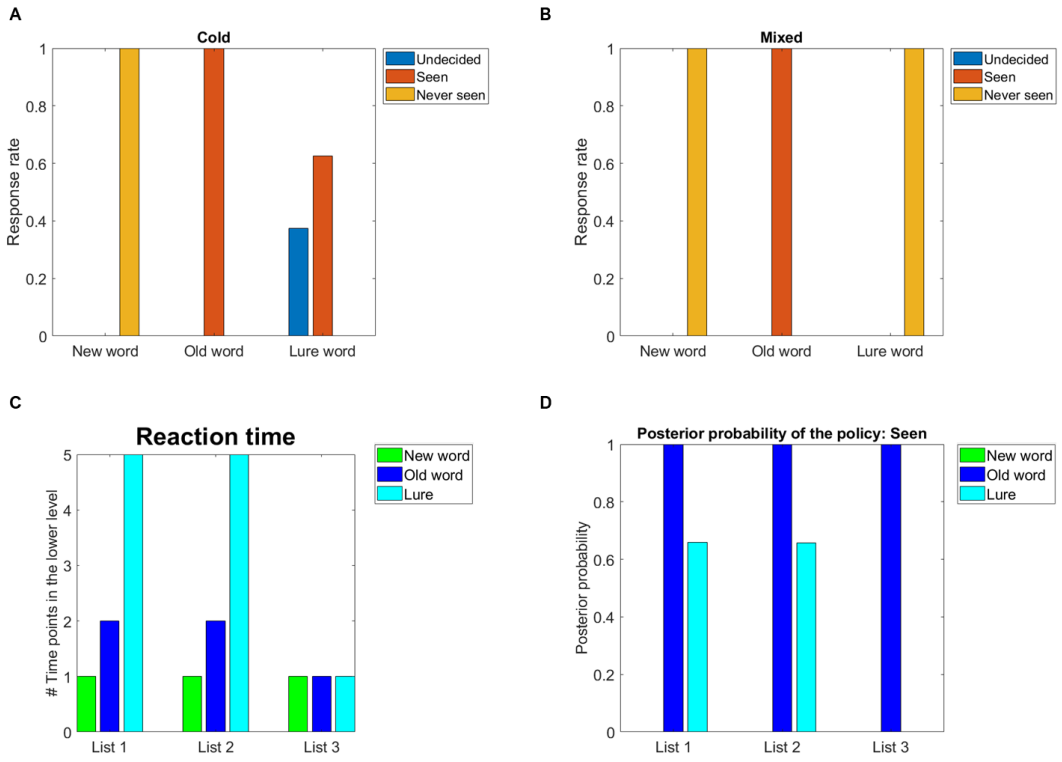


Figure 2.5A2. Simulated behavioural responses for implicit Model 1. A) Average behavioural responses (i.e., whether the agent answers 'yes', recognising the word as seen before, 'no' for never seen before words, or undecided) for probes 'New word', 'Old word' (i.e. words presented in the list at encoding) and 'Lure word (i.e. false memory) in list 1 (cold-related). The average is calculated over 32 trials. In these simulations, the value of the context-related prior was set to 1, while the Dirichlet parameters for the identity mapping was set to 2 in the lower level likelihood matrix. The new word and old context words are reported as never seen and seen 100% of the trials, respectively. The lure (new context) word is reported as seen before on 63% of the trials. B) Average behavioural responses for list 3 (mixed word list). In this case, the New context word is not semantically related to the words in the list, and the performance is indeed consistent with the one seen for new words. C) This panel shows the reaction times for the three probe words are the same. The reaction time was calculated as the number of time points required at the lower level of the model to minimise uncertainty over states (please see software note). The simulations at the lower level were terminated once the uncertainty about the word states was

minimised. This was implemented by computing the entropy of the posterior distribution over the word states at each unit time and terminating the simulations once the entropy is smaller than a pre-determined confidence threshold (chi parameter set to 1.5). D) These panels show the posterior probability of the policy 'Yes' (i.e., 'Seen') when three kinds of probe words are presented for each list. The posterior probability of the policy 'Yes' is 1 for old words and 0 for new words in all lists. The posterior probability of the policy 'Yes' is non-zero only when a new word is semantically related to the lists (i.e., list 1 and 2, where the false memory effect is reproduced).

The rate of false memory can be modulated by changing the value of the context-related Dirichlet parameters relative to the strength of the diagonal (identity) concentration parameter in the likelihood mapping: the stronger the association between semantically related words (context-related prior), the higher the rate of false memory (Figure 2.6A2). In this case, inference entails recognising the specific list presented during encoding, and on the identity of the word tested as the probe, which could be part, or not, of the presented list. The relative strength of the word identity (diagonal) vs associated words (context-related) Dirichlet parameters over the likelihood matrix modulates how shallow the attractor states afforded by the likelihood matrix are, i.e., the higher the context-related parameters, the easier it is for the model to mistakenly fall into an associated word state and recognise (infer) the lure word as part of the presented list.

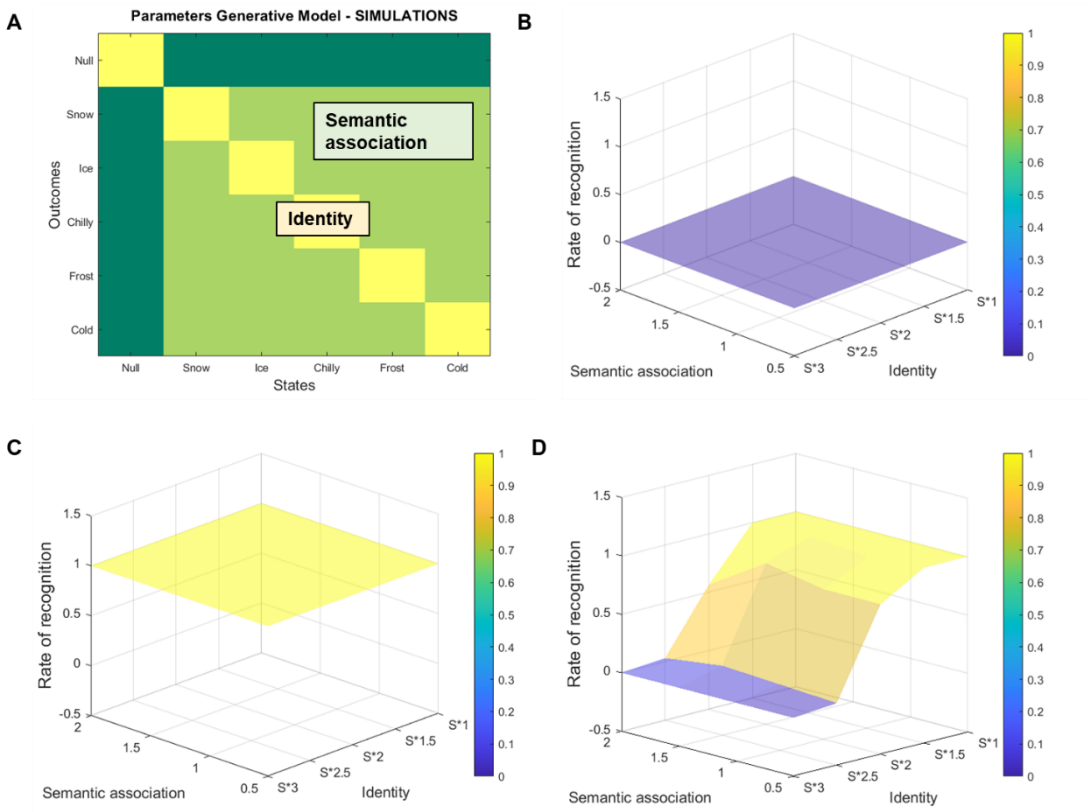


Figure 2.6A2. (A) Visual representation of the parameters' change in the simulation for list 'cold'. The 'identity' parameter modulates the strength of the diagonal in the generative model, while the 'Semantic association' parameter modulates the Dirichlet parameter strength for semantically associated words. Rate of recognition ('Yes' answer) for new words (B), old words (C) and lure word (false memory) (D) for list 1 (cold-related). The recognition rate is shown as functions of the context-related prior (off diagonal non-zero entries) and of the identity prior (diagonal entries) of the likelihood matrix at lower level. The identity prior is changed proportionally to the context-related prior: context-related prior ×1 (S*1), ×1.5 (S*1.5), ×2 (S*2), ×2.5 (S*2.5) and ×3 (S*3). The context-related prior took values: [0.5 1 1.5 2]. Rates of recognition for old and new words (B and C) was not affected by the change of priors. However, the rate of recognition for lure word (false memory) indeed showed a dramatic change (D). In particular, the relative strength of word identity prior vs (semantic) context-related prior modulates the rate of false recognition: the bigger the ratio between identity prior and context prior, the lower the rate of false memory.

Finally, we simulated electrophysiological responses (LFPs). The basic assumption is that depolarisation of distinct neuronal populations reflects belief updating about hidden states when new evidence (i.e., a stimulus) is presented. These simulated neuronal activities are used to produce simulated

LFPs (Friston, FitzGerald et al., 2017). Technically, these synthetic neuronal responses correspond to a gradient flow on variational free energy, which underscores the neuronal plausibility of the belief updating scheme used to simulate active inference. Each trial comprises eight epochs (5 for encoding and 3 for testing). The amplitude of the simulated LFPs is associated with the extent of belief updates, c.f., prediction errors over hidden states, where technically, free energy gradients correspond to prediction errors.

Convergence time reflects the time needed for new posterior beliefs to be formed, when the LFPs settle to zero: i.e., free energy is minimised such that free energy gradients are zero at the minima of free energy attractors. Higher peaks and faster convergence times result from more precise beliefs. In Figure 2.7A2, we report exemplar simulated LFPs for the three different kinds of probes (new words, old words and lure word). These simulations are run for 'cold'-related list, with identity parameter set to 2 and semantic association parameter set to 1, as in Figure 2.5A2.

Interestingly, the LFPs for new words yield higher peaks, reflecting greater confidence. Old word and false memory have similar amplitude peaks. However, the false memory probe shows multiple word-associated responses with relatively higher peaks. The word-related curves converge slower for the lure word compared to the old words, with peaks at later time points for the lure word. The slower convergence for the lure words compared to the old context word might be due to the increased Dirichlet parameters (i.e., increase in synaptic efficacy) — at the lower level of the model — at the end of the encoding phase for old words compared to lure word at the time point when the probe is presented. For semantically related words — presented at encoding (not for the lure word) — the model has a more precise mapping compared to the lure word at the lower level.

This reflects an inherently different kind of simulated neuronal behaviour that distinguishes between correct responses for old words and false memory (incorrect recognition for lure word). Namely, the model infers that the lure word was indeed part of the presented list, but this is driven via accumulated evidence for semantically associated words (i.e. the context-associated priors

affect the hidden state mapping so that an outcome can be explained — although less strongly — by multiple hidden states representing associated words). During encoding, the Dirichlet parameters associated with the presented words (old words) increase in the lower level of the model, leading to a more precise mapping from word states to outcomes. This is not the case with the lure word.

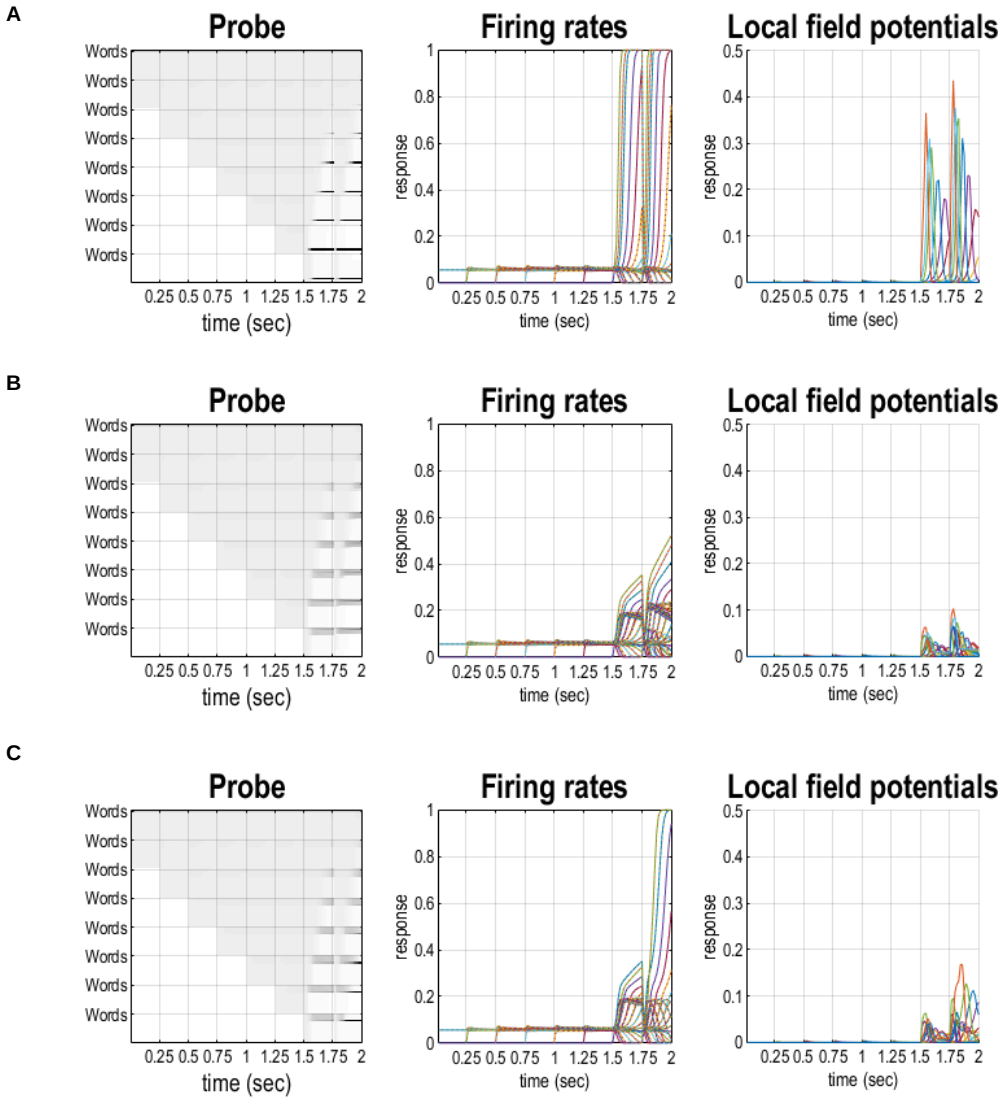


Figure 2.7A2. Simulated electrophysiological responses Model 1, for new word probes (A), old word probes (B) and false memory probe (C). In particular, the relevant hidden state responsible for the model's answer is the recognition of the probe word, as part of the presented list, or not. Column 1 (left) shows the expectation about hidden states at each unit time and epoch. An epoch represents a time that could be in the past, present or the future and these are shown on the y-axis. The above

diagonal entries represent the beliefs about the states in the past epochs, while the below diagonal entries represent the beliefs in the future epochs compared to the diagonal entries which represent the beliefs at the present. In this panel, each cell shows the way the posterior beliefs about a set of 18 words change during gradient descent for a given epoch and time point. The x-axis shows the progression of time as the agent makes new observations. The central column represents the firing rates associated with each possible hidden state (word) entry, i.e., each word from each epoch is assigned a different colour and line. Finally, the right column LFP is simulated as the rate of change in beliefs on hidden states; in other words, as the gradient of the firing rate curves (central column). The rate of belief update for new word (A) shows higher peaks compared to both old words (B) and false memory (C). This reflect higher precision in expectation when the word presented as probe was not part of the list at encoding. Interestingly, false memory (C) has higher firing rate (central column) than old word probes (B). However, in case of false memory, the local field potentials (right column) appear to have a slower convergence implying, possibly, a lower level of confidence in beliefs.

Explicit Model 2

As with the first model, the second series of simulations reproduces the behavioural findings from the DRM task. This model's responses depend on its beliefs about the mapping from the *temporal point* (i.e., time) states to *word* outcomes. The presented words during the encoding (old words) are strongly associated with an earlier moment in time, compared to the word presented at retrieval only (new words). The presence of a hidden state — that reflects a form of biologically plausible temporal context (Umbach et al., 2020) — is the key element that allows the model to perform the recognition task, depending on whether the probe word is most likely to be associated with one of the first eight time steps which encodes the previously seen words or the last three (only seen at retrieval, as new word).

This model updates the Dirichlet parameters of the likelihood mapping from time states to word outcomes at the end of the encoding phase. As visualised in Figure 2.8A2, the model learns the time points when the words from the list 'cold' are presented. The model starts the testing phase with updated

concentration parameters. The presence of a delay period after encoding (null word at $t = 6$) shows that the agent can retain information about the first part of the task that underwrites her answer to the probe word.

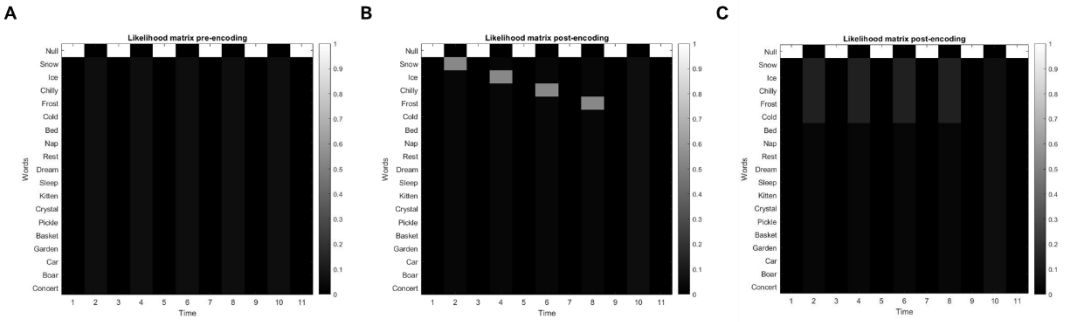


Figure 2.8A2. Dirichlet parameters associated with the likelihood mapping from time states to word outcomes at the higher level. A) This matrix represents the Dirichlet parameters of the likelihood mapping before the trial starts. The mapping from time states to word outcomes are initialised with small but uniform concentration parameters except for the time points when the null words are presented. The model was equipped with precise beliefs that there would be null words (fixation cross) between words (i.e., high concentration parameters). This was introduced only to resemble the typical structure of the DRM paradigm and to test whether the model could reproduce the behavioural responses even with the presence of time delays between learnt stimuli. B) An example of a learned likelihood mapping at the end of the encoding phase for weak semantic association parameter (0.5). In this case, the temporal order of the presented words is clearly encoded. C) Another example of the learned likelihood mapping at the end of the encoding phase, with a stronger semantic association parameter (2). The model has stronger expectations that the four words presented in the list and the lure word (i.e., cold) were indeed associated with the encoding time points of the trial. After learning, the model starts the testing phase with the updated Dirichlet parameters (B and C) to discern whether a word was seen before or not.

We ran simulations of 32 trials for each kind of tested probe for each of the 3 lists. Examples of the behavioural responses are summarised in Figure 2.9A2.

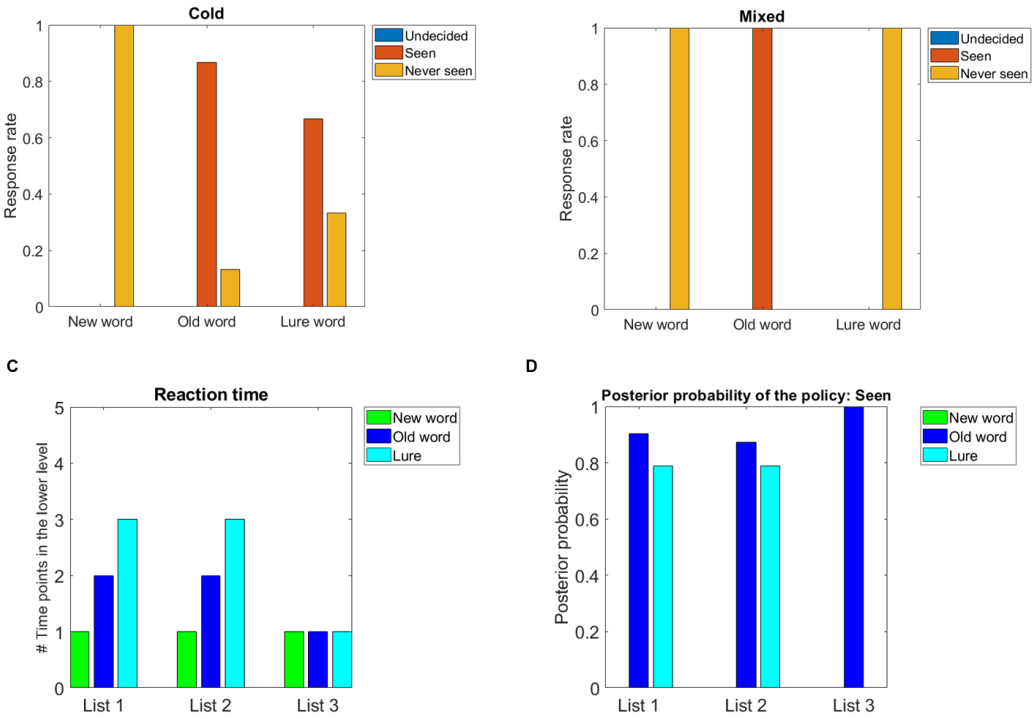


Figure 2.9A2. Simulated behavioural responses for Model 1. A) Average behavioural responses for probes 'New word', 'Old context word' (i.e. words presented in the list at encoding) and 'New context word' (i.e. lure word, or false memory) in list 1 (cold-related). The average is calculated over 32 trials. In these simulations, the value of the context-related prior was set to 2, while the identity prior was set to 2.5. The model has a 100% accuracy for new words, 85% accuracy for old words, while it recognised the lure word as seen before in 65% of trials. B) Average behavioural responses for list 3 (mixed word list). In this case, the New context word is not semantically related to the words in the list, and the performance is consistent with the new words. C) Reaction times for the lure words were longer than the old context words, which were longer than the new words for list 1 and 2 (cold- and sleep-related), while they were the same for all probe words in list 3. D) This panel shows the posterior probability of the policy 'Seen' (i.e. 'Yes') for three kinds of probes for list 1 (cold-related), list 2 (sleep-related) and list 3 (mix list). The posterior probability of the policy 'Seen' was 0.9 for the old words in semantically coherent lists, and 1 for old words in the mixed list, while it was 0 for the new words in all lists. The posterior probability of 'Seen' was about 0.65 when presented with the lure words from lists 1 and 2 (i.e., lists with semantically related words) and 0 for list 3 (semantically unrelated words).

To simulate how both learning rate and relative strengths of the identity and context-related prior Dirichlet parameters affect the simulated responses of the model (Figure 2.10A2), we set the context-related prior to values [0.5 1 1.5 2], and

modulate the identity prior proportionally to the context-related prior: context-related prior ×1 (S*1), ×1.5 (S*1.5), ×2(S*2), ×2.5 (S*2.5) and ×3 (S*3), as per model 1.

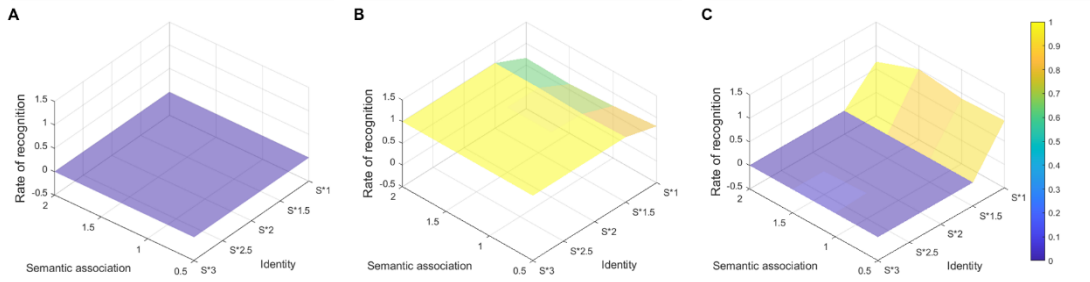


Figure 2.10A2. Rate of recognition ('Yes' answer) for new words (A), old words (B) and lure words (false memory, C) for list 1 (cold-related). The recognition rate is shown as a function of the context-related prior (off diagonal non-zero entries) and of the identity prior (diagonal entries) of the likelihood matrix at the lower level. The identity prior is changed proportionally to the context-related prior: context-related prior x1 (S*1), ×1.5 (S*1.5), ×2 (S*2), ×2.5 (S*2.5) and ×3 (S*3). The context-related prior Dirichlet parameter (semantic association) took values of [0.5 1 1.5 2]. Rates of recognition for old words (A) did not change as a function of the priors. However, the rate of recognition for the old and lure words changed (B and C), depending on both parameters described above. A weaker difference between identity prior and semantic associations with other words (see S*1) reduced and increased the rate of recognition for old words (B) and lure words (C), respectively.

Finally, we simulated electrophysiological responses (Figure 2.11A2). The agent has precise beliefs that the new word was presented in the testing phase (top left panel). The agent believes that the old word was presented both in the encoding and testing phases, but the combined probability associated with the encoding phase outweighs the testing phase (middle left panel). In the case of the lure word, there is a considerable probability mass associated with the time points when a semantically relevant word was presented in the encoding phase, and a substantial probability mass at the time point when the lure word was presented in the testing phase (bottom left panel).

Focusing on the encoding phase, the model specifically believes that the old word was presented at the fourth time point. In contrast, there are roughly equal probabilities assigned to the time points when semantically relevant

words were presented when testing with the lure word. This shows that the model recollects the episode when the old word was presented but there is no unique episode with which the model can associate the lure word. The firing rates for the new word (A) shows higher peaks between $0.5 - 1\ (s)$ compared to both old words (B) and false memory (C), expressing higher confidence about when the word was presented (compare centre panels). Future studies employing brain imaging techniques could potentially test the validity of the simulated responses and the subtle differences between probe words as shown here.

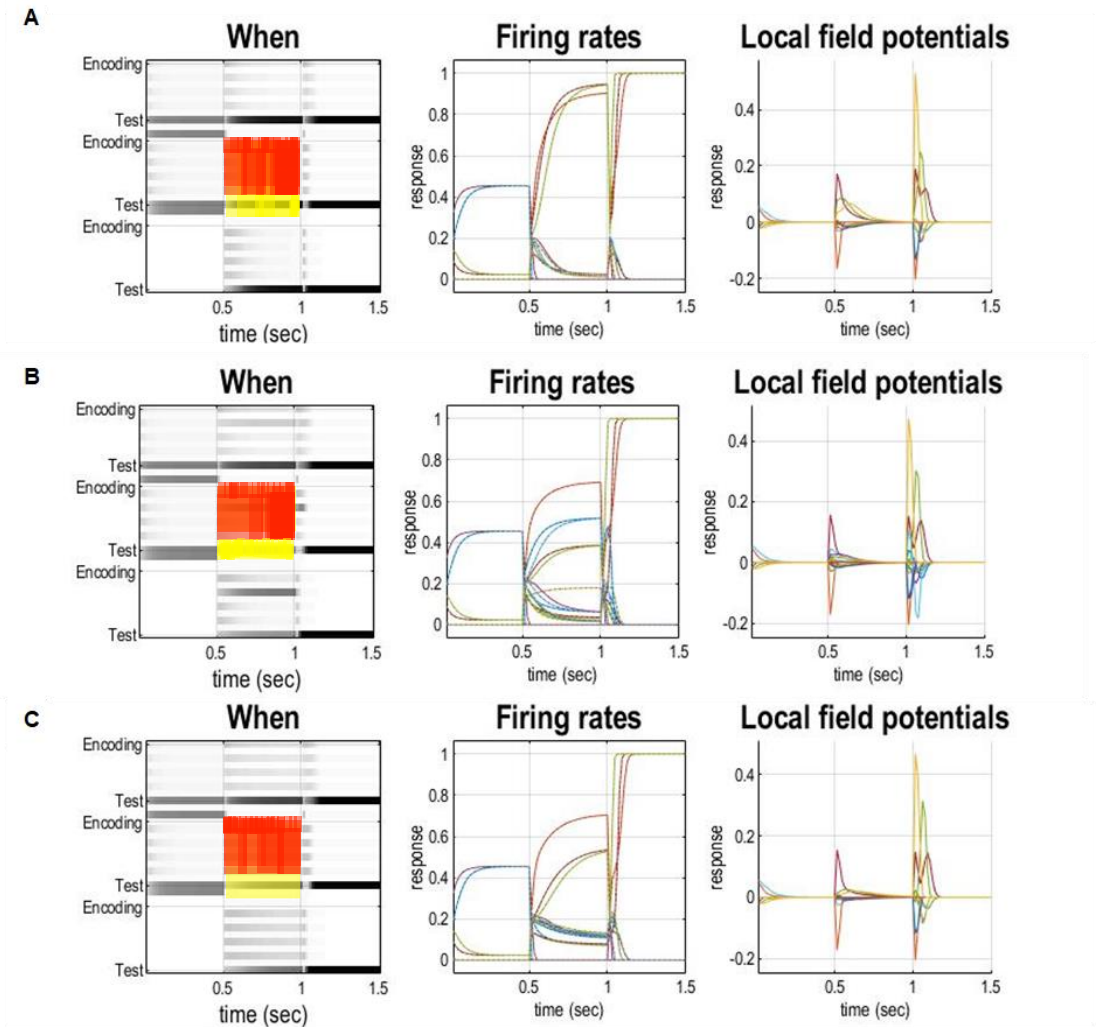


Figure 2.11A2. Simulated electrophysiological responses of the second model are illustrated during the testing phase (i.e., the last three time points) for the new word (A), old word (B) and false memory (C). The agent observes the probe word at the second time point (on the x-axis), reports if the word was seen before and gets

feedback at the third time point. The panels on the left show the expectations about the *when* hidden states. The panels at the centre represent the firing rates associated with the *when* states where each state and epoch is shown with a different colour and line. Finally, the right column shows simulated LFPs in terms of the rate of change in beliefs on hidden states; in other words, as the gradient of the firing rate curves (central column). The highlighted areas in the left plots show the model's beliefs about when the probe word was presented, where the red and yellow shades correspond to the time points in the encoding and testing phases, respectively.

Discussion

In this work, we reproduced the phenomenology elicited by the DRM task for semantics-induced false memory. These models are minimal models of recognition memory (the first or implicit model) and episodic memory (the second or explicit model), with only one episode per trial. Both are set up to accommodate further hierarchical extensions, to accommodate more expressive models of memory recollection and consolidation.

In these hierarchical models, the first level is used to account for the semantics-induced misremembering. Future work on episodic memory *per se* could build upon the structure of the second level; for example, extending the number of episodes. One can also imagine a further hierarchical extension where a third level of the hierarchy provides empirical priors over temporal states of the lower levels. In this way, the high-level states may become an explicit representation of multiple episodes (using a similar hidden state to the 'temporal points' states in the explicit model), which index a particular succession of states that are unique to the episode in question at the lower level.

In this kind of generative model, high-level states provide an index for a particular episode. Technically, episodes become latent states upon which state transitions at a lower level are conditioned. One can see how episodes of episodes can be constructed to any hierarchical depth. The special aspect of this deep model — for episodic memories — is that they contextualise learning. In other words, given a particular episode, the state transitions (i.e.,

succession of states) — and likelihood mappings to the lower-level or outcomes — are uniquely encoded for the episode in question. In exactly the same way that outcomes are ‘remembered’ in the first nine time steps of the second model, higher levels can ‘remember’ sequences of episodes via a temporal encoding by superordinate hidden states. Furthermore, if we enable learning of the probability transitions, the ordinal or sequential unfolding of episodes can be remembered.

In the first (implicit) model, the likelihood matrices had no explicit encoding of temporal mapping related to a ‘time’ hidden state in the second (explicit) model. In the implicit model, we considered a simplistic narrative model such that observing a single word (e.g., snow) was enough for the model to infer the list state in the encoding phase. Because the list factor encodes which words were presented in the encoding phase, the model could report the other words from the list (e.g., ice, chilly, frost), as seen before, even if it has not seen them before. A more complex version of this model could account for which words were presented, from a given list, by adding a further hidden state factor at the higher level and providing feedback accordingly. This additional hidden state factor would have as many levels as the combination of words in the list, so that the model can infer the seen words (for example, this additional hidden state would have factors ‘snow null null null’, ‘snow ice null null’, ‘snow ice chilly null’, ‘snow ice chilly frost’ for the list of cold-related words).

In the second model, the explicit representation of the mapping between the moment in time and word (‘when’ and ‘what’) reproduces a similar sort of mapping to the one proposed by the temporal context model, associated with the medial temporal lobe (Hasselmo et al., 2005). Interestingly, the structure of the likelihood matrix of level one can be interpreted as a simplified version of the auto-associative network in area CA3 of the hippocampus. Our preliminary behavioural results from human participants, suggest evidence for its contribution in item-to-item associations in semantic spaces.

The above formulation foregrounds a challenging issue: namely, if every new episode is represented as a unique episode, then we would need a large number of hidden states to encode every episode encountered. This suggests

that episodes are amalgamated or consolidated, if sufficiently similar, such that there is a finite repertoire of episodes available for encoding the current experience. This further suggests that it will become impossible to disambiguate similar episodes in the past, if they have become consolidated under the same superordinate hidden (episodic) state.

Technically, the issue of when to induce a completely new state — as opposed to reusing a previously indexed episode — is closely related to structure learning, of the sort addressed by nonparametric Bayes (Goldwater, 2006; Salakhutdinov et al., 2013; Gershman et al., 2017; Tervo et al., 2016). Indeed, the consolidation referred to above is an emergent property of active inference at the level of Bayesian model selection. This has been illustrated previously in the context of abstract rule learning using Bayesian model reduction (Friston, Lin et al., 2017). In brief, the marginal likelihood of observations can be increased by removing redundant parameters, which looks as if certain representations are merged or consolidated, to provide a simpler model of the sensorium. This reading of structure learning or Bayesian model reduction offers a simple account for consolidation: it is simply the optimisation of model evidence (a.k.a., marginal likelihood) through the minimisation of complexity inherent in the removal of redundant model parameters. This underwrites the ability of any generative model to generalise to new sensory data — at the price of providing an accurate account of previously observed data. In the present setting, this means that there is an inevitable forgetting of episodic memories that is entirely Bayes optimal, in accord with Occam's principle — and entailed by the minimisation of variational free energy.

Future steps related to this work could focus on model fitting to human behavioural data and deep hierarchical models of multiple episodes (as noted above), where different levels in the hierarchical structure can account for a range of time scales and temporal encoding. In other words, we can adjust the parameters and priors of the generative model until the likelihood of empirical choices is maximised. This enables one to computationally phenotype any given subject in terms of their prior beliefs (Schwartenbeck et al., 2016, Parr et al., 2018). Future developments of this work will entail constructing a deep

hierarchical model for retrieval of multiple episodes. A hierarchical extension of these models could emulate the encoding and indexing of multiple episodes, their reduction to simpler narratives over time (Bayes model reduction) and how experiences that happened before and after the episodes can affect the way in which we reconstruct them.

Software Note

The (MATLAB) routines used in the numerical experiments reported above can be downloaded as part of the open (academic) SPM software from: SPM · GitHub. These are generic routines that simulate belief updating and the accompanying neuronal responses using standard (variational) schemes: in this instance, the variational message passing implemented in spm_MDP_VB_X.m. The only thing that the user has to supply is a generative model specified in terms of the A, B, C ,and D matrices described in the main text. MATLAB routines specifying these models are available on request from the corresponding author.

Chapter 3. Reduced grid-like theta modulation in schizophrenia

Introduction

In this chapter, I focus on how spatial memory and specialised neuronal populations involved in spatial coding can be studied in virtual reality (VR) using magnetoencephalography (MEG), in healthy controls and neuropsychiatric populations.

O'Keefe and Dostrovsky (1971) provided fundamental advances in the field with the discovery of place cells, specialised pyramidal cells located in the CA1 and CA3 areas of the hippocampus (O'Keefe, 1976; Kjelstrup et al., 2008; Henze et al., 2000), with the ability to fire in correspondence with the animal's location in the environment. Place cell populations have been found across mammals (Rotenberg et al., 1996, Ulanovsky and Moss, 2007; Yartsev and Ulanovsky, 2013), and persist in humans (Ekstrom et al., 2003). Each place cell covers a specified area of the environment, called place field. Grid cells populations in the hippocampus present a dorso-ventral gradient organisation, with incrementally bigger place fields moving from the dorsal to the ventral HPC (Jung et al., 1994; Kjelstrup et al., 2008) to represent the environment at different resolutions (van Strien et al., 2009). Interestingly, place fields change across environments, with the same cell firing at different locations depending on the context (remapping, Muller and Kubie, 1987). Thanks to their unique function, place cells encode a map of the environment and contribute to the creation of cognitive maps (O'Keefe and Nadel, 1978; O'Keefe and Nadel, 1979).

The study of neuronal space mapping saw another breakthrough with the discovery of another specialised cell population, in the dorsolateral entorhinal cortex (ERH): grid cells. These cells firing pattern maps the environment by having each cell firing at multiple locations in a periodic fashion following a

hexagonal grid (Fyhn et al., 2004; Hafting et al., 2005, Moser et al., 2014; Sargolini et al., 2006). The space mapping provided by grid cells allow for computing environmental metrics (Bush et al., 2015; Kraus et al., 2015; Dordek et al., 2016). Similarly to place cells, grid cells fields' dimensions are organised over the dorso-ventral axis of the ERH, with anatomical connections between ERH grid cells and HPC place cells with similar scales (Zhang et al., 2013). The scale, in other words the distance between grid fields, is thus organised to represent the environment at different magnitudes. Crucially, the input from HPC place cells to ERH seems to be responsible for the formation of the grid code in ERH (Bonnevie et al., 2013).

Recent advance into the study of neural representations of spatial context and spatial navigation in humans comes from the use of fMRI in healthy volunteers (Doeller et al. 2010), and of intra-cranial recording in epilepsy patients performing a spatial virtual navigation task (Jacobs et al., 2013; Nadasdy et al., 2017), to detect grid-like processing. With this study, I aim to investigate whether grid like coding in healthy volunteers can be detected non-invasively as a modulation of theta power using MEG. Moreover, I ask how this could be used to gain insight into the differences in cognitive processing in patients with diagnosis of schizophrenia.

Overall, the aim of this chapter is to coherently translate recent discoveries in basic neuroscience to improve mechanistic understanding of the neuronal and cognitive processes impaired in schizophrenia.

Schizophrenia is characterised by distortion of thoughts and perception including delusions, hallucinations, disorganised or catatonic behaviour, and diminished emotional expression or motivation (DSM-5, 2013). Several studies suggest a role for the hippocampal formation in the pathophysiology of schizophrenia (Adams et al. 2020, Harrison 2004, Heckers 2011, Heckers 2002, Lieberman *et al.* 2018). Specifically, patients exhibit structural changes in entorhinal cortex (Baiano et al., 2008; Prasad et al., 2004; Roalf et al., 2016) and reduced functional connectivity between the medial temporal lobe (MTL) and medial prefrontal cortex (mPFC) (Adams et al., 2020; Dickerson et al., 2010; Ellison-Wright et al., 2009; Sigurdsson et al., 2010; Weinberger et al.,

1992). The hippocampal formation plays a fundamental role in episodic memory and spatial navigation (Bird et al., 2008; Burgess et al., 2002). Consistent with this, patients with schizophrenia also exhibit impaired performance in a range of spatial navigation tasks (Ledoux et a., 2014; Mohammadi et al., 2018; Salgado-Pineda et al., 2016; Wilkins et al., 2019; Wilkins et al., 2013).

Spatial cognition appears to depend on specialised populations of neurons including grid cells (Hafting et al., 2005), originally identified in the rodent medial entorhinal cortex and subsequently found in the human entorhinal cortex and mPFC in direct intracranial recordings (Jacobs et al., 2013). Grid cells exhibit periodic spatial firing fields with six-fold (or 'hexadirectional') rotational symmetry. Grid cells are thought to support accurate spatial navigation (Bush et al., 2015; Gil et al., 2018; Tennant et al., 2018) and may also contribute to relational memory (Aronov et al., 2017; Constantinescu et al., 2016) and the acquisition of structural knowledge (Behrens et al., 2018). Hence, we examined whether grid cell activity patterns might be disrupted in schizophrenia.

In rodents, grid cell firing patterns appear to depend on movement-related theta band oscillations (Brandon et al., 2011; Koenig et al., 2011; Winter et al., 2015). There is also evidence for movement-related theta oscillations in human intracranial local field potentials (Aghajan et al., 2017; Bohbot et al., 2017; Kahana et al., 1999)., particularly during movement initiation (Bush et al. 2017). Hexadirectional modulation of theta band activity, consistent with the presence of grid cell firing patterns, has also been observed in intracranial EEG recordings from the entorhinal cortex during virtual navigation (Chen et al., 2018; Maidenbaum et al., 2018), building on observations of similar patterns in BOLD signal throughout the default mode network (Doeller et al., 2010). We therefore asked participants with a diagnosis of schizophrenia (half of whom were unmedicated) and a matched control group to complete an established spatial navigation task inside a magnetoencephalography (MEG) scanner (Adams et al., 2020; Kaplan et al., 2014; Kaplan et al., 2012). We then looked for hexadirectional modulation of theta band oscillatory activity during virtual movement.

Materials and methods

Participants

This study re-analyses MEG data first presented in Adams et al. (2020). The study was approved by the local NHS research ethics board (REF: 17/LO/0027), and all participants gave informed consent. Age, sex, IQ, digit span, handedness, and years in education information was collected from all participants. Participants with a schizophrenia diagnosis also completed the Positive and Negative Symptoms Scale (Kay et al., 1987; First et al., 2007), a saliva recreational drugs test (see Supplementary Table 3.1 in Adams et al., 2020), and documented their medication. To be included, participants must have been educated in English, not be using benzodiazepines or anticonvulsants, have normal (or corrected to normal) vision, and be under 60 years old. The patient group was recruited based on DSM-IV criteria for schizophrenia, with 18 participants in total. Patients had no other psychiatric diagnoses, based on the structured clinical interview for DSM-IV-TR axis I disorders[41]. The control group were recruited to match the age, sex, and IQ of the patient group as closely as possible, with 35 participants in total. Controls were excluded if they had history of a psychiatric or neurological condition. In addition, one patient and twelve control participants were excluded due to excessive MEG artefacts, interruption of the experiment due to nausea or sleep, or loss of fiducial markers. This left 17 patients (14 males) and 23 controls (17 males). All participants were asked not to consume caffeine or smoke on the testing day.

Spatial memory task

Inside the MEG scanner, participants performed a spatial memory task in a virtual reality environment (Doeller et al., 2008). constructed using the Unity

game engine (Unity Technologies Ltd). During the task, participants navigated freely around up to three different virtual reality environments and were asked to learn – and subsequently recall – the locations of four different objects in each environment (Fig. 3.1A). Movement was directed using three buttons controlling left and right rotation and forward translation (via rapid acceleration to a fixed maximum speed). The environments were 100 virtual metre (vm) square arenas delineated by a solid boundary and surrounded by distant landmarks. Each environment was distinguished by the surface textures used for the floor and boundary, the location and identity of distal cues, and the location and identity of the objects being memorised. At the start of each block (in each different environment), participants were placed in the centre of the environment facing in the same direction (north).

During encoding, one of four objects was visible in the environment in each trial, and participants were instructed to remember the location of that object. Once they were happy that they had remembered its location, they collided with the object to move to the next trial. There were two encoding trials for each object, in a pseudorandom order, giving eight encoding trials in each environment. Object locations were selected from 16 possible locations, so that each environment contained two objects close to the middle of the arena, one close to a corner and one near the middle of a boundary, to match difficulty across environments (with object locations not used more than once across environments).

During retrieval, each trial began with a 3 s fixation cross, followed by a 3 s cue period in which a single target object was presented on screen. Participants were then placed at a random location and orientation within the environment and asked to navigate to the location of that object and make a button press response. Participants subsequently received feedback on their performance, i.e. the cued object appeared in its correct location, and the next trial began when they collided with the object. Performance in each trial was quantified using the inverse of the distance between the remembered object location and its actual location (such that larger values correspond to better performance, as used in Doeller et al., 2010). There were eight retrieval trials for each object, giving 32 retrieval trials in each environment. Controls and

patients completed 2.70 ± 0.56 and 2.88 ± 0.33 (mean ± SD) task blocks (i.e. environments), respectively.

MEG data collection and pre-processing

MEG data were acquired using a 275-channel axial gradiometer system (CTF Omega, VSM MedTech) at a sample rate of 480 Hz. During the recording, head position coils (attached to nasion and left and right pre-auricular sites) were used for anatomical co-registration, and eye tracking was performed using an Eyelink 1000 system (SR Research). Raw MEG data were imported into SPM12 (Litvak et al., 2011) and downsampled to 200 Hz before eye blink and heartbeat artefacts were manually identified and removed using ICA implemented in FieldTrip (Oostenveld et al., 2011) and EEGLAB (Delorme et al. 2004). Finally, a fifth order, zero phase Butterworth filter was used to remove slow drift (1 Hz high-pass) and mains noise (48-52 Hz notch) from the recordings.

Our analyses focussed on periods of movement onset and complete immobility in the virtual environment. Movement onset 'epochs' were defined as [-3 3] s windows around the onset of continuous translational movements that lasted ≥1 s and were preceded by ≥1 s of complete immobility (consistent with previous studies (Bush et al., 2017)). This captured 25.4 ± 6.9% and 25.5 ± 6.4% of the task data for controls and patients, respectively. Stationary 'epochs' were defined as [-2.5 3.5] s windows around the onset of ≥2 s periods during which no translational movement occurred. This captured 51.4 ± 8.9% and 49.8 ± 7.2% of the task data for controls and patients, respectively (see Table 3.1 for trial numbers). Importantly, although these epochs could overlap, the overlapping time periods were not included in any of our analyses (see Fig. 3.1 and further details below). Once the MEG data had been divided into movement onset and stationary epochs, artefact trials were automatically identified and removed using an underlying outlier test (with a threshold of α=0.05).

Table 3.1: Number of movement and stationary periods (or 'epochs') in controls and patients

	Total movement epochs (mean ± SD)	Bad movement trials (mean ± SD, %)	Included movement trials (mean ± SD, range)	Stationary epochs (mean ± SD)	Bad stationary trials (mean ± SD, %)	Included stationary trials (mean ± SD, range)
CONTROLS	122.6 ± 35.3	3.39 ± 4.22%	119.0 ± 36.4, 61-192	241.7 ± 75.5	3.03 ± 3.92%	234.9 ± 76.6, 110-43
PATIENTS	142.8 ± 45.3	6.95 ± 6.89%	133.5 ± 45.8, 38-246	278.4 ± 82.5	5.91 ± 5.11%	262.9 ± 81.4, 86-408

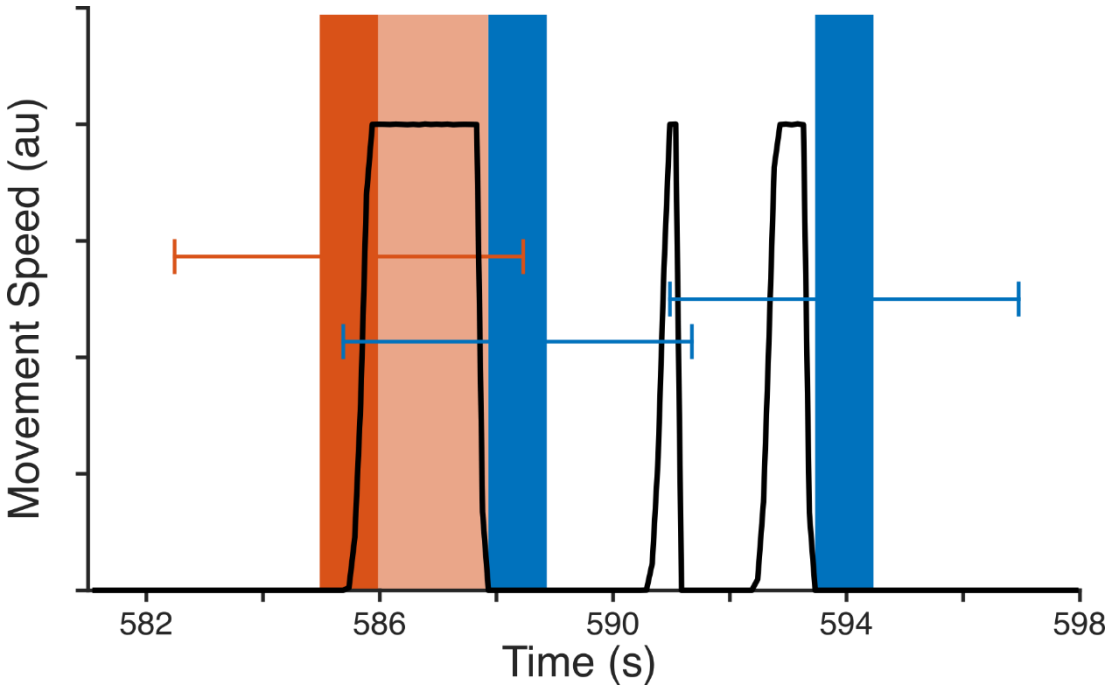


Figure 3.1: Example time course of task conditions. Analyses of movement related changes in oscillatory power (shown in Fig. 3.2 and S3A) compare [-0.5 0.5] s windows around the onset of movements that last ≥1 s and are preceded by ≥1 s stationary periods (shown in dark red) with [0 1] s windows around the onset of stationary periods which last ≥2 s (shown in dark blue). In both cases, the data is extracted from wider 6 s 'epochs' (indicated by the coloured error bars) that can overlap, and provide padding to avoid edge effects in signal processing. Importantly, however, the time windows of interest (shown as dark coloured boxes) are always separated by ≥0.5 s due to the duration thresholds for movement and stationary periods described above. Analyses of oscillatory power modulation by movement direction (shown in Fig. 3.3 and S3B) focus on the full period of translational movement following movement onset in each 6 s epoch (shown in light red). All other task periods (i.e. including movement or stationary periods that do not meet our duration thresholds, stationary periods from 1 s after movement cessation to 0.5s before movement onset, and movement periods from 3 s after movement onset) are unused.

MEG data analysis

To examine changes in low frequency power associated with the onset of virtual movement, we generated a time frequency spectrogram for each movement and stationary period in the 2-70 Hz range using a five cycle Morlet wavelet transform for 40 equally logarithmically spaced frequencies. The resulting power values were log transformed and normalised by the sum of power values across frequencies at each time point. Finally, power values were averaged across epochs for each participant, and power in the [-0.5 0.5] s window around movement onset was baseline corrected by average power in the [0 1] s window during stationary periods. Inspection of the resultant power spectrum, averaged across all participants in both groups, revealed a peak in the 4-10 Hz theta band on which subsequent analyses were focussed. Source localisation of 4-10 Hz theta power was performed in SPM12 using the Linearly Constrained Minimum Variance beamformer from the DAiSS toolbox, with a single-shell forward model and sources evenly distributed on a 10mm grid co-registered to MNI coordinates. This resulted in a set of linear weights for each participant that could generate 4-10 Hz band-pass filtered time series in source space from sensor-level data in each movement onset epoch (Barnes et al. 2003).

To look for the hexadirectional modulation of theta power, we first isolated the continuous period of translational movement following movement onset in each epoch. Next, for each task block (i.e. each virtual environment), we extracted continuous movement direction from the corresponding behavioural data and a measure of theta power by applying the Hilbert transform to band-pass filtered data in each voxel and Z-scoring the resultant time series (to match signal amplitude across voxels and participants). We then estimated grid orientation independently for each voxel using a quadrature filter (Doeller et al. 2010) applied to alternate movement onset epochs from that block. Finally, we estimated the strength of hexadirectional modulation in each voxel for the remaining movement onset epochs by linearly regressing continuous theta power against the cosine of the angular deviation from that grid

orientation, with six-fold periodicity (see Fig. 3.2 for a schematic). We repeated this analysis, reversing the use of alternate epochs for estimating orientation and modulation, and averaged the regression coefficients across the two folds and then across task blocks to provide a single metric indicating the strength of hexadirectional theta modulation for each participant in each voxel ($p<0.001$ uncorrected for setting the cluster, and FWE $p<0.05$ on the cluster size). The same analysis was also performed for other rotational symmetries (specifically: four-, five-, seven-, and eight-fold) and hexadirectional modulation in other oscillatory bands (specifically: 2-4 Hz delta, 12-20 Hz alpha, 20-35 Hz beta and 40-70 Hz gamma). For anatomically defined region of interest (ROI) analyses, we used probabilistic masks from the Julich-Brain Cytoarchitectonic Atlas (Amunts et al., 2020) thresholded at a probability value of 40%.

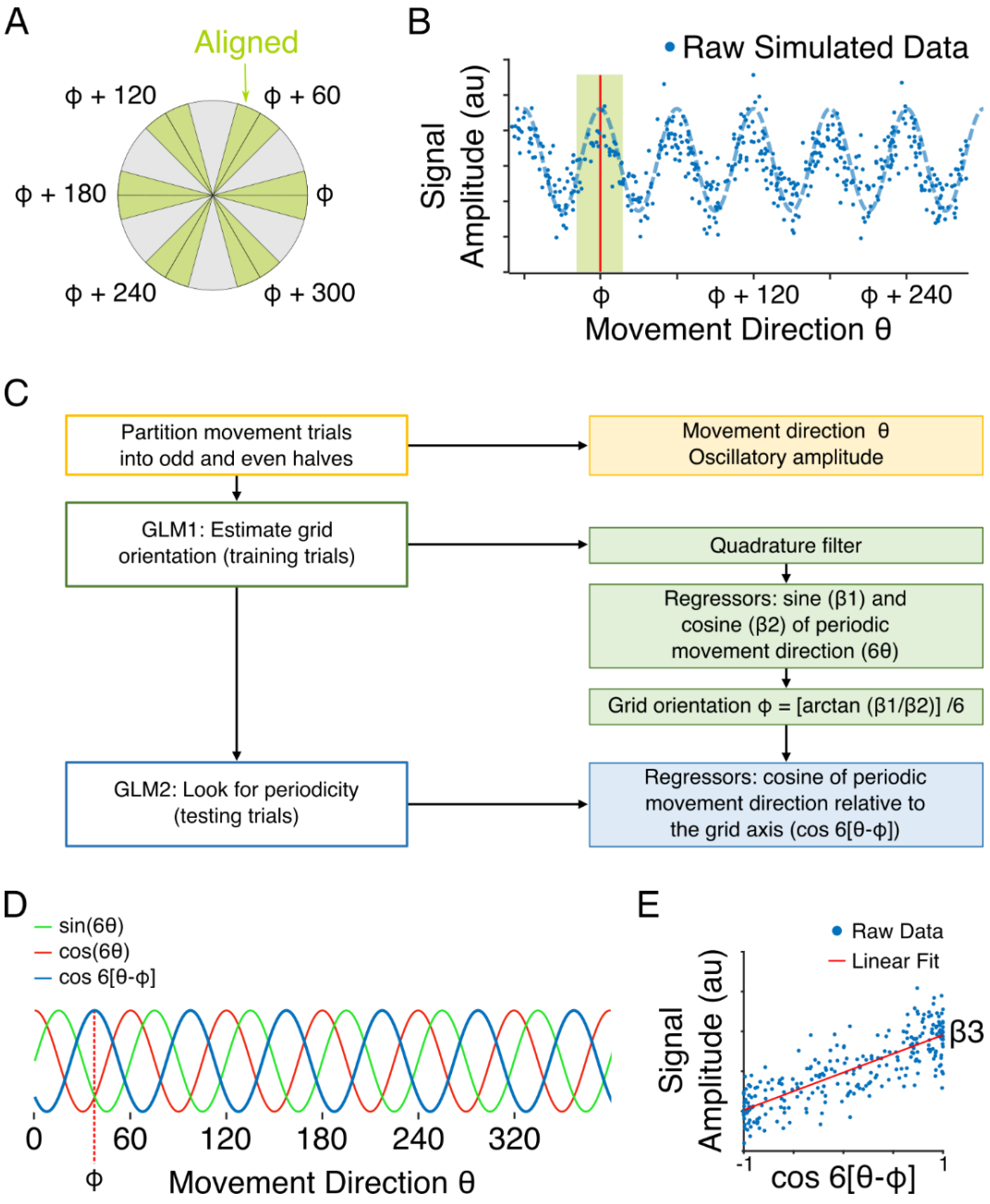


Figure 3.2: Identifying hexadirectional modulation of theta band activity in MEG. **A)** Each movement direction in the virtual environment is either aligned (green) or misaligned (grey) with the orientation φ of an underlying grid firing pattern. **B)** We hypothesise that the amplitude of theta band activity in the MEG signal is sinusoidally modulated by movement direction θ with six-fold (hexadirectional) rotational symmetry. **C)** To test this hypothesis, we first partitioned movement epochs into alternate even and odd trials. For each movement epoch, we then extracted the direction of movement in the VR environment θ and power in the theta band at each time step. In half of the movement epochs, we used a quadrature filter to estimate the orientation of the underlying grid. Specifically, we calculated the sine and cosine of

movement direction θ with 6-fold periodicity and fit those cos(6θ) and sin(6θ) regressors to oscillatory power in a first GLM (GLM1). This produced regressor coefficients β1 and β2 which could be used to estimate grid orientation φ = [arctan(β2/β1)] / 6. We then tested whether the amplitude of oscillatory power in the other half of movements epochs was sinusoidally modulated at this orientation. Specifically, we fit the cosine of movement direction with 6-fold periodicity and orientation φ - cos(6θ-φ) - to oscillatory power in a second GLM (GLM2). This produced a regressor coefficient β3 which characterises the strength of hexadirectional modulation. **D)** Graphic representation of the cos(6θ) and sin(6θ) regressors used in GLM1; and cos(6θ-φ) regressor used in GLM2. **E)** Simulated data are used to show the linear fit between cos(6θ-φ) and oscillatory power in GLM2. The slope of the linear fit corresponds to the strength of hexadirectional modulation. Panels A and D adapted from Doeller et al. (2010).

Results

We asked participants with a diagnosis of schizophrenia (half of whom were unmedicated) and an age, sex and IQ matched control group to perform an established spatial navigation task (Adams et al., 2020, Kaplan et al. 2014, Kaplan et al., 2012; Doeller et al., 2008) using desktop virtual reality (VR) inside a magnetoencephalography (MEG) scanner (Fig. 3.3A). Consistent with previous reports (Ledoux et a., 2014; Mohammadi et al., 2018; Salgado-Pineda et al., 2016; Wilkins et al., 2019; Wilkins et al., 2013), spatial memory performance was significantly better in the control group ($t(38)=2.10$, $p=0.042$, Hedge's $g=0.66$, *CI* [0.028 1.32]; Fig. 3.3B).

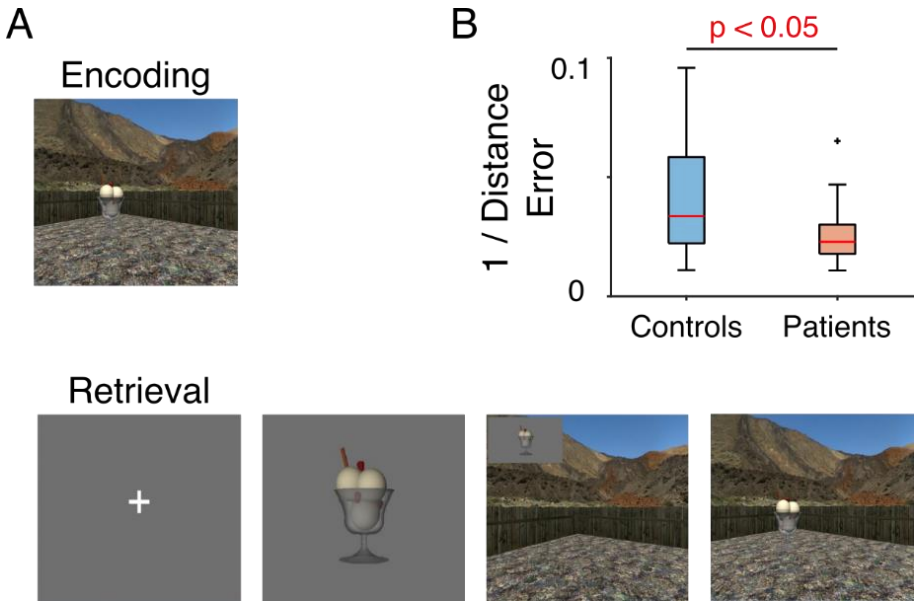


Figure 3.3: Spatial Memory Task. **A)** Schematic. Participants navigate through the environment and make responses using a button box. During encoding, they are asked to remember the locations of four objects (one object being visible in each trial). During retrieval, a fixation cross on a grey screen is followed by an image of one object (cue period). The participants are then asked to navigate from a random start location to the retrieved location of that object and make a response. During navigation, the object image remains visible in the top left corner of the screen. Following a response, the object appears in its correct location to provide feedback. The next trial begins when the participants collide with the object. **B)** Performance, quantified as the inverse of the average distance between remembered and actual object locations, for controls (in blue) and patients (in red). Each red line indicates the median, box edges the 25th and 75th percentiles, whiskers extend to the most extreme datapoints not considered to be outliers (defined as values more than 1.5 times above or below the 75th and 25th percentile, respectively), and outliers are plotted individually. Spatial memory accuracy was significantly higher in the control group.

To look for evidence of grid-like activity during translational movement within the VR environment, we first investigated changes in oscillatory power associated with movement onset versus stationary periods. Power spectra for

both groups, averaged across all sensors, showed a peak in the theta band during movement onset (Fig. 3.4A). Specifically, 4-10 Hz theta power was greater during movement onset than stationary periods in both controls ($t(22)=5.58$, $p<0.001$) and patients ($t(16)=2.39$, $p=0.03$), and greater in controls than patients ($t(38)=2.02$, $p=0.05$, $g=0.63$, CI [0.0014 1.29]; Fig. 3.5A). This is illustrated by time-frequency spectrograms of movement onset periods (Fig. 3.4B), which show a clear increase in theta power in the control group beginning ~0.5 s prior to movement onset (consistent with previous reports (Bush et al., 2017; Kaplan et al., 2012)) that is markedly reduced in patients.

Scalp plots (showing normalised power differences between movement onset and stationary periods) illustrate that 4-10 Hz theta power increases arise over bilateral frontal and temporal sensors in both groups, with controls showing greater movement-related theta power than patients over left frontal sensors (Fig. 3.4C). Importantly, we found no evidence for differences in movement statistics between control and patient groups in the virtual environment that could account for these differences. Specifically, there were no differences in the average duration of movements between patients (mean ± SD = 2.29 ± 0.43 s) and controls (2.18 ± 0.5 s; $t(38)=-0.723$, $p=0.47$) or preference to navigate close to the boundaries of the environment (patients: 79.2 ± 4.8%; controls: 79.9 ± 6.1%; $t(38)=0.37$, $p=0.71$), and movement speed accelerated rapidly to a fixed top speed for all participants.

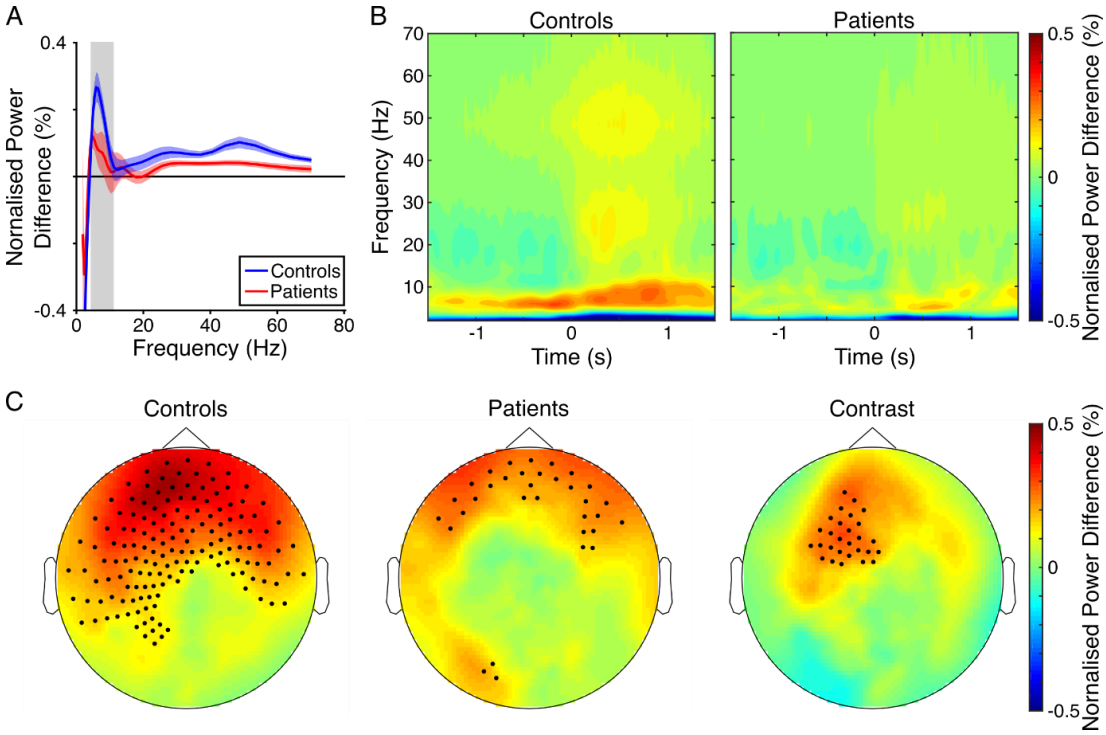


Figure 3.4: Movement-related 4-10Hz theta power increases in controls and patients. **A)** Power spectra showing normalised power during movement onset epochs (i.e. [-0.5 0.5] s around the onset of ≥1 s translational movements that were preceded by ≥1 s immobility), baseline corrected by average power during stationary periods (i.e. [0 1] s around the onset of ≥2 s periods of immobility) for controls (in blue) and patients (in red, shading indicates standard error). The grey bar delineates the 4-10 Hz theta band. **B)** Time-frequency spectrograms showing normalised power during movement onset, baseline corrected by average power during stationary periods. Controls show a marked increase in theta power beginning ~0.5 s prior to movement onset that is reduced in patients. **C)** Scalp plots of normalised 4-10 Hz theta power during movement onset epochs, baseline corrected by average theta power during stationary periods for controls, patients, and for the contrast between groups. Highlighted channels show significant positive power differences at a threshold of $p<0.01$ (uncorrected).

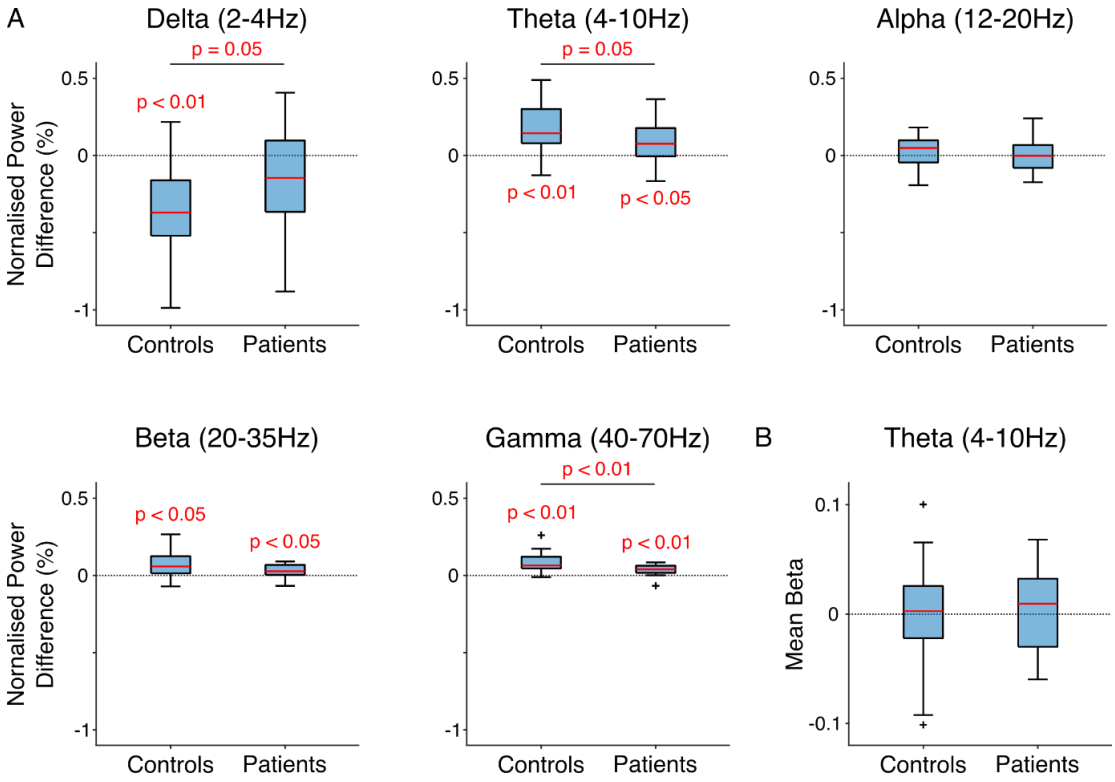


Figure 3.5: Movement related changes in oscillatory power and hexadirectional modulation of theta power in left entorhinal cortex. **A**) Normalised oscillatory power during movement onset epochs (i.e. [-0.5 0.5] s around the onset of ≥1 s translational movements that were preceded by ≥1 s immobility), baseline corrected by average power during stationary periods (i.e. [0 1] s around the onset of ≥2 s periods of immobility) for controls (in blue) and patients (in red) in the delta (2-4Hz), theta (4-10Hz), alpha (12-20Hz), beta (20-35Hz) and gamma (40-70Hz) frequency bands, averaged across all sensors. In addition to the significant changes in theta power described in the main text, we find evidence for movement related decreases in delta power in both controls ($t(22)=-5.93$, $p<0.001$) and patients ($t(16)=-2.11$, $p=0.05$), as well as a significant difference between groups ($t(38)=2.02$, $p=0.05$, $g=0.633$, *CI* [0.0024 1.29]); movement related increases in beta power in both controls ($t(22)=3.74$, $p=0.001$) and patients ($t(16)=2.76$, $p=0.014$), with no difference between groups ($t(38)=1.57$, $p=0.13$); and movement related increases in gamma power in both controls ($t(22)=6.39$, $p<0.001$) and patients ($t(16)=4.12$, $p<0.001$), as well as a significant difference between groups ($t(38)=2.74$, $p=0.0092$, $g=0.86$, CI [0.221 1.54]); but no changes in alpha power in either group (both $p>0.15$) or difference between groups ($t(38)=0.557$, $p=0.58$). **B)** Absence of hexadirectional theta modulation inside an anatomically-defined left entorhinal ROI for controls ($t(22)=-0.184$, $p=0.856$) and patients ($t(16)=0.45$, $p=0.659$). There is no significant difference in the strength of hexadirectional modulation between groups in this ROI ($t(38)=-0.419$, $p=0.677$). Each

red line indicates the median, box edges the 25th and 75th percentiles, whiskers extend to the most extreme datapoints not considered to be outliers (defined as values more than 1.5 times above or below the 75th and 25th percentile, respectively), and outliers are plotted individually.

Next, we looked for hexadirectional modulation of movement-related theta power across the whole brain using established methods (Doeller et al., 2010) (see Fig. 3.2 for further details). Remarkably, the control group showed a single significant cluster of hexadirectional theta modulation in the vicinity of right entorhinal cortex (Fig. 3.6A). In contrast, the patient group showed no clusters that passed our threshold of $p<0.05$ FWE corrected across the whole brain.

To further characterise this effect, we extracted the strength and orientation of hexadirectional theta power modulation from each voxel in an anatomically-defined right entorhinal region of interest (ROI) for each participant (Fig. 3.6B). Consistent with the whole brain results, this revealed significant hexadirectional modulation of 4-10 Hz theta power for controls ($t(22)=3.04$, $p=0.0059$) but not patients ($t(16)=-0.04$, $p=0.97$), and significantly stronger hexadirectional modulation for controls than patients ($t(38)=2.08$, $p=0.044$, $g=0.65$, *CI* [0.02 1.31]; Fig. 3.6C). Similarly, theta power in this ROI was greater during movement aligned versus misaligned with the grid axes for controls (i.e. within ±15° of the fitted grid orientation versus other movement directions; $t(22)=2.82$, $p=0.01$; Fig. 3.6D), despite no difference in the proportion of movement samples with aligned versus misaligned directions ($t(22)=-0.70$, $p=0.49$; Fig. 3.6E). Importantly, theta power from this ROI was not significantly modulated by four, five, seven or eight fold movement direction in the control group (although we note a trend towards significance for eight-fold modulation ($t(22)=2.03$, $p=0.055$; all others $p>0.27$; Fig. 3.6F), nor was there any evidence for hexadirectional modulation of delta, alpha, beta, or gamma frequency band activity in this region (all $p>0.26$; Fig. 3.6G). In addition, we found no evidence for the hexadirectional modulation of theta power within a corresponding anatomically-defined left entorhinal ROI (Supplementary Fig. 3.6B).

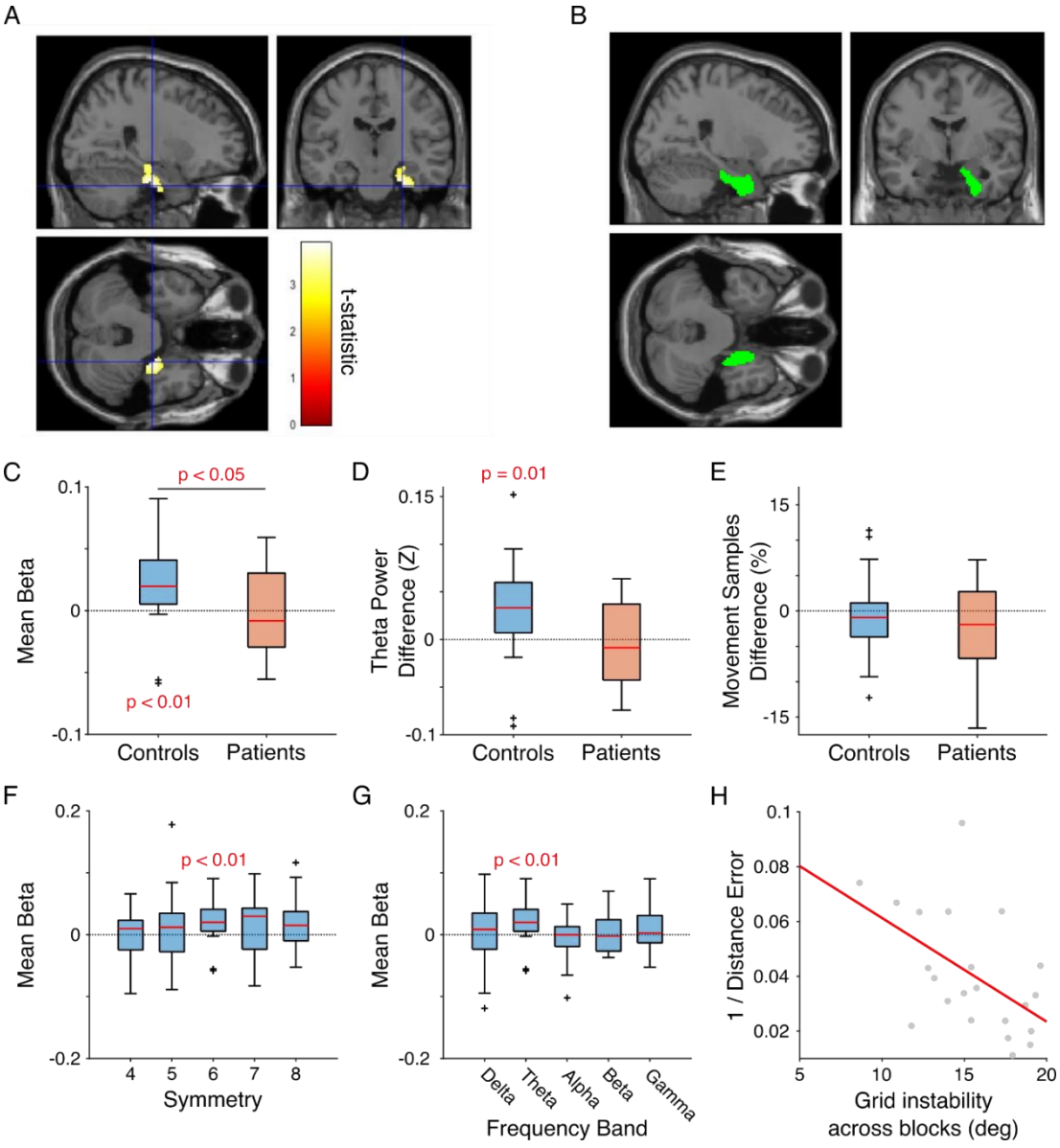


Figure 3.6: Modulation of oscillatory power by movement direction in right entorhinal cortex. **A)** Regions showing significant hexadirectional modulation of 4-10Hz theta power at the whole brain level. Only one cluster in right entorhinal cortex (peak at [18 -22 -44], Z=4.05) passes our significance threshold of $p<0.05$ FWE corrected (image shown at $p<0.005$ uncorrected, for display purposes). **B)** Image of the anatomically defined right entorhinal cortex region of interest (ROI). **C)** Strength of hexadirectional theta modulation inside the ROI for controls and patients, with 19/23 controls (82.6%) and 8/17 patients (47.1%) showing a positive beta coefficicent. **D)** Difference in theta power between on vs off axis movement inside the ROI for controls and patients, with 19/23 controls (82.6%) and 7/17 patients (41.2%) showing greater on vs off axis theta power. **E)** Difference in the percentage of movement samples that occurred during on vs off axis movement for controls and patients. **F)** Theta modulation by 4-8 fold

movement direction inside the ROI for controls. **G)** Strength of hexadirectional modulation of delta (2-4Hz), theta (4-10Hz), alpha (12-20Hz), beta (20-35Hz) and gamma (40-70Hz) frequency bands inside the ROI for controls. **H)** Correlation between performance, quantified as the inverse of the average distance between remembered and actual object locations, and grid (in)stability across task blocks for controls. Each red line indicates the median, box edges the 25th and 75th percentiles, whiskers extend to the most extreme datapoints not considered to be outliers (defined as values more than 1.5 times above or below the 75th and 25th percentile, respectively), and outliers are plotted individually.

Reassuringly, grid orientation across voxels inside the right entorhinal ROI (within each task block and data partition) was more consistent than expected by chance (5.33 ± 2.25°, chance=15°; $t(22)=-20.7$, $p<0.001$), as was grid orientation across data partitions, each including half of the trials (within each task block and ROI voxel; 12.9 ± 3.55°; $t(22)=-2.85$, p=0.0093). However, grid orientation across blocks (within each data partition and voxel inside the ROI) was no more consistent than expected by chance (15.5 ± 3.05°, $t(21)=0.71$, $p=0.49$), suggesting that grid patterns randomly realigned with the visually distinct square environment encountered in each task block. Importantly, we found no evidence for a relationship between theta power during movement onset (averaged across all sensors) and the strength of hexadirectional modulation inside the ROI (Pearson's $r=0.32$, $p=0.14$); or between theta power during movement onset (averaged across all voxels within the ROI) and the strength of hexadirectional modulation in the same region ($r=0.25$, $p=0.25$). This suggests that differences in the magnitude of hexadirectional modulation across participants did not arise simply from differences in the power of the underlying theta oscillation.

Finally, we looked for a relationship between the hexadirectional modulation of 4-10 Hz theta power inside the ROI and our behavioural data. Although we found no evidence for a correlation between the strength of hexadirectional modulation and task performance across controls ($r=0.15$, $p=0.49$), we did find a significant relationship between the consistency of the grid orientation across blocks and task performance ($r=-0.52$, $p=0.013$; Fig. 3.3H). This indicates that

control participants with grid patterns that were more consistent across task blocks tended to more accurately remember object locations in the VR environments.

Discussion

This is the first demonstration of hexadirectional theta modulation in MEG, building on previous studies showing similar patterns in BOLD signal throughout the default mode network (Constantinescu et al., 2016; Doeller et al., 2010), in high frequency activity from the anterior temporal lobe in both MEG and intracranial EEG recordings (Staudigl et al., 2018), and in entorhinal theta power from intracranial EEG recordings (Chen et al, 2018; Maidenbaum et al., 2018). Crucially, however, the relationship between grid cell activity at the neural level, network level modulations of theta or high frequency power in the local field potential or in MEG, and the BOLD signal measured using fMRI are not clear, and merit further attention. These findings show hexadirectional modulation of theta power in right entorhinal in healthy volunteers, which is consistent with the presence of stable grid cell firing patterns. Importantly, the stability of grid orientation across task blocks in the control population correlated positively with their performance in the spatial memory task, suggesting a functional relationship between grid firing patterns and spatial memory.

Previous studies have reported impaired spatial navigation associated with hippocampal anomalies in schizophrenia (Ledoux et a., 2014; Mohammadi et al., 2018; Salgado-Pineda et al., 2016; Wilkins et al., 2019; Wilkins et al., 2013). In particular, people with schizophrenia are selectively impaired in spatial navigation strategies based on cognitive mapping, rather than single-landmark (response-based) strategies (Wilkins et al., 2019; Wilkins et al., 2013). These results demonstrate that people with schizophrenia show worse spatial memory and less movement-related theta power during a virtual spatial

navigation task than a matched control group. They also lack the hexadirectional modulation of theta power cortex observed in the control group.

Schizophrenia is also associated with impairments in associative inference and acquisition of relational knowledge (Adams et al., 2020; Armstrong et al., 2018; Armstrong, Kose et al., 2012; Armstrong, Willams et al., 2012), in which the hippocampal formation - and grid cells in particular - are thought to play a key role (Behrens et al., 2018). Our findings therefore suggest that dysfunctional grid coding may underlie atypical inference and poor acquisition of relational knowledge in schizophrenia. Grid firing patterns may be supported by attractor network dynamics (McNaughton et al., 2006), and attractor states are thought to be more unstable in schizophrenia (Adams et al., 2018; Hamm et al., 2018), potentially due to reduced α5-GABA-A receptor density in the MTL (Marques et al., 2021). We speculate that this may increase reliance on striatal learning mechanisms, making inferences more dependent on individual landmarks (or, perhaps, events) than structured relational knowledge of the world.

It is interesting to note the discrepancy between movement related theta power increases at the sensor level, where differences between groups are most prominent over left frontal regions; and the hexadirectional modulation of theta power by movement direction, which is restricted to the right entorhinal cortex in control participants. This suggests independent underlying mechanisms, which is supported by the absence of any correlation between theta power and the strength of hexadirectional modulation across our control group. Similarly, we find no evidence for the hexadirectional modulation of theta power in an anatomically-defined left entorhinal cortex ROI, in contrast to some previous studies (Jacobs et al., 2013; Chen et al., 2018; Maidenbaum et al., 2018), although our results are not sufficient to make strong claims about laterality.

In summary, in healthy volunteers performing a virtual spatial navigation task, we have shown grid-like modulation of MEG theta power localised to the right entorhinal cortex whose consistency of orientation across virtual environments correlates with spatial memory performance. Relative to this baseline, we have

shown that people with a diagnosis of schizophrenia have impaired spatial memory performance, reduced movement-related theta oscillations and disrupted grid-like modulation of theta power. This extends previous work showing structural and functional impairment of the hippocampal formation in schizophrenia and selective deficits of hippocampus-dependent strategies in spatial navigation. Future studies could address a possible role of grid cell populations in impaired structural knowledge and inference in schizophrenia.

Overall, this chapter provides additional insight on the role of grid cells in the representation of spatial context in memory for spatial locations. Moreover, it validates MEG analysis of theta power as a non-invasive method to study grid cell activity in humans.

Chapter 4. Navigating Memory through Semantics and Time

Introduction

In this chapter, I further investigate the interaction of semantic and temporal contexts in memory for word lists, as in Chapter 2. This superficially simple task, however, is finely controlled and implicitly structured to open a window into the underlying neural mechanisms responsible for integrating conceptual dimensions that are usually assigned to separate semantic and episodic aspects of declarative memory function.

The hippocampus (HPC) and para-hippocampal areas in the medial temporal lobe (MTL) have a well-established fundamental role in episodic memory (Tulving 1983), spatial memory, and navigation, in the mammalian brain, including the human brain (Moser et al., 2015; Burgess et al., 2002; Doeller et al., 2010).

These structures host different neuronal populations that play key roles in memory encoding and retrieval, and in encoding spatial information for efficient navigation . As summarised in Chapter 3 (Introduction), grid cells have been at the core of a rich body of research aimed at better understanding how the human brain flexibly navigates the environment and creates cognitive maps of both the real world and abstract spaces (O'Keefe and Nadel, 1978; O'Keefe and Nadel, 1979). Grid cells were originally found in the dorsolateral entorhinal cortex (ERH) (Fyhn et al., 2004; Hafting et al., 2005, Moser et al., 2014; Sargolini et al., 2006), where they map the environment via hexadirectionally modulated firing activity (Bush et al., 2015; Kraus et al., 2015; Dordek et al., 2016).

In both rodents and humans, grid cell activity appear to be associated with theta oscillations during movement (Brandon et al., 2011; Koenig et al., 2011; Winter et al., 2015, Aghajan et al., 2017; Bohbot et al., 2017; Kahana et al.,

1999, Bush et al. 2017, Chen et al., 2018; Maidenbaum et al., 2018). While keeping hexagonal symmetry, grid cell populations share their orientation. Movements in the environment are associated with an angle of orientation that can fall either in phase with the angle of orientation of the grid cell population (aligned with the axis) or out of phase (misaligned with the axis) (Barry et al., 2007, Stensola et al., 2012). This shared population orientation of the grid cell acitivity allowed scientists to develop new methods to detect the bulk six-fold (hexagonal) symmetry grid activity in humans using non-invasive neuroimaging via fMRI (Doeller et al., 2010; Kunz et al., 2015; Bellmund et al., 2016; Constantinescu et al., 2016; Horner et al., 2016) and MEG (Staudigl et al., 2018, Giari et al., 2023; Convertino et al., 2023, see Chapter 3). These studies found grid cell-like activity in humans,not only in the ERH but also in other areas of the default mode network (medial prefrontal cortex, medial parietal cortex and lateral temporal areas).

Over the last decades, new insights from human and primate neuroscience have confirmed the hypothesis that cognitive maps play an important role in non-spatial navigation and support flexible cognition and generalisation across abstract contexts (Tolman, 1948; O'Keefe and Nadel, 1978; Buzsáki and Moser, 2013; Eichenbaum and Cohen, 2014; Farzanfar et al., 2023). Recent studies have found evidence of grid cell activities in primate and human navigation across abstract conceptual spaces (Killian et al., 2012; Constantinescu et al. 2016; Bao et al., 2019; Theves et al., 2019, 2020; Vigano et al., 2021; Park et al., 2021; Raithel et al. 2023).

Moreover, different new theories have proposed a broad inter-domain role for the MTL via predictive mapping, which would provide the underline machinery for reward-based behaviour, as well as sensory expectations and concept search (Stachenfeld et al., 2017; Mok and Love, 2019; Mok and Love, 2023; Epstein et al., 2017; Solomon et al., 2018). In this direction, recent work by Solomon et al. (2019) found that theta power in the hippocampus predicted the semantic distances between words during free recall in a cohort of epilepsy patients using intracranial recording (iEEG). This result opens new interesting investigations on the role of the MTL and cognitive maps in semantic aspects of declarative memory and concept navigation.

Based on the work of Solomon et al. (2019), we used 3-T fMRI to investigate whether grid-like activity in the ERH cortex, mPFC and potentially other areas of the default mode network could be involved in learning of lists of words. To do so, we created an implicit abstract 2-D space, where one axis represents the semantic distance between words in word2vec embedding space, and the other axis represents the ordinal distances between the locations of the words in a studied list. 'Moving' from thinking of one word to another would result in an abstract movement in this environment.

The goal of this chapter is to provide preliminary results and insight into the mechanism of interaction of temporal and semantic contexts in declarative memory. This aims to guide and inform further confirmatory analysis and new hypotheses testing in future work. Although the intent of this chapter is not to deliver an exhaustive understanding of the experimental findings, or finalised wholistic analysis, I hope to lay the foundation blocks for new insights on the neuronal mechanisms responsible for the integration of multidimensional and metacontextual information in declarative memory.

Material and Methods

Participants

The study was approved by UCL ethics board (REF: 1338/009). Thirty-four participants were recruited via Sona System. Inclusion criteria were age between 18- and 35-year-old, native English speaker (mother tongue), right-handed. Age, sex, handedness and years in education data and informed consent to participate in the experiment were collected for all participants. One participant was later excluded due to brain anatomical abnormalities found during the fMRI scan, and three participants were excluded due to inability to complete the task during scan. Thirty participants were included in the analysis, of which 19 female and 11 males with age mean $M = 25.1, SD = 4.1$; with ages spanning between 19- and 35-year-old. Before scanning, on the same day, subjects were trained to perform the sequence memory task outside

the scan on a computer; they then participated in one fMRI scan session. Finally, outside scan, they were asked to perform two similarity judgment test tasks post-scanning and debriefed. The experiment took place at the Wellcome Centre for Human Neuroimaging (FIL, 12 Queen Square, UCL).

Word list learning task

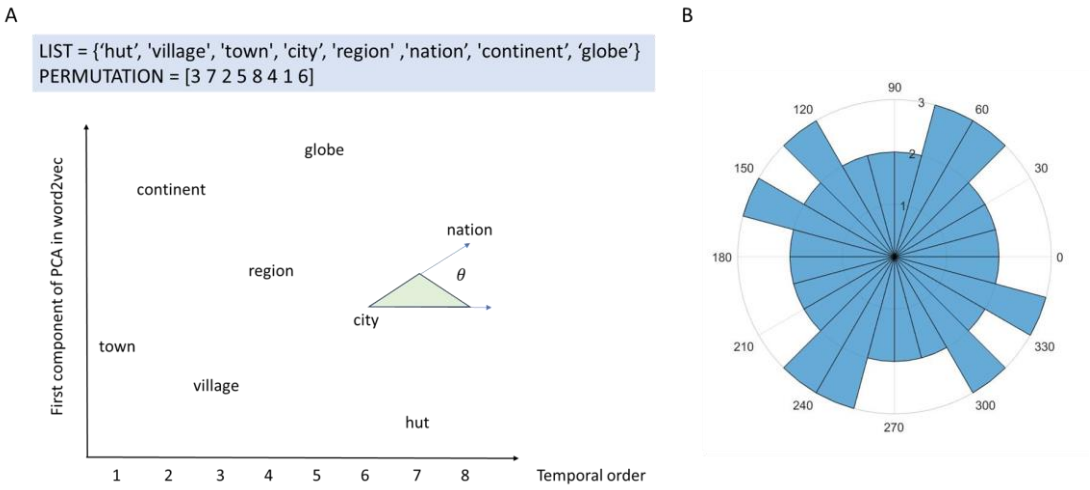


Figure 4.1. A. Example of the distribution of words in the abstract 2-D space for an example permutation of the words in the list. An example of angle θ can be seen between the word 'city' and the word 'nation'; moving between the word 'city' (6^{th} word in the list on the temporal order axis, and 4^{th} word over the first PCA component on the semantic axis, i.e. coordinates 6,4) and the word 'nation' (coordinates 8,6), would create the presented angle on the 2D abstract space. B. Distribution of all the possible angles created by all the possible paired associations of words for the permutation used in figure A. To avoid biases and confounds in the trajectories between words, I controlled for a balanced distribution of all angles tested in the experiments.

The experimental task consisted in a list of eight words that the participants were asked to learn in the correct order. The task was repeated twice, with two different conditions corresponding to two different lists of words. The two lists differentiate form each other based on how the words in each list were selected. While the first list included words from the same taxonomic group,

the second list included words across a variety of taxonomic semantic domains, as better explained in the next paragraph and illustrated in Fig. 4.2. Without the participants knowing, both lists were built so that the words in each list were distributed evenly in a 2-D space with dimension (Fig.4.1): 1st PCA component of a PCA analysis of the word vectors in word2vec, and word position (order) within the list. The order of the words in the list was organised so that the two dimensions, i.e. the order of the words over the first PCA component in word2vec and the temporal order of the words in the list, were not correlated with each other (Pearson correlation coefficient = 0). To achieve this, every word was transformed into its vector form via word2vec MATLAB implementation, and the resulting word vectors were organised in order over the first principal component for each list. Critically, each trajectory between two words in the 2-D space is associated to an angle . This will be fundamental to investigate hexadirectional modulation of brain activity. Moreover, I controlled for the even distribution of angles in the 2D space for each trajectory between all possible combination of words (Fig.4.1 B) to avoid direction biases. I then selected two different permutation of the words that provided the preferred occupancy of the words in the 2-D space, and the orthogonality of semantic and order dimensions. Each condition was tested with both permutations, counterbalancing across participants (Fig.4.2).

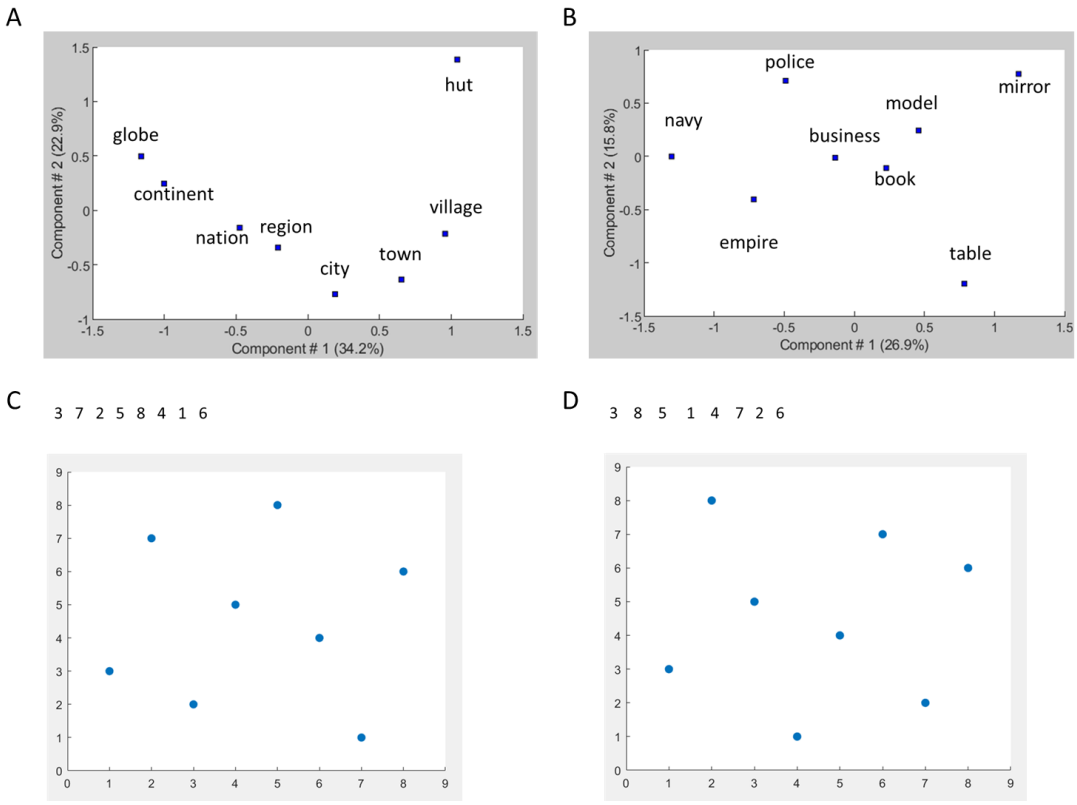


Figure 4.2. Projection over the first and second component of PCA analysis from word2vec word vectors for the continuous (A) and the mixed (B) lists. Each list was tested with one of two possible permutations for each participant (C and D). For each participant, both permutations were used, one for each list type, and the order of the two lists and of the two permutations during the task was counterbalanced across participants.

For the first condition (continuous list, Fig.4.2 A), I selected eight words related with each other via a taxonomic semantic grouping, where each word referred to a human populated human community aggregate. The first component of the PCA for the first list organised the words in the same order as the size-based one associated with their meanings.. The semantic domain that better suited these criteria was the one of community aggregates (hut, village, town, city, region, nation, continent, globe). These words can beorganised over a size dimension, which corresponds to the order of the words over the first PCA of their vectors.

For the second condition (Fig.4.2 B), the words were selected from the iEEG Free Recall pool of high frequency nouns of the Computational memory laboratory, University of Pennsylvania (https://memory.psych.upenn.edu/Word_Pools). The criteria of inclusion for the words in both lists controlled for comparable string length (maximum difference in word length of two letters) and frequency of the words and aimed to maximise the difference in spelling between words in the same list (Levenshtein distance greater than two). Moreover, the first PCA component of the word vectors had to explain at least 25%, and the difference between first and second PCA components had to be higher than 10%. Finally, the words within each list had to be evenly spread over the first PCA component. The mix list condition was characterised by the lack of a taxonomical common group , which made it more difficult to organise the words over a continuous dimension based on an evident descriptive semantic characteristic (such as size or colour brightness).

Before the scan, the participants were asked to perform a trial version of the experiment with random words (not included in the fMRI task) on a testing room computer. This training did not involve any relevant learning for the task in the fMRI scan but was carried solely to support the participant familiarise with the structure of the task, the nature of the stimuli and the response buttons apparatus. Once the training was completed, the participants performed the task inside the fMRI scan for both lists (encoding and retrieval) in one single session of the total duration of about one hour, including structural data acquisition.

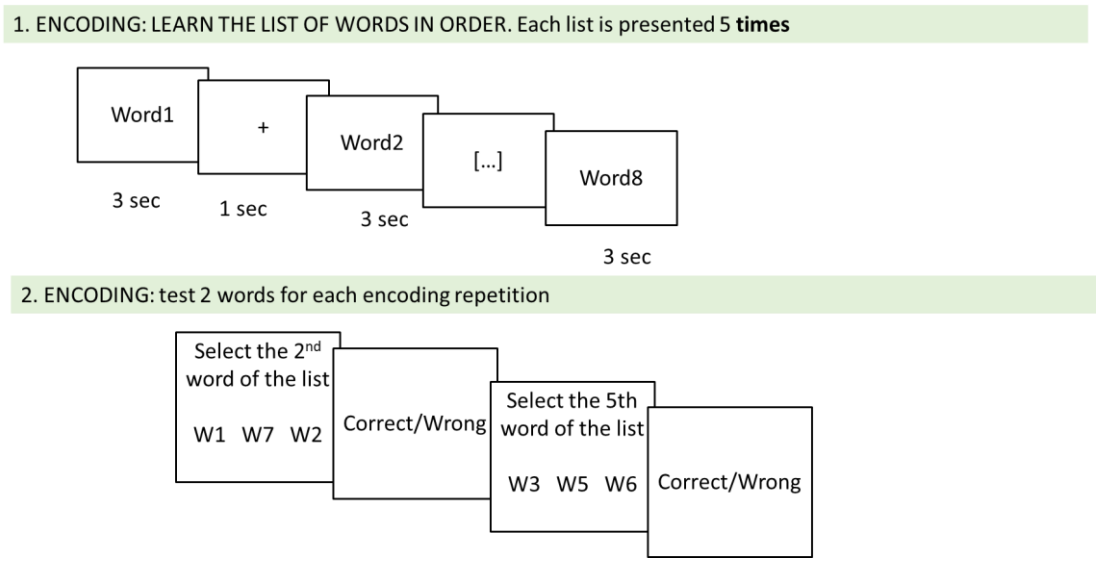


Figure 4.3. Structure of the encoding phase of the experiment in the scan. Each word was presented on the screen for 3 seconds, followed by one second fixation cross between words. After the 8 words in the list were presented, the participants were asked to select the right word for a specific position in the presented list. After each presentation of the list, the participants were asked two questions and received feedback for their answer. This process was repeated five times to support learning.

The task was organised in an encoding and retrieval phase. During the encoding phase (Fig.4.3), the participants were asked to remember the words in the list in the correct order in which they were presented. Each word was presented in white on a dark background for 3 seconds on the screen, followed by a fixation cross. After list presentation, the participants were tested with two questions. They were asked to select the word corresponding to a specific position in the list among three possible options; for each answer, they received right or wrong feedback. This was introduced to support the learning of the list in the correct order. The encoding phase, including both list presentation and two memory questions, was repeated 5 times.

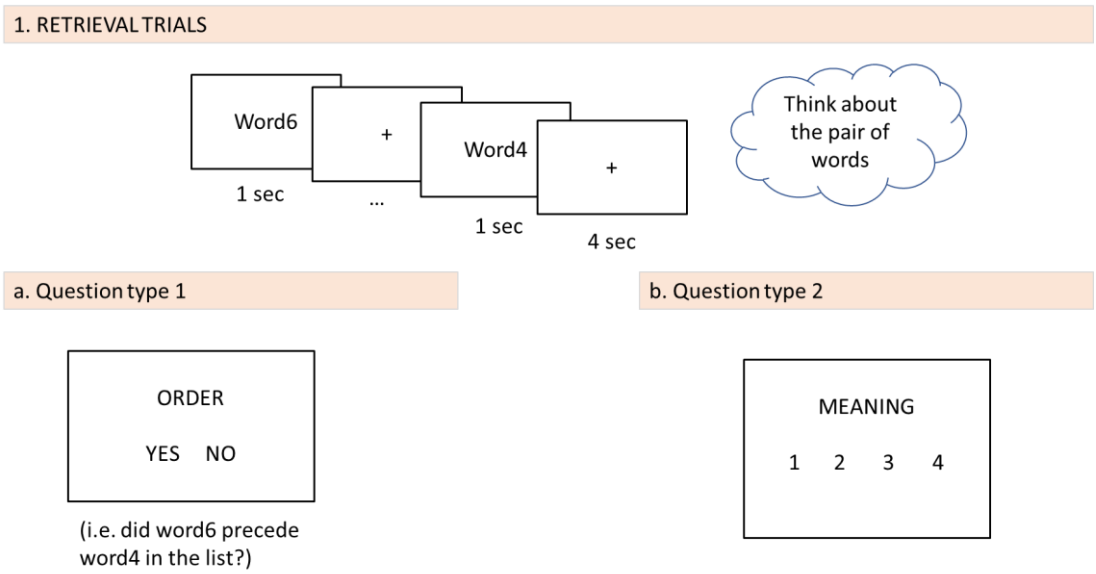


Figure 4.4. Structure of the testing phase of the experiment. The participants were presented with different pairs of words, with one word presented on the screen at the time. They then had 4 seconds to think about the two words they just saw and their position (order) in the learnt list. For one third of the trials, the participants were randomly presented with a question that could either test their memory for the order of the words in the list (question type 1) or ask for their subjective rating for the similarity of the meanings of the two words (question type 2). They had two seconds to answer to either question. The participants were instructed had the chance to familiarise themselves with the structure of the experiment and the meaning of the two questions before starting the experiment in the scanner.

During the retrieval (testing) phase (Fig.4.4), the participants were presented with 128 trials structured as followed. Each trial consisted in a pair of words, so that each word in the pair was presented individually on the screen for one second, followed by fixation cross, and by the second word. The jittered inter-word interval (ITI) generated from a truncated (at 7 s) Poisson distribution with a mean of 2 s. The 128 pairs were made of twenty-eight unique pairs from a list of eight items, presented in both possible orders, twice (116 trials), and of an additional 16 trials of the same word repeated twice within the pair (two trials for each word in the list). The order of presentation of the pairs was semi-randomised to adjust for an even distribution of temporal (order) distances between words across trials. After the pair of words was presented, the

participants were exposed to a blank screen with dark background and no fixation cross for 4 seconds.

The experiment, including both encoding and retrieval phases, was repeated twice, once for each list. The order of the lists and the semi-randomized presentation of pairs during retrieval were counterbalanced across participants, utilizing eight possible permutations of list order and pair presentation.

Participants were instructed to focus on the position (i.e. first, second etc in the list of eight) of the words just presented in the pair, relative to their position in the original list of eight words shown during encoding. This task was performed during a four-second interval of blank screen exposure. In one third of the pairs, randomly distributed, the blank screen was followed by a question designed to ensure participants' sustained attention and task engagement. Two types of questions were presented with equal frequency (21 randomized trials each). The first type of question asked whether the first word in the current pair had appeared before the second word in the original list of eight, requiring a yes or no response. The second type of question asked the semantic similarity between the two words on a scale from 1 to 4, with '4' indicating the highest level of similarity. Participants had two seconds to respond. The trial advanced either when the participant provided a response or automatically after the two-second response window, followed by a jittered inter-trial interval (ITI) from a truncated Poisson distribution with a mean of 2 s.

The experiment, including both encoding and retrieval phases, was conducted twice, once for each list. The order of the lists and the semi-randomised presentation f word pairs during retrieval were counterbalanced across participants, resulting in eight possible permutations of list order and pair presentation presentation.

Post- scan: Free recall and Similarity Judgment Tests

Following the scanning session, participants were debriefed and subsequently completed a free recall test along with two similarity judgment tests (SJTs). In

the free recall test, participants were instructed to write down as many words as they could recall from each list, maintaining the correct order of presentation as seen during encoding.

.

The similarity judgment tests consisted of two components: a continuous rating of semantic similarity and a forced-choice selection task. In the continuous rating test, participants rated the semantic similarity of two words displayed on the screen on a scale from 1 to 9. In the second test, participants were shown three words simultaneously—one at the top of the screen and two at the bottom—and were asked to select which of the two bottom words was closest in meaning to the top word. Both tasks were self-paced, with no time limits imposed.

In both SJTs, participants were presented with all possible combinations of words, including similarity test between words across two different lists comparisons between words from different lists (i.e., words that had never been presented or tested together during the initial list-learning task).

Both the word list learning task and the SJTs were programmed using Psychtoolbox 3 for MATLAB (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007).

fMRI acquisition

Structural and functional magnetic resonance imaging (fMRI) data were collected with a 3-Tesla scanner, with a 32-channel head coil. First, we acquired a structural brain scan (T1-weighted image) for each participant using MPRAGE (magnetisation-prepared rapid gradient echo sequence) with parameters set to: TR = 2530 ms, TE = 3.34 ms, flip angle = 7°, FoV = 256mm, voxel size = 1 x 1 x 1 mm^3. We then moved to the acquisition of functional data. To minimise the signal loss in the medial temporal lobe and the orbitofrontal cortex region, set the slice angle at 30° relative to the anterior-posterior commissure line. We used the following parameters for data acquisition: repetition time (TR): 3360 ms, echo time (TE) = 30ms, filp angle =90°, field of view (FoV) = 192mm. 48 slices were acquired in ascending order;

the slices were 2.5mm thick and the distance factor between slices was set to 20%, with voxel sixe = 3 x3 x3 mm^3. Between the two repetitions of the task, we also acquired a whole brain field map with dual echo-time images to correct for signal distortion, with parameters: TR = 1020 ms, TE1 = 10 ms, TE2 = 12.46 ms, flip angle = 90°, FoV = 192mm, voxel size = 3 x 3 x 2 mm^3 . During the task acquisition, we time-locked the stimuli presentation and button presses to the fMRI data.

fMRI pre-processing

The acquired data were first pre-processed using SPM12 (RRID:SCR_007037). Functional imaging data were realigned and co-registered to each subject's structural scan to account for motion correction. A high-pass filter with a cut-off of 128 seconds was applied to address slow signal drift. The data were then normalised to MNI space and smoothed with a 7 mm Gaussian kernel. To facilitate further first and second level whole-brain analysis using FMRIB's Software library (FSL), I re-ran pre-processing data analysis using FSL (Constantinescu et al., 2016). Starting again from the raw MRI data, I removed motion artefacts and applied a high-pass filter at 6 1/100 Hz. Smoothing was performed with a gaussian filter of 7 mm, and I corrected for slice time acquisition differences. I used the acquired field maps to correct the geometric distortions in the EPI images. The EPI images were registered to the structural T-1 acquisition using boundary-based reconstruction, and subsequently normalised into standard space via non-linear registration (Montreal Neurological Institute – MNI152).

fMRI statistical analysis

In the following analyses, uncorrected results are reported with uncorrected cluster threshold of $p<0.001$ for explorative whole brain analyses, while a more lenient threshold of $p<0.01$ is used when there is a strong prior hypothesis for ERH and other small-volume para-hippocampal regions of interest, consistent

with previous studies (see Constantinescu et al. 2016 for further methodological discussion).

Clusters of voxels showing significant uncorrected activity differences in whole brain analyses were corrected for multiple comparisons using family-wise error correction at the cluster level (FWEc). For clusters within the ERH and other small-volume ROIs, multiple comparisons were corrected using family-wise error correction at the peak level (FWEp), accounting for the small volume and the high risk of signal loss in fMRI data collection from these regions (see Constantinescu et al., 2016). Further methodological details are provided in the following sub-sections for each individual analysis.

Effect of temporal distance, semantic distance, and 2-D distance measures

For each subject, I conducted first-level general linear models (GLMs) and second-level group analyses on the retrieval phase of the experiment to investigate the effect of temporal distance (i.e. the distance between words' positions in each list at encoding), semantic distance (i.e. distance between words over the first PCA component), and Euclidean distance between words in the 2D space.

The analysis focused on the 4-seconds period of blank screen follwoing the presentation of each pair of words, during which participants were instructed to consider the positions of the words in the learnt list.. To avoid confounds, epochs corresponding to trials in which the same word repeated twice were excluded from the GLM model for parametric regressor; however, their main effect was still modelled as part of the GLM. I ran three separate GLMs: one with parametric regressors for temporal and semantic distances, one with a parametric regressor for Euclidean distance in the 2-D space, and one with a parametric regressor for cosine distance in the 2-D space. Each GLM was independently ran for each list at the first level, while at the second level the 2 lists were kept as separate conditions for each subjects specifying not-independency in the model.

At the second level, I performed a one-sample t-test across subjects and applied a p<0.001 uncorrected cluster-forming threshold to identify clusters of interest. For the sake of clarity and representation, a more lenient threshold of p<0.005 was also applied for some small ROIs (Figure 4.7).. Correction for multiple comparisons across voxels was performed at the cluster level, with a family-wise error FWE corrected threshold of p=0.05.

Hexadirectional modulation analysis for ROIs

I employed the methodology of Constantinescu et al. (2016) to identify brain regions susceptible to hexadirectional modulation. Using the pre-processed data in FSL, I conducted a first-level general linear model (GLM) analysis for each subject, incorporating parametric regressors for sine and cosine of the angle θ(t) formed by each trajectory between the two words presented in each trial during testing phase, with a 6-fold periodicity (60°): sin(6θ(t)) and cos(6θ(t)). The same GLM also included regressors for the main effects of each phase of the trial (stimuli, blank screen, questions, and response). I first divided the trials in even and odd, and modelled parametric coefficients for each half of the trials separately in the same GLM. This approach aimed to identify brain regions whose activity is modulated by hexagonal symmetry, with these regressors providing coefficients with magnitude $\sqrt{\beta_{sin}^2 + \beta_{cos}^2}$, representing the overall activity modulated by hexadirectional periodicity. To identify the brain areas modulated by a linear combination of the two regressors, $\beta_{sin*\sin(6\theta)} + \beta_{cos*\cos(6\theta)}$, I used an F test. The result was transformed to a Z-statistic from F test for each subject via asymptotic approximation (http://www.fmrib.ox.ac.uk/analysis/techrep/tr00mj1/tr00mj1/). Subsequently, I conducted a one-sample t-test across subjects on the Z-transformed F-scores in each voxel. I then applied a 3.1 cluster-forming threshold on the Z-transformed t-statistic (equivalent to p<0001 uncorrected) to identify clusters that were significantly modulated by the linear combination of cosine and sine regressors with hexagonal symmetry, correcting for multiple comparisons across voxels at the cluster level with a family-wise error FWE corrected threshold of p=0.05. To explore the modulation in the entorhinal

cortex - a region with a strong prior hypothesis of involvement and susceptibility to higher signal loss and limited voxel availability due to its small size and location - I used a more lenient cluster-forming threshold of Z=2.3 (equivalent to $p<0.01$ uncorrected, refer to Constantinescu et al. 2016, supplementary material). Using this method, I identified regions of interest (ROIs) for further GLM analyses to test for hexadirectional modulated activity. The significant clusters of interest where then binarized to obtain ROIs masks to extract the orientation of the grid cell population, as detailed in the subsequent section.

fMRI ROIs analysis

Hexadirectional modulation analyses

To test for hexadirectional modulation of fMRI activity, I first utilised binarized probabilistic entorhinal cortex masks (left and right), thresholded at 20% probability, from the Julich-Brain Cytoarchitectonic Atlas (Amunts et al., 2020). As in the previous analysis, the epoch of interest for each trial (i.e. each pair of words at retrieval) was the four seconds of blank screen following the presentation of the second word in the pair. Trials were divided into odd and even groups. Each epoch of interest was associated with a specific 'movement' between words in the 2-D abstract space. Trialswhere the same word was presented twice were not modelled with parametric regressors, due to their ambiguous direction of 'movement', but were still included in the GLM.)

Following the methodology of Doeller et al. (2010), I conducted two separate GLMs. In the first GLM, for each half of the trials, I modeled parametric regressors for the sine and cosine of θ with a 6-fold periodicity (cos(6θ) and sin(6θ)). This GLM produced two regressor coefficients for each half of the trials, β1 and β2. I then averaged these regressor coefficients within the binarized probabilistic ROIs (right and left entorhinal cortex) and used them to calculate the orientation angle of the grid population: φ = [arctan(β2/β1)] / 6. This process yielded two estimated angles of grid activity, one for each half of the trials.

The direction of the line connecting the words in each trial corresponds to an angle θ specific to the trial. Following the methodology of Doeller et al. (2010), I conducted two separate GLMs. In the first GLM, for each half of the trials, I modelled parametric regressors for sine and cosine of θ in 6-fold periodicity (cos(6θ) and sin(6θ)). This GLM produced two regressor coefficients for each half of the trials, β1 and β2. I then averagedthese regressor coefficients from the binarized probabilistic ROIs (right and left enthorinal cortex), and used them to calculate the angle of orientation of the grid population: φ = [arctan(β2/β1)] / 6. With this step, This process yielded two estimated angles of grid activity, one for each half of the trials. In the second GLM, I used the estimated angles from one half of the trials (φ) to assess the strength of

hexadirectional modulation in the second half, using the parametric regressor: $\cos[6(\theta - \varphi)]$. The output of this second GLM was a regressor coefficient that quantified the degree of hexadirectional modulation of fMRI activity. The GLMs were separately for coefficient extracted from left and right ERH ROIs.The same approach was extended to a whole-brain analysis in to determine whether the activity in any area of the brain was significantly modulated by the grid angle, calculated using the binarized empirical mask previously obtained via the z-F statistic in FSL. This mask was applied in the same manner as the binarized probabilistic masks of the entorhinal cortex to determine the grid population angle.

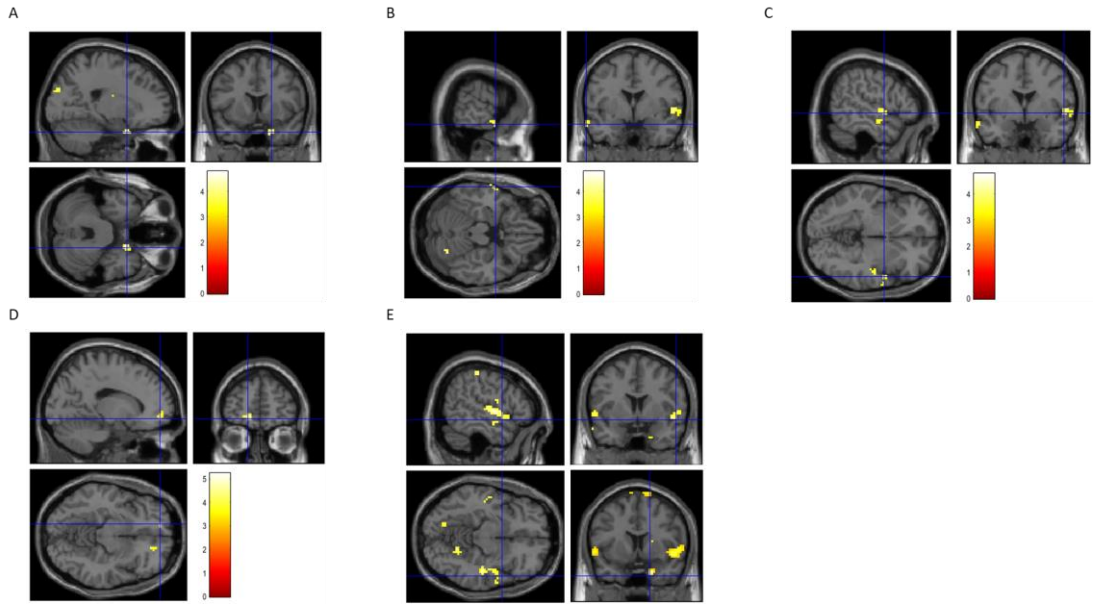


Figure 4.5. Whole brain analysis for distance over the temporal order axis in both lists, with cluster-forming-threshold $p<0.001$ uncorrected, and FWEp (family-wise error peak) corrected threshold $p<0.05$. Activity was found in the right entorhinal cortex (A), left middle temporal gyrus (B) and right superior temporal gyrus (C). These effects were mainly driven by activity in these areas in the mixed list condition (E), while the continuous list condition had significant activity in mPFC (D). Colour bars show the t-statistic. The whole brain analysis revealeda significant effect of temporal distance between words - defined as the difference in the position of the two words within each pair in the encoding list - in the right entorhinal cortex, right superior temporal gyrus and left middle temporal gyrus in both lists (Fig. 4.5 A, B and C). However, when analysed separately, the effect was stronger in mixed lists (condition 2, Fig. 4.5 E) compared to the continuous list (condition 1, Fig. 4.5 D), where the only cluster surviving p<0.001 uncorrected was found in the left mPFC.

From the same GLM, I identified brain activity modulated by the semantic distance between words in each pair, calculated as the difference in position between words along the first PCA component of word2vec. This modulation was observed in the left cerebellum (Fig.4.6 A), right thalamus (Fig. 4.6 B) across both conditions.The Euclidean distance in the 2D space significantly

modulated clusters in the left hippocampus (Fig. 4.7 A) and in both right and left cerebellum (Fig.4.7 B). When examining the lists separately, condition 1, but not condition 2, showed a particularly strong effect of Euclidean distance on BOLD activity in the left hippocampus and left entorhinal cortex (Fig. 4.7 C).

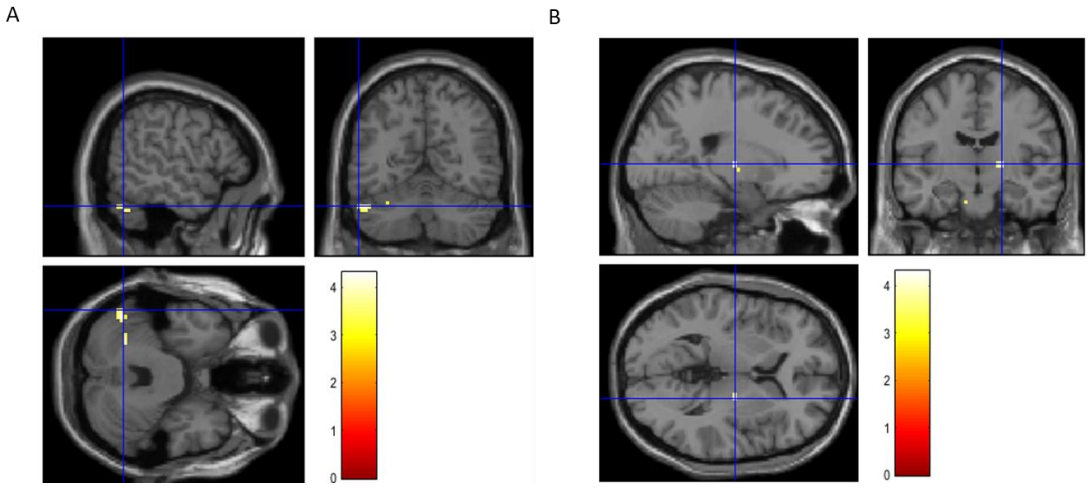


Figure 4.6. Whole brain analysis for distance over the first semantic PCA component axis in both lists, cluster-forming-threshold $p<0.001$, FWEp corrected ($p<0.05$). Activity was found in the left cerebellum (A) and right thalamus (B). No other significant areas of activity modulated by semantic distance regressor were found checking for list types separately. Colour bars show the T-statistic.

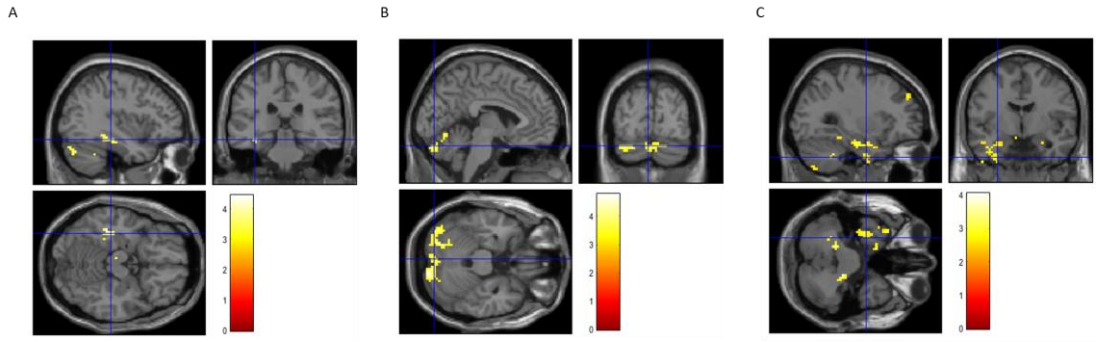


Figure 4.7. Whole brain analysis for Euclidean distance - in 2D (temporal and semantic) space - parametric regressor. For both lists, I found significant cluster of activity in the left hippocampus (A), cluster-forming-threshold $p<0.001$, FWEp corrected ($p<0.05$) and in the cerebellum (B), cluster threshold $p<0.001$, FWEc (family-wise-error cluster) corrected at $p<0.05$. The continuous list alone showed significant activity in the left hippocampus and left entorhinal cortex (C), cluster threshold $p<0.005$, FWEp corrected ($p<0.05$). Colour bars show the t-statistic.

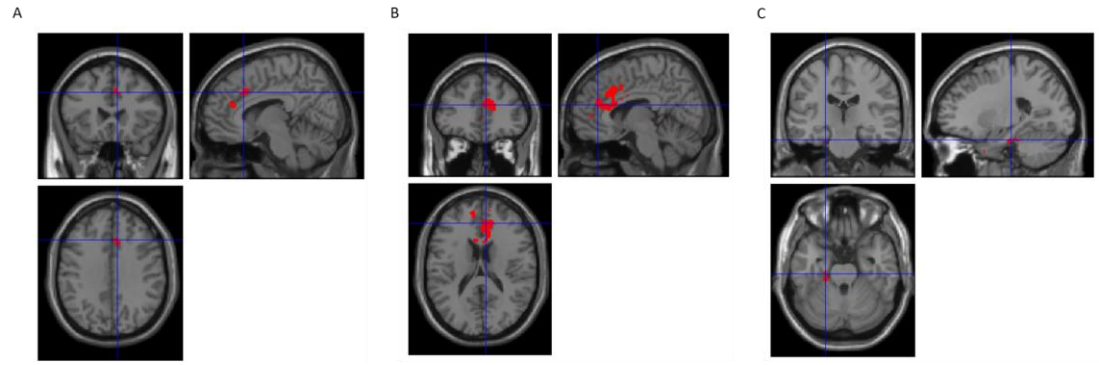


Figure 4.8. Whole brain analysis for hexadirectional modulation using the z-F statistic. A shows the (binarized) cluster corrected at a cluster threshold Z = 3.1 (p=0.001) and p< 0.05 in mPFC (clusters-forming-threshold from Z-statistic = 3.1. Supra-threshold clusters were corrected for family-wise error using a cluster significance threshold of p = 0.05); B shows this at a cluster corrected threshold Z = 2.3 (p=0.01) and p < 0.05; C shows the ERH cluster at cluster forming threshold Z = 2.3, not thresholded at cluster level.

The z-transformed F-statistic test for significant modulation of the linear combination of sin(6θ) and cos(6θ) of the angle θ between words in the 2D abstract space revealed significant clusters across participants and across lists in the mPFC (Fig. 4.8 A and 4.8 B) and in the ERH (Fig.4.8 C). The mPFC cluster was then binarized and used as a mask to extract grid population angles in whole-brain analyses for grid cell activity (Doeller et al. 2010, Constantinescu et al. 2016). The ERH cluster was used to validate the use of anatomical probabilistic ERH masks; however, it was not utilised as a binarized mask due to the small number of voxels included, likely a consequence of signal loss and the small size of the region.

Hexadirectional modulation: Probabilistic (anatomical) mask of left entorhinal cortex

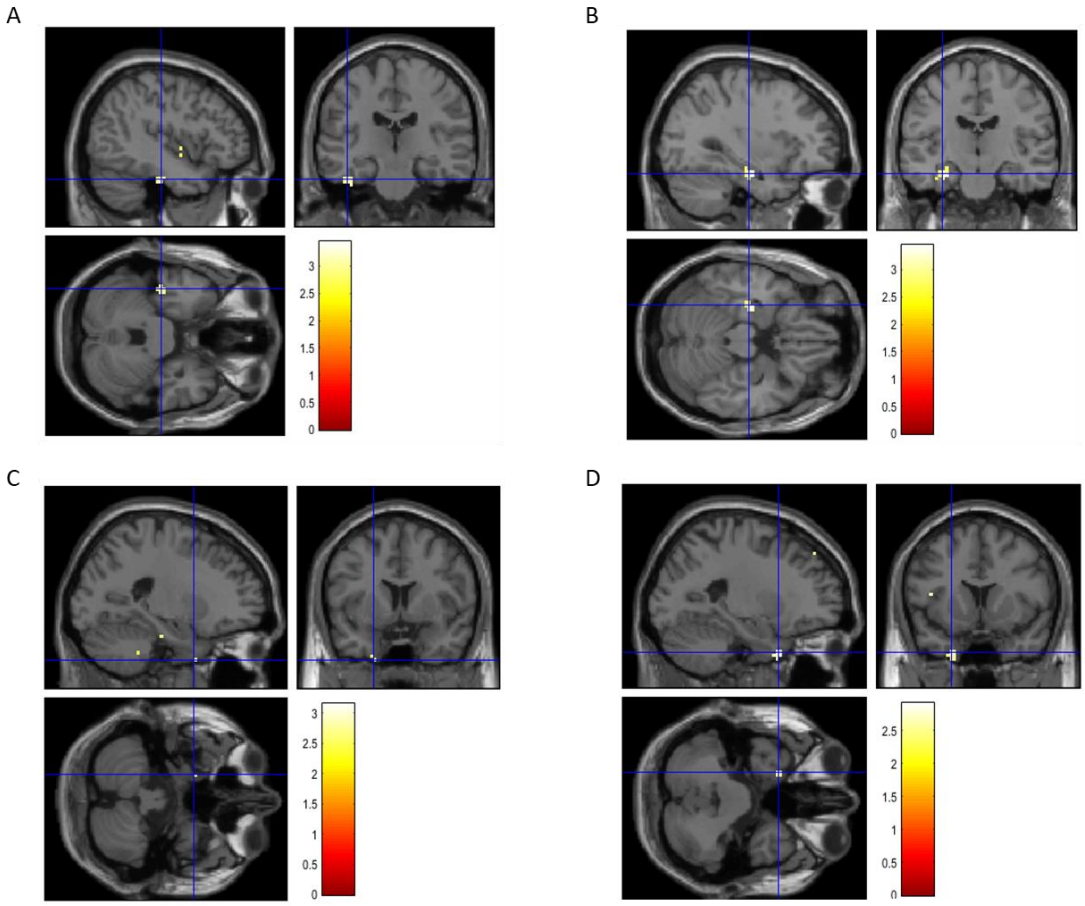


Figure 4.9. Regions showing significant hexadirectional modulation at the whole brain level using grid population angle from the left ERH anatomical mask (Amunts et. Al, 2020). In the continuous list in the left ERH (A) and left HPC (B) (cluster forming threshold p<0.005 FWEp corrected at p <0.05), and in both lists in left ERH (cluster forming threshold p <0.005 in C and p <0.01 in D, not cluster corrected). Colour bars show the t-statistic.

I ran the hexadirectional modulation analysis using binarized probabilistic anatomical masks (Amunts et al., 2020) of the right and left ERH to calculate the grid population angle. In whole-brain analysis, I found clusters of activity significantly modulated by 6-fold symmetry in the left ERH and left HPC in list 1 (Fig.4.9 A, B), and in left ERH across both lists (Fig. 4.9 D, E).

Hexadirectional modulation: Empirical mask of mPFC

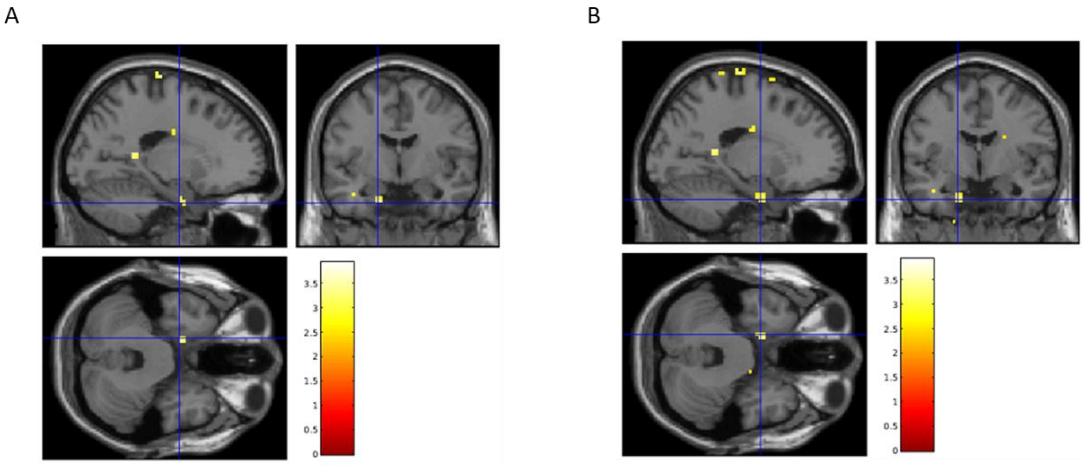


Figure 4.10. Regions showing significant hexadirectional modulation at the whole brain level using grid population angle from empirical binarised mPFC mask at threshold Z=3.1 and FWEc *p* <0.05. In both lists in the left ERH, displayed with *p* <0.001 (A) and *p* <0.005 (B) thresholds, uncorrected. Colour bars show the T-statistic.

The same analysis was ran using the empirical mask of mPFC at threshold 3.1 (Fig. 4.8 A) to calculate the grid angle. Whole-brain analysis found significant clusters of hexadirectionally modulated activity in the left ERH across lists (Fig. 4.10).

Discussion

With this work, I investigated how temporal and semantic domains are integrated within declarative memory. The interaction between temporal and semantic contexts was conceptualised as a specific case of bidimensional space, allowingme to hypothesise and test whether the same brain area and neuronal mechanism fundamental to spatial memory and navigation might also be involved in integrating two different abstract contextual dimensions. While previous studies on navigation in abstract spaces required the participants to undergo task-specific training to establish the two relevant spatial dimensions (Constantinescu et al. 2016, Vigano et al. 2021), this study aimed to investigate whether the similar processes are engaged in multidomain spaces (time and semantics) during a naturalistic task, without the need for a task-specific training. Although preliminary, the results suggest a potential involvement of grid cell populations and 6-fold symmetry modulation during the retrieval of one-dimensional word lists. This work contributes to the growing body of research on abstract navigation and cognitive maps across various domains. Notably, this study is the first to investigate how two separate domains, namely temporal order and word semantics, can be integrated into a unified 2D abstract map of concepts. Interestingly, the preliminary findings indicate hexadirectional modulation of fMRI signal in regions known to play an important role in spatial navigation (ERH, HPC and mPFC) during the recollection of sequences of words. Moreover, this effect appears to be more pronounced in lists of words from the same taxonomical domain with a more obvious one-dimensional order over the size domain (condition one). This may suggest that the involvement of grid cell population and bidimensional organisation of conceptual spaces might appear in the real world – without any previous task-specific training – when the dimensions aresalient for the specific task tested. Additionally, a left-lateralization of most results aligns with the known hemispheric language dominance in right-handed participants. Examining temporal and semantic modulation independently, additional ROIs were identified in the whole-brain analysis. Temporal modulation of activity in the right superior temporal gyrus

aligns with previous findings that highlight this area's role in coding time and durations (Cantarella et al., 2023; Morillon et al., 2009), while the left middle temporal gyrus is involved in reading and speech production. Regarding semantic distance modulation of fMRI activity, activation in the left cerebellum and right hippocampus suggests that the non-dominant brain may support language semantic processing with visuospatial organisational information (on which the first PCA component of semantic distances is built), which is primarily encoded in non-dominant areas (right hippocampus and left cerebellum). Additional representation similarities analyses might be beneficial to further investigate how different measures of semantic similarity and temporal distances affect fMRI brain activity.*Further planned analyses*

The aim of this chapter was to *outline a novel approach to studying grid cell activities outside the domain of spatial navigation*. While the results presented here suggest a possible involvement of hippocampal, parahippocampal areas and mPFC in supporting grid-like hexadirectional modulation during abstract navigation of a multidomain 2D space, *they are not intended to be conclusive*. The analyses and results presented are only the first exploratory steps in a broaderinvestigation and hypothesis testing. To build upon this work, I plan ot conduct additional tests to validate the preliminary findings and better understand the underlying mechanisms.

Further analyses will be required to validate the preliminary results and explore the mechanisms underlying the reported findings. First, the same hexadirectional modulation analysis will be completed to test for other symmetries (4-, 5-, 7- and 8- fold symmetries), and verify that the results are specific for 6-fold (hexadirectional) symmetries. Additionally, I plan to run a GLM to test for on-axes versus off-axes (aligned vs misaligned) modulation of fMRI activity (Doeller et al. 2010, Fig. 3e). This latter analysis aims to determine how the average fMRI signal for ROIs changes for angles aligned with the computed grid population angle versus misaligned angles compared to baseline, thereby enhancing our understanding of the mechanisms underlying the results, if confirmed.

To improve statistical power and refine the model design, I plan to model the preliminary analyses as a unified GLM at the first level for each participant, where each list is modelled as a different condition, rather than running two separate first-level GLMs per subject (one for each list). Moreover, correlation analyses will be conducted to investigate whether the strength of the hexadirectional modulatory effect correlates with learning performance across participants.

The data from the similarity judgment tests (SJTs) can be used to create subject-specific semantic graphs. This approach will allow for the generation of subject-tailored semantic dimensions to replace the generic word2vec first PCA component, enabling the creation of personalized 2D graphs for each participant and the re-running of analyses based on individual semantic structures. These results can then be compared with those derived from the generic word2vec-informed semantic structure.Thus far, all discussed analyses have focused on the retrieval phase of the experiment. However, the encoding phase, during which participants learned the word lists, can also be of interest. In this phase, I aim to observe how learning the words affected the representation similarity between them. By including trails where the same word was repeated twice during retrieval, it will be possible to run representation similarity analysis on both encoding and retrieval data to analyse how the temporal structure provided by the order of words in the list influences the brain's representation of words. Additionally, the same analysis can be used to identify representational similarities based on measures of semantic distance (first principal component, cosine and Euclidean distance in word2vec space), and determine which brain areas are responsible for semantic representation for words from the same and different semantic domains.

If confirmed, these findings represent a substantial and innovative contribution to the study of cognitive maps and abstract navigation in humans, particularly in the context of combining different domains. By using a participant-specific similarity judgment tests to create individual semantic structures, we can better understand whether explicit conscious access to semantic structure of concepts more accurately reflects the map-like organisation of stimuli. By

testing alternative methods to detect hexadirectional modulation (Bellmund et al., 2016), we will also be able to explore how different methodologies might influence thefindings presented here. Interestingly, both empirical masks of the mPFC (Constantinescu et al., 2016) and probabilistic maps of the ERH (Convertino et al., 2023) provided coherent results, suggesting that the observed effects are robust across different approaches.Further studies would be beneficial to investigate the involvement of special coding in real-world semantic and temporal information independentl. The rich nature of natural language processing and the potential of nested temporal structures open the field to the possibility of testing complex hypotheses and a variety of experimental conditions.

In summary, these preliminary findings offer new insight into the role of spatial coding in abstract cognitive maps across domains in real-world conditions. Additionally, they suggest a potential contribution of cell populations and brain structures typically involved in spatial cognition to language processing and mnemonic strategies for linguistic stimuli.

Chapter 5. General Discussion and Future Directions

This work provided a multi-modality approach to the experimental investigation and theoretical conceptualisation of context-bound memory reconstruction and navigation. Thanks to the use of different computational frameworks, MEG and fMRI studies, this thesis challenges the idea of semantic and episodic memory as fully separated aspects of declarative memory; here, I explored the possibility of a more integrated and nuanced system, which makes use of overlapping computational strategies and brain structures to flexibly access information.

In Chapter 2.1, I implemented a modified version of the DRM model and investigated whether behavioural measures of pattern completion could provide better understanding of the mechanism involved in the false memory task. The findings showed that false memory can be understood as pattern completion of lists reflecting influence from both temporal and semantic context. This work proposes an easy behavioural implementation of the DRM paradigm, which can be flexibly manipulated to better understand the underline mechanisms of interference and interaction between episodic and semantic aspects of declarative memory. Then, I developed different computational models of false memory in the frameworks of auto-associative Hopfield dynamics, successor representation and active inference. These models succeed in reproducing the phenomenology of the DRM false memory task in a controlled mechanistic manner. Crucially, at different levels of computational description, these models coherently support the hypothesis of a pattern completion mechanism, typically associated with the hippocampal formation and episodic memory, involved in the processes of integration and interference of semantic and episodic memory. The auto-associative Hopfield network model (Chapter 2.2) reproduces the pattern completion effect driven by neurons in the CA3 area of the hippocampus, re-building the synaptic structure of the area and its mechanism at a neuronal level.

Moving towards a higher level of abstraction, the temporal context model, in its successor representation (SR) form (Chapter 2.3), allows to bring the potential mechanism of integration between continuous time-based associations and discretised pre-acquired semantic knowledge e to the broader field of reward-driven algorithms. This opens to exciting opportunities to further implement SR algorithms to account for the role of different aspects of declarative memory and pre-acquired knowledge in planning and goal-directed behaviour. Finally, I made use of the active inference framework to better understand the false memory effect as Bayesian optimal behaviour (Chapter 2 Appendix 2). Overall, these models provide a multi-perspective approach to an integrated declarative memory system. In future work, I plan to further implement these models to account for different time scales (from item integration to episodes for narrative construction), and to test them with hypothesis -driven experimental work and model fitting.

Thus far this thesis successfully built the foundation to study temporal and semantic contexts in a rigorous, mechanistic manner. However, the most fundamental form of context, i.e. space, provided the key to explore the foundational cognitive processes and underline brain computations that account for memory and navigations in the human brain. I first validated MEG as a powerful non-invasive method to study grid cell activity in humans in theta frequency. Using an MEG VR experimental approach (Chapter 3), we were also able to develop cutting edge computational data analysis to study grid population activity in MEG recording. Moreover, the stability of the grid activity correlated with spatial memory performance in healthy volunteers.

Schizophrenia offers an interesting pathological model to explore the potentially overlapping mechanisms supporting spatial navigation and abstract inference. I verified that individuals with schizophrenia display diminished spatial memory and reduced theta power associated with movement in a virtual spatial navigation task, compared to a matched control group. I also showed for the first time that schizophrenia is associated with a reduced degree of hexadirectional modulation of theta power in the right entorhinal cortex, a phenomenon linked to the stable firing patterns of grid cells. This research marks the inaugural identification of hexadirectional theta modulation

in magnetoencephalography (MEG), expanding upon prior findings of similar patterns observed through various neuroimaging techniques, thereby highlighting the intricate connections between grid cell function, theta oscillations, and spatial cognition.

These findings are also suggestive of a potential causative role of impaired grid coding in inferential mechanisms and relational knowledge impairment seen in neuropsychiatric conditions. This dysfunction could be attributed to instabilities in attractor network dynamics, potentially exacerbated by alterations in receptor densities affecting neural stability and reliance on different learning mechanisms. Although further work is needed to better understand the pathophysiological involvement of navigation systems in a variety of multi-domain cognitive functions in schizophrenia, this work successfully showed how overlapping mechanisms and brain cell populations flexibly account for multiple aspects of declarative memory and can be directly studied with a variety of neuroimaging approaches.

The different explorative computational and experimental approaches previously explored in this work informed the final experimental chapter of this thesis, Chapter 4, where I built an fMRI experiment to investigate how semantic, temporal, and spatial information is processed and integrated into a coherent unified map by brain dynamics. To do so, I built a simple controlled task of sequence memory for word lists, where the word presented in each list were implicitly organised to cover a 2-dimensional abstract map built on semantic and temporal distance coordinates. Unlike previous investigations that necessitated task-specific training for abstract space navigation, this research investigated the naturalistic integration of time and semantics without such prerequisites.

My preliminary findings indicate that grid cell populations and their hexadirectional signal modulation may play a crucial role in the retrieval of sequentially listed words, particularly within semantically coherent domains. This suggests that the cognitive mechanisms underlying spatial navigation may also support the organization of abstract conceptual spaces in a bidimensional multidomain context, even in tasks without explicit training.

Significantly, these observations reveal hexadirectional modulation in brain regions previously linked to spatial navigation (entorhinal cortex, hippocampus, and medial prefrontal cortex) during word sequence recollection. These findings hint at the spontaneous emergence of bidimensional conceptual maps in real-world settings, particularly when task dimensions are intuitively salient. Further work will focus on validating these findings and exploring additional measures to calculate the extent and localisation of grid-like activities in different brain regions in encoding as well as retrieval. Overall, this work positions itself at the forefront of research into cognitive maps and abstract navigation, proposing new hypothesis on how spatial coding principles are applied across diverse cognitive domains.

In this work, I attempted to investigate the computational, cognitive and neuronal mechanisms involved in different types of contexts, spanning across spatial, temporal and semantic contexts. I tried to contribute to implementing our understanding of each individual context and of their interactions in memory. This journey was guided by a particular interest in the role of the hippocampus, entorhinal cortex, parahippocampal areas and medial prefrontal cortex in memory and navigation, which informed both hypotheses and experimental designs, as well as data analysis strategies.

Overall, this thesis proposes alternative approaches, computational mechanisms, and experimental techniques to investigate the role of multi modal contexts in declarative memory, and the integration of traditionally differentiated aspects of it, accounting for the nuanced complexity of overlapping cognitive dynamics in the human brain.

References

Abbott, L. F., & Nelson, S. B. (2000). Synaptic plasticity: taming the beast. Nature Neuroscience, 3(Suppl), 1178-1183.

Adams RA, Bush D, Zheng F et al. Impaired theta phase coupling underlies frontotemporal dysconnectivity in schizophrenia. Brain. 2020; 143(4): 1261-1277.

Adams RA, Napier G, Roiser JP, Mathys C, Gilleen J. Attractor-like dynamics in belief updating in schizophrenia. J Neurosci. 2018; 38(44): 9471-9485.

Aghajan Z, Schuette P, Fields TA, Tran ME, Siddiqui SM, Hasulak NR et al. Theta Oscillations in the Human Medial Temporal Lobe during Real-World Ambulatory Movement. Curr Biol. 2017; 27(24): 3743-3751.

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. Frontiers in Psychology, 4, Article 519.

Amunts K, Mohlberg H, Bludau S, Zilles K. Julich-Brain (2020) A 3D probabilistic atlas of the human brain's cytoarchitecture. Science.; 369(6506): 988-992.

Armstrong K, Avery S, Blackford JU, Woodward N, Heckers S. (2018). Impaired associative inference in the early stage of psychosis. Schizophr Res.; 202: 86-90.

Armstrong K, Kose S, Williams L, Woolard A, Heckers S. (2012) Impaired associative inference in patients with schizophrenia. Schizophr Bull.; 38(3): 622-629.

Armstrong K, Williams LE, Heckers S. (2012). Revised associative inference paradigm confirms relational memory impairment in schizophrenia. Neuropsychology.; 26(4), 451-458.

Aronov D, Nevers R, Tank DW. (2017). Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. Nature.; 543(7647): 719-722.

Attias, H. (2003). Planning by probabilistic inference. In Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics.

Backus, A. R., et al. (2016). Hippocampal-Prefrontal Theta Oscillations Support Memory Integration. Current Biology, 24(5), 450–457. DOI: 10.1016/j.cub.2016.01.023

Baiano M, Perlini C, Rambaldelli G, Cerini R, Dusi N, Bellani M et al. (2008) Decreased entorhinal cortex volumes in schizophrenia. Schizophr Res.; 102(1-3): 171-180.

Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. Neuron, 102(5), 1066–1075.e5. https://doi.org/10.1016/j.neuron.2019.03.034

Barnes GR, Hillebrand A. (2003). Statistical flattening of MEG beamformer images. Human Brain Mapping.; 18(1): 1-12.

Barry, C., Hayman, R., Burgess, N., & Jeffery, K. J. (2007). Experience-dependent rescaling of entorhinal grids. Nature neuroscience, 10(6), 682–684. https://doi.org/10.1038/nn1905

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or Surprise? Frontiers in Psychology, 4. doi:10.3389/fpsyg.2013.00907.

Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL et al. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron.; 100(2): 490-509.

Bellmund, J. L., Deuker, L., Navarro Schröder, T., & Doeller, C. F. (2016). Grid-cell representations in mental simulation. eLife, 5, e17089. https://doi.org/10.7554/eLife.17089

Binte Mohd Ikhsan, S. N., Bisby, J. A., Bush, D., Steins, D. S., & Burgess, N. (2020). EPS mid-career prize 2018: Inference within episodic memory reflects pattern completion. Quarterly Journal of Experimental Psychology, 73(12), 2047-2070. https://doi.org/10.1177/1747021820959797

Bird CM, Burgess N.(2008). The hippocampus and memory: insights from spatial processing. Nat Rev Neurosci.; 9(3): 182-194.

Boggio, P., et al. (2009). Temporal Lobe Cortical Electrical Stimulation during the Encoding and Retrieval Phase Reduces False Memories. PLoS ONE, 4(3), e4959. DOI: 10.1371/journal.pone.0004959

Bohbot, V., Copara, M., Gotman, J. et al. Low-frequency theta oscillations in the human hippocampus during real-world and virtual navigation. Nat Commun 8, 14415 (2017). https://doi.org/10.1038/ncomms14415

Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J. L., Roudi, Y., Moser, E. I., & Moser, M. B. (2013). Grid cells require excitatory drive from the hippocampus. Nature neuroscience, 16(3), 309–317. https://doi.org/10.1038/nn.3311

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. Psychological Review, 113, 201-233.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. Trends in Cognitive Sciences, 16(10), 485-488.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. Journal of General Psychology, 49, 229–240. https://doi.org/10.1080/00221309.1953.9710088

Bragin, A., Jando, G., Nadasdy, Z., Hetke, J., Wise, K., & Buzsaki, G. (1995). Gamma (40-100 Hz) oscillation in the hippocampus of the behaving rat. Journal of Neuroscience, 15, 47-60.

Brandon, M. P., Bogaard, A. R., Libby, C. P., Connerney, M. A., Gupta, K., & Hasselmo, M. E. (2011). Reduction of theta rhythm dissociates grid cell spatial periodicity from directional tuning. Science (New York, N.Y.), 332(6029), 595–599. https://doi.org/10.1126/science.1201652

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. Cognition Psychology, 57(3), 153-178. doi: 10.1016/j.cogpsych.2007.12.002

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. Neuron, 35(4), 625–641. https://doi.org/10.1016/s0896-6273(02)00830-9

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using Grid Cells for Navigation. Neuron, 87(3), 507–520. https://doi.org/10.1016/j.neuron.2015.07.006

Bush, D., Bisby, J. A., Bird, C. M., Gollwitzer, S., Rodionov, R., Diehl, B., McEvoy, A. W., Walker, M. C., & Burgess, N. (2017). Human hippocampal theta power indicates movement onset and distance travelled. Proceedings of the National Academy of Sciences of the United States of America, 114(46), 12297–12302. https://doi.org/10.1073/pnas.1708716114

Buzsaki, G. (1998). Memory consolidation during sleep: a neurophysiological perspective. Journal of Sleep Research, 7(Suppl 1), 17-23.

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nature neuroscience, 16(2), 130–138. https://doi.org/10.1038/nn.3304

Cai, D., Aharoni, D., Shuman, T., et al. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. Nature, 534, 115–118. https://doi.org/10.1038/nature17955

Cantarella G., Vianello G., Vezzadini G., Frassinetti F., Ciaramelli E., Candini M. (2023). Time bisection and reproduction: Evidence for a slowdown of the internal clock in right brain damaged patients. Cortex, Volume 167, 303-317, ISSN 0010-9452, https://doi.org/10.1016/j.cortex.2023.05.024.

Chen, D., Kunz, L., Wang, W., Zhang, H., Wang, W. X., Schulze-Bonhage, A., Reinacher, P. C., Zhou, W., Liang, S., Axmacher, N., & Wang, L. (2018). Hexadirectional Modulation of Theta Power in Human Entorhinal Cortex during Spatial Navigation. Current biology : CB, 28(20), 3310–3315.e4. https://doi.org/10.1016/j.cub.2018.08.029

Chien, H. S., & Honey, C. J. (2020). Constructing and Forgetting Temporal Context in the Human Cerebral Cortex. Neuron, 106(4), 675-686.e11. doi: 10.1016/j.neuron.2020.02.013

Coane, J. H., et al. (2007). False Memory in a Short-Term Memory Task. Experimental Psychology, 54, 62-70. DOI: 10.1027/1618-3169.54.1.62

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science (New York, N.Y.), 352(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Convertino, L., Bush, D., Zheng, F., Adams, R. A., & Burgess, N. (2023). Reduced grid-like theta modulation in schizophrenia. Brain : a journal of neurology, 146(5), 2191–2198. https://doi.org/10.1093/brain/awac416

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. Neural Computation, 5(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Delorme A, Makeig S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods.; 134(1): 9-21.

Dickerson DD, Wolff AR, Bilkey DK. (2010). Abnormal long-range neural synchrony in a maternal immune activation animal model of schizophrenia. J Neurosci.; 30(37): 12424-12431.

Doeller CF, Barry C, Burgess N. (2010). Evidence for grid cells in a human memory network. Nature; 463(7281): 657-661.

Doeller CF, King JA, Burgess N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. Proc Natl Acad Sci U S A.; 105(15): 5915-5920.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. Nature, 463(7281), 657–661. https://doi.org/10.1038/nature08704

Dordek, Y., Soudry, D., Meir, R., & Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. eLife, 5, e10094. https://doi.org/10.7554/eLife.10094

Dragoi, G., Harris, K. D., & Buzsaki, G. (2003). Place representation within hippocampal networks is modified by long-term potentiation. Neuron, 39, 843-853.

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. Nature Reviews Neuroscience, 15, 732–744. https://doi.org/10.1038/nrn3827

Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function?. Neuron, 83(4), 764–770. https://doi.org/10.1016/j.neuron.2014.07.032

Ekman, M., Kusch, S., & de Lange, F. P. (2023). Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. eLife, 12, e78904.

Ekstrom, A. D., & Ranganath, C. (2017). Space, time, and episodic memory: The hippocampus is all over the cognitive map. DOI: 10.1002/hipo.22750

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. Nature, 425(6954), 184–188. https://doi.org/10.1038/nature01964

El-Kalliny, M. M., Wittig, J. H., Sheehan, T. C., et al. (2019). Changing temporal context in human temporal lobe promotes memory of distinct episodes. Nature Communications, 10, 203. https://doi.org/10.1038/s41467-018-08189-4

Ellison-Wright I, Bullmore E. (2009). Meta-analysis of diffusion tensor imaging studies in schizophrenia. Schizophr Res.; 108(1-3): 3-10.

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: spatial navigation and beyond. Nature neuroscience, 20(11), 1504–1513. https://doi.org/10.1038/nn.4656

Fang C., Aronov D., Abbott L.F., Mackevicius E.L. (2023) Neural learning rules for generating flexible predictions and computing the successor representation eLife 12:e80680

Farzanfar, D., Spiers, H. J., Moscovitch, M., & Rosenbaum, R. S. (2023). From cognitive maps to spatial schemas. Nature reviews. Neuroscience, 24(2), 63–79. https://doi.org/10.1038/s41583-022-00655-9

First MB, Williams JB, Spitzer RL, Gibbon M. (2007). Structured clinical interview for DSM-IV-TR axis I disorders, clinical trials version (SCID-CT). New York: Biometrics Research, New York State Psychiatric Institute.

Freeman, J. H., & Steinmetz, A. B. (2011). Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning & Memory (Cold Spring Harbor, N.Y.), 18, 666-677.

Friston, K. J., & Dolan, R. J. (2010). Computational and dynamic models in neuroimaging. NeuroImage, 52(3), 752-765. https://doi.org/10.1016/j.neuroimage.2009.12.068

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. Neuroscience and Biobehavioral Reviews, 68, 862–879.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference, Curiosity and Insight. Neural Computation, 29(10), 2633-2683. doi:10.1162/neco_a_00999.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. Neuroscience and Biobehavioral Reviews, 77, 388-402. doi:10.1016/j.neubiorev.2017.04.009.

Friston, K., & Buzsaki, G. (2016). The functional anatomy of time: What and when in the brain. Trends Cogn Sci, 20(7), 500-511. doi:10.1016/j.tics.2016.05.001.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. Neural Computation, 29, 1–49.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of Neurophysiology, 61, 331.

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M. B. (2004). Spatial representation in the entorhinal cortex. Science (New York, N.Y.), 305(5688), 1258–1264. https://doi.org/10.1126/science.1099901

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences, 2, 493-501.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. Memory & Cognition, 38, 833.

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. Proceedings of the Royal Society B, 285(1891), 20181645. doi: 10.1098/rspb.2018.1645

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. eLife, 6, e17086. doi: 10.7554/eLife.17086

Geerts, J. P., Chersi, F., Stachenfeld, K. L., & Burgess, N. (2020). A general model of hippocampal and dorsal striatal learning and decision making. Proceedings of the National Academy of Sciences of the United States of America, 117(49), 31427-31437. doi: 10.1073/pnas.2007981117

Geerts, J. P., Gershman, S. J., Burgess, N., & Stachenfeld, K. L. (2023). A probabilistic successor representation for context-dependent learning. Psychological Review. doi: 10.1037/rev0000414

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. PLoS Computational Biology, 5(10), e1000532. doi:10.1371/journal.pcbi.1000532. Epub 2009 Oct 9. PMID: 19816557; PMCID: PMC2749218.

George T.M, de Cothi W., Stachenfeld K.L., Barry C. (2023) Rapid learning of predictive maps with STDP and theta phase precession eLife 12:e80663

Gershman, S. J. (2017). Predicting the Past, Remembering the Future. Current Opinion in Behavioral Sciences, 17, 7-13. doi:10.1016/j.cobeha.2017.05.025.

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. Journal of Neuroscience, 38(33), 7193-7200. doi: 10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. Neural Computation, 24(6), 1553–1568. https://doi.org/10.1162/NECO_a_00282

Giari, G., Vignali, L., Xu, Y., & Bottini, R. (2023). MEG frequency tagging reveals a grid-like code during attentional movements. Cell reports, 42(10), 113209. https://doi.org/10.1016/j.celrep.2023.113209

Gil M, Ancau M, Schlesiger MI et al. (2018). Impaired path integration in mice with disrupted grid cell firing. Nat Neurosci.; 21: 81–91.

Glanzer, M. (1969). Distance between related words in free recall: Trace of the STS. Journal of Verbal Learning & Verbal Behavior, 8(1), 105–111. https://doi.org/10.1016/S0022-5371(69)80018-6

Goldwater, S. (2006). Nonparametric Bayesian Models of Lexical Acquisition. Brown University. Retrieved from http://www.stanford.edu/~sgwater/papers/thesis_1spc.pdf.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. Journal of the International Neuropsychological Society, 16(5), 748–753.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. Nature, 436(7052), 801–806. https://doi.org/10.1038/nature03721

Hamid, O. H., Wendemuth, A., & Braun, J. (2010). Temporal context and conditional associative learning. BMC Neuroscience, 11, 45. doi: 10.1186/1471-2202-11-45

Hamm JP, Peterka DS, Gogos JA, Yuste R. (2017). Altered Cortical Ensembles in Mouse Models of Schizophrenia. Neuron.; 94(1): 153-167.

Harrison PJ. (2004). The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. Psychopharmacology (Berl).; 174(1): 151-162.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. Trends in Cognitive Sciences, 11, 299-306.

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364, 1263-1271.

Heckers S, Konradi C. (2002). Hippocampal neurons in schizophrenia. J Neural Transm (Vienna).; 109(5-6): 891-905.

Heckers S. (2001). Neuroimaging studies of the hippocampus in schizophrenia. Hippocampus.; 11(5): 520-528.

Henze, D. A., Borhegyi, Z., Csicsvari, J., Mamiya, A., Harris, K. D., & Buzsáki, G. (2000). Intracellular features predicted by extracellular recordings in the hippocampus in vivo. Journal of neurophysiology, 84(1), 390–400. https://doi.org/10.1152/jn.2000.84.1.390

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks, Science, vol. 268, pp. 1158-1161

Hinton, G. E., & Zemel, R. S. (1993). Autoencoders, minimum description length and Helmholtz free energy. In Proceedings of the 6th International Conference on Neural Information Processing Systems (pp. 3-10). Morgan Kaufmann Publishers Inc.

Hohwy, J. (2016). The Self-Evidencing Brain. Noûs, 50(2), 259-285. doi:10.1111/nous.12062.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Science, 79(8), 2554-2558. DOI: 10.1073/pnas.79.8.2554

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. Journal of Experimental Psychology: General, 142(4), 1370-1383. DOI: 10.1037/a0033626

Horner, A. J., Bisby, J. A., Wang, A., Bogus, K., & Burgess, N. (2016). The role of spatial boundaries in shaping long-term event representations. Cognition, 154, 151–164. https://doi.org/10.1016/j.cognition.2016.05.013

Horner, A., et al. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. Nature Communications, 6, 7462. DOI: 10.1038/ncomms8462

Howard, M. W. (2017). Temporal and spatial context in the mind and brain. Current Opinion in Behavioral Sciences, 17, 14–19.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(4), 923–941. doi: 10.1037//0278-7393.25.4.923

Howard, M. W., & Kahana, M. J. (2001). A Distributed Representation of Temporal Context. Journal of Mathematical Psychology. DOI: 10.1006/jmps.2001.1388

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. Journal of Mathematical Psychology, 46(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? Journal of Memory and Language, 46(1), 85–98. https://doi.org/10.1006/jmla.2001.2798

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. Psychological Review, 112(1), 75-116

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually-changing representation of

temporal context. Topics in Cognitive Science, 3(1), 48–73. doi: 10.1111/j.1756-8765.2010.01112.x

Hsieh, L. T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. Neuron, 81(5), 1165-1178. doi: 10.1016/j.neuron.2014.01.015

Huh, D., & Todorov, E. (2009). Real-time motor control using recurrent neural networks. 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, pp. 42-49.

Jacobs J, Weidemann CT, Miller JF, Solway A, Burke JF, Wei X-X et al. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. Nat Neurosci.; 16(9): 1188-1190.

Johnson, M. K., Kounios, J., & Reeder, J. A. (1994). Time-course studies of reality monitoring and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1409–1419. https://doi.org/10.1037/0278-7393.20.6.1409

Jung MW, Wiener SI, McNaughton BL (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. J Neurosci 14: 7347–7356.

Kahana, E., & Kahana, B. (1996). Conceptual and empirical advances in understanding aging well through proactive adaptation. In V. L. Bengtson (Ed.), Adulthood and aging: Research on continuities and discontinuities (pp. 18–40).

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. Memory & Cognition, 30(6), 841–849. doi: 10.3758/BF03195770

Kahana, M. J., Sekuler, R., Caplan, J. B., Kirschen, M., & Madsen, J. R. (1999). Human theta oscillations exhibit task dependence during virtual maze navigation. Nature, 399(6738), 781–784. https://doi.org/10.1038/21645

Kaplan R, Bush D, Bonnefond M, Bandettini PA, Barnes GR, Doeller CF et al. (2014). Medial prefrontal theta phase coupling during spatial memory retrieval. Hippocampus.; 24(6): 656-665.

Kaplan R, Doeller CF, Barnes GR, Litvak V, Düzel E, Bandettini PA et al. (2012). Movement-related theta rhythm in humans: coordinating self-directed hippocampal learning. PLoS Biol.; 10(2): e1001267.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. Biological Cybernetics, 112(4), 323-343. doi:10.1007/s00422-018-0753-2.

Kay SR, Fiszbein A, Opler LA. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull.; 13(2): 261-276.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2009). Perception and hierarchical dynamics. Frontiers in Neuroinformatics, 3, 20.

Killian, N. J., Jutras, M. J., & Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. Nature, 491(7426), 761–764. https://doi.org/10.1038/nature11587

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., Moser, E. I., & Moser, M. B. (2008). Finite scale of spatial representation in the hippocampus. Science (New York, N.Y.), 321(5885), 140–143. https://doi.org/10.1126/science.1157086

Koenig, J., Linder, A. N., Leutgeb, J. K., & Leutgeb, S. (2011). The spatial periodicity of grid cells is not sustained during reduced theta oscillations. Science (New York, N.Y.), 332(6029), 592–595. https://doi.org/10.1126/science.1201685

Kojima, S., & Goldman-Rakic, P. S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Research, 248, 43-49.

Kraus, B. J., Brandon, M. P., Robinson, R. J., 2nd, Connerney, M. A., Hasselmo, M. E., & Eichenbaum, H. (2015). During Running in Place, Grid Cells Integrate Elapsed Time and Distance Run. Neuron, 88(3), 578–589. https://doi.org/10.1016/j.neuron.2015.09.031

Kunz, L., Schröder, T. N., Lee, H., Montag, C., Lachmann, B., Sariyska, R., Reuter, M., Stirnberg, R., Stöcker, T., Messing-Floeter, P. C., Fell, J., Doeller,

C. F., & Axmacher, N. (2015). Reduced grid-cell-like representations in adults at genetic risk for Alzheimer's disease. Science (New York, N.Y.), 350(6259), 430–433. https://doi.org/10.1126/science.aac8128

Ledoux AA, Boyer P, Phillips JL, Labelle A, Smith A, Bohbot VD. (2014). Structural hippocampal anomalies in a schizophrenia population correlate with navigation performance on a wayfinding task. Front Behav Neurosci.; 8: 88.

Li, S.-c., & Lewandowsky, S. (1993). Intralist distractors and recall direction: Constraints on models of memory for serial order. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(4), 895–908. doi: 10.1037/0278-7393.19.4.895

Li, S.-c., & Lewandowsky, S. (1995). Forward and backward recall: Different retrieval processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(4), 837–847. doi: 10.1037/0278-7393.21.4.837

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13-22.

Lieberman JA, Girgis RR, Brucato G, Moore H, Provenzano F, Kegeles L et al. (2018). Hippocampal dysfunction in the pathophysiology of schizophrenia: a selective review and hypothesis for early detection and intervention. Molecular Psychiatry; 23(8): 1764-1772.

Lisman, J., & Buzsaki, G. (2008). A neural coding scheme formed by the combined function of gamma and theta oscillations. Schizophrenia Bulletin, 34, 974-980.

Lisman, J., & Redish, A. D. (2009). Prediction, sequences, and the hippocampus. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364, 1193-1201.

Litvak V, Mattout J, Kiebel S, Phillips C, Henson R, Kilner J et al. EEG and MEG Data Analysis in SPM8. Computational Intelligence and Neuroscience. 2011; 852961.

M Aghajan, Z., Schuette, P., Fields, T. A., Tran, M. E., Siddiqui, S. M., Hasulak, N. R., Tcheng, T. K., Eliashiv, D., Mankin, E. A., Stern, J., Fried, I., & Suthana,

N. (2017). Theta Oscillations in the Human Medial Temporal Lobe during Real-World Ambulatory Movement. Current biology : CB, 27(24), 3743–3751.e3. https://doi.org/10.1016/j.cub.2017.10.062

Maass, W., Natschlager, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14, 2531-2560.

MacKay, D. J. (1995). Free-energy minimisation algorithm for decoding and cryptoanalysis. Electronics Letters, 31, 445-447.

Maidenbaum, S., Miller, J., Stein, J. M., & Jacobs, J. (2018). Grid-like hexadirectional modulation of human entorhinal theta oscillations. Proceedings of the National Academy of Sciences of the United States of America, 115(42), 10798–10803. https://doi.org/10.1073/pnas.1805007115

Marini, et al. (2012). True and false DRM memories: Differences detected with an implicit task. Frontiers in Psychology, 3, Article 310.

Marques TR et al. (2021). GABA-A receptor differences in schizophrenia: a positron emission tomography study using [11C]Ro154513. Mol Psychiatry.; 26(6): 2616-2625.

Marr, D. (1971). Simple memory: A theory for archicortex. Philosophical Transactions of the Royal Society B, 262(841), 23-81. DOI: 10.1098/rstb.1971.0078

Marr, D. (1982). Vision: A Computational Approach. San Francisco: Freeman & Co.

Martinet, L. E., Sheynikhovich, D., Benchenane, K., & Arleo, A. (2011). Spatial learning and action planning in a prefrontal cortical network model. PLoS Computational Biology, 7, e1002045.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (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), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McGeoch, J. A., & McGeoch, G. O. (1936). Studies in retroactive inhibition. VI. The influence of the relative serial positions of the interpolated synonyms. Journal of Experimental Psychology, 19(1), 1–23. doi: 10.1037/h0060051

McNaughton BL, Battaglia FP, Jensen O, Moser EI, Moser M-B. (2006). Path integration and the neural basis of the 'cognitive map'. Nat Rev Neurosci.; 7(8): 663-78.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene Construction, Visual Foraging, and Active Inference. Frontiers in Computational Neuroscience, 10, 56. doi:10.3389/fncom.2016.00056.

Mirza, M. B., Adams, R. A., Parr, T., & Friston, K. J. (2018). Impulsivity and Active Inference. Journal of Cognitive Neuroscience, 31(2), 202–220.

Mohammadi A, Hesami E, Kargar M, Shams J. (2018). Detecting allocentric and egocentric navigation deficits in patients with schizophrenia and bipolar disorder using virtual reality. Neuropsychol Rehabil.; 28(3): 398-415.

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. Nature communications, 10(1), 5685. https://doi.org/10.1038/s41467-019-13760-8

Mok, R. M., & Love, B. C. (2023). A multilevel account of hippocampal function in spatial and concept learning: Bridging models of behavior and neural assemblies. Science advances, 9(29), eade6903. https://doi.org/10.1126/sciadv.ade6903

Monfils, M. H., & Teskey, G. C. (2004). Induction of long-term depression is associated with decreased dendritic length and spine density in layers III and V of sensorimotor neocortex. Synapse, 53, 114-121.

Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Three stages and four neural systems in time estimation. The Journal of neuroscience : the official journal

of the Society for Neuroscience, 29(47), 14803–14811. https://doi.org/10.1523/JNEUROSCI.3222-09.2009

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. Journal of Memory and Language, 86, 119–140. doi: 10.1016/j.jml.2015.10.002

Morton, N. W., et al. (2021). Semantic knowledge of famous people and places is represented in the hippocampus and distinct cortical networks. Journal of Neuroscience. DOI: 10.1523/JNEUROSCI.2034-19.2021

Moser, E., Roudi, Y., Witter, M. et al. (2014). Grid cells and cortical representation. Nat Rev Neurosci 15, 466–481. https://doi.org/10.1038/nrn3766

Moser, M. B., Rowland, D. C., & Moser, E. I. (2015). Place cells, grid cells, and memory. Cold Spring Harbor perspectives in biology, 7(2), a021808. https://doi.org/10.1101/cshperspect.a021808

Muller, R. U., & Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. The Journal of neuroscience : the official journal of the Society for Neuroscience, 7(7), 1951–1968. https://doi.org/10.1523/JNEUROSCI.07-07-01951.1987

Murdock, B. B., Jr. (1962). The serial position effect of free recall. Journal of Experimental Psychology, 64(5), 482–488. doi: 10.1037/h0045106

Nadasdy, Z., Nguyen, T. P., Török, Á., Shen, J. Y., Briggs, D. E., Modur, P. N., & Buchanan, R. J. (2017). Context-dependent spatially periodic activity in the human entorhinal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), E3516–E3525. https://doi.org/10.1073/pnas.1701352114

Naim, M., et al. (2020). Fundamental Law of Memory Recall. Physical Review Letters, 124, 018101. DOI: 10.1103/PhysRevLett.124.018101

O'Keefe J. & Nadel L. (1978) The Hippocampus as a Cognitive Map, Oxford University Press.

O'Keefe J. (1976). Place units in the hippocampus of the freely moving rat. Experimental neurology, 51(1), 78–109. https://doi.org/10.1016/0014-4886(76)90055-8

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain research, 34(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

O'Keefe, J., & Nadel, L. (1979). Précis of O'Keefe and Nadel's The Hippocampus as a Cognitive Map.. Behavioral and Brain Sciences, 2(4), 487–533. https://doi.org/10.1017/S0140525X00063949

Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Computational Intelligence and Neuroscience. 2011; 156869.

Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. Frontiers in Neurorobotics, 1, 6.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22–27. DOI: 10.1016/j.jbef.2017.12.004

Park, A. J., Harris, A. Z., Martyniuk, K. M., Chang, C. Y., Abbas, A. I., Lowes, D. C., Kellendonk, C., Gogos, J. A., & Gordon, J. A. (2021). Reset of hippocampal-prefrontal circuitry facilitates learning. Nature, 591(7851), 615–619. https://doi.org/10.1038/s41586-021-03272-1

Parr, T. (2019). 'The computational neurology of active vision', UCL (University College London).

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. Nature Scientific Reports, 7, 14678.

Parr, T., et al. (2020). Prefrontal computation as active inference. Cereb Cortex, 30(2), 682-695.

Parr, T., Rees, G., & Friston, K. J. (2018). Computational neuropsychology and Bayesian inference. Front Hum Neurosci, 12, 61.

Parr, T., Rikhye, R. V., Halassa, M. M., & Friston, K. J. (2019). Prefrontal Computation as Active Inference. Cerebral Cortex. doi:10.1093/cercor/bhz118.

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

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. Nature Communications, 12, 6587. https://doi.org/10.1038/s41467-021-26731-9

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. Science, 310(5756), 1963–1966. doi: 10.1126/science.1117645

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. Psychological review, 116(1), 129–156. https://doi.org/10.1037/a0014420

Prasad KM, Patel AR, Muddasani S, Sweeney J, Keshavan MS. (2004). The entorhinal cortex in first-episode psychotic disorders: a structural magnetic resonance imaging study. Am J Psychiatry.; 161(9): 1612-1619.

Pu, Y., Kong, X. Z., Ranganath, C., et al. (2022). Event boundaries shape temporal organization of memory by resetting temporal context. Nature Communications, 13, 622. https://doi.org/10.1038/s41467-022-28216-9

Raithel, C. U., Miller, A. J., Epstein, R. A., Kahnt, T., & Gottfried, J. A. (2023). Recruitment of grid-like responses in human entorhinal and piriform cortices by odor landmark-based navigation. Current biology : CB, 33(17), 3561–3570.e4. https://doi.org/10.1016/j.cub.2023.06.087

Ranjan, S., & Odegaard, B. (2024). Reality monitoring and metacognitive judgments in a false-memory paradigm. *Neuroscience research*, *201*, 3–17. https://doi.org/10.1016/j.neures.2023.11.007

Raskin, E., & Cook, S. A. (1937). The strength and direction of associations formed in the learning of nonsense syllables. Journal of Experimental Psychology, 20(4), 381–395. doi: 10.1037/h0061612

Ravassard, P., Pachoud, B., Comte, J.C., Mejia-Perez, C., Scoté-Blachon, C., Gay, N., Claustrat, B., Touret, M., Luppi, P.H., & Salin, P.A. (2009). Paradoxical (REM) sleep deprivation causes a large and rapidly reversible decrease in long-term potentiation, synaptic transmission, glutamate receptor protein levels, and ERK/MAPK activation in the dorsal hippocampus. Sleep, 32, 227-240.

Ribeiro, S., Mello, C. V., Velho, T., Gardner, T. J., Jarvis, E. D., & Pavlides, C. (2002). Induction of hippocampal long-term potentiation during waking leads to increased extrahippocampal zif-268 expression during ensuing rapid-eye-movement sleep. Journal of Neuroscience, 22, 10914-10923.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2(1), 79–87. https://doi.org/10.1038/4580

Roalf DR, Quarmley M, Calkins ME, Satterthwaite TD, Ruparel K, Elliott MA et al. (2016). Temporal Lobe Volume Decrements in Psychosis Spectrum Youths. Schizophr Bull.; 43(3): 601-610.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(4), 803–814. doi: 10.1037/0278-7393.21.4.803

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. Psychological Science, 4, 28–34.

Rotenberg A., Mayford M., Hawkins R.D., Kandel E.R., Muller R.U. (1996). Mice Expressing Activated CaMKII Lack Low Frequency LTP and Do Not Form Stable Place Cells in the CA1 Region of the Hippocampus. Cell, Volume 87, Issue 7, Pages 1351-1361, ISSN 0092-8674. https://doi.org/10.1016/S0092-8674(00)81829-2.

Rubin, A., Geva, N., Sheintuch, L., & Ziv, Y. (2015). Hippocampal ensemble dynamics timestamp events in long-term memory. eLife, 4, e12247.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2021). Neural evidence for the successor representation in choice evaluation. bioRxiv 2021.08.29.458114; doi: https://doi.org/10.1101/2021.08.29.458114

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Computational Biology, 13(9), e1005768. https://doi.org/10.1371/journal.pcbi.1005768

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2013). Learning with hierarchical-deep models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1958-1971.

Salgado-Pineda P, Landin-Romero R, Portillo F, Bosque C, Pomes A, Spanlang B et al. (2016). Examining hippocampal function in schizophrenia using a virtual reality spatial navigation task. Schizophr Res.; 172(1-3), 86-93.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M. B., & Moser, E. I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. Science (New York, N.Y.), 312(5774), 758–762. https://doi.org/10.1126/science.1125572

Schacter, D. L. (2012). Constructive memory: Past and future. Dialogues in Clinical Neuroscience, 14(1), 7-18.

Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). IEEE Transactions on Autonomous Mental Development, 2(3), 230-247.

Schwartenbeck, P., & Friston, K. (2016). Computational phenotyping in psychiatry: A worked example. eNeuro, 3(4), 0049-16.2016.

Shine, J. M., Müller, E. J., Munn, B., Cabral, J., Moran, R. J., Breakspear, M. (2021). Computational models link cellular mechanisms of neuromodulation to

large-scale neural dynamics. Nature Neuroscience, 24(6), 765-776. https://doi.org/10.1038/s41593-021-00824-6

Sigurdsson T, Stark KL, Karayiorgou M, Gogos JA, Gordon JA. (2010). Impaired hippocampal-prefrontal synchrony in a genetic mouse model of schizophrenia. Nature.; 464(7289): 763-767.

Smith, D. E., Moore, I. L., & Long, N. M. (2022). Temporal Context Modulates Encoding and Retrieval of Overlapping Events. Journal of Neuroscience, 42(14), 3000-3010. doi: 10.1523/JNEUROSCI.1091-21.2022

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. Journal of Mathematical Psychology, 107, 102632. https://doi.org/10.1016/j.jmp.2021.102632

Smith, T. A., Hasinski, A. E., & Sederberg, P. B. (2013). The context repetition effect: predicted events are remembered better, even when they don't happen. Journal of Experimental Psychology: General, 142(4), 1298-1308. doi: 10.1037/a0034067

Solomon, E. A., Lega, B. C., Sperling, M. R., & Kahana, M. J. (2019). Hippocampal theta codes for distances in semantic and temporal spaces. Proceedings of the National Academy of Sciences of the United States of America, 116(48), 24343–24352. https://doi.org/10.1073/pnas.1906729116

Solomon, E.A., Kragel, J.E., Gross, R. et al. (2018). Medial temporal lobe functional connectivity predicts stimulation-induced theta power. Nat Commun 9, 4437. https://doi.org/10.1038/s41467-018-06876-w

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. Nature Neuroscience, 20(11), 1643-1653. doi: 10.1038/nn.4650

Stadler, M.A., Roediger, H.L. & McDermott, K.B. (1999). Norms for word lists that create false memories. Memory & Cognition 27, 494–500. https://doi.org/10.3758/BF03211543

Staresina, B. P., & Wimber, M. (2019). A Neural Chronometry of Memory Recall. Trends in Cognitive Sciences, 23(12), 1071-1085. DOI: 10.1016/j.tics.2019.09.011

Staudigl, T., Leszczynski, M., Jacobs, J., Sheth, S. A., Schroeder, C. E., Jensen, O., & Doeller, C. F. (2018). Hexadirectional Modulation of High-Frequency Electrophysiological Activity in the Human Anterior Medial Temporal Lobe Maps Visual Space. Current biology : CB, 28(20), 3325–3329.e4. https://doi.org/10.1016/j.cub.2018.09.035

Stensola, H., Stensola, T., Solstad, T. et al. (2012). The entorhinal grid map is discretized. Nature 492, 72–78. https://doi.org/10.1038/nature11649

Stoewer, P., Schlieker, C., Schilling, A., et al. (2022). Neural network based successor representations to form cognitive maps of space and language. Scientific Reports, 12, 11233. https://doi.org/10.1038/s41598-022-14916-1

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

Tennant SA, Fischer L, Garden DLF, Gerlei KZ, Martinez-Gonzalez C, McClure C et al. (2018). Stellate Cells in the Medial Entorhinal Cortex Are Required for Spatial Learning. Cell Rep.; 22(5): 1313-1324.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. Curr Opin Neurobiol, 37, 99-105.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. Behavioral Neuroscience, 100, 147-154.

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. Current biology : CB, 29(7), 1226–1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Theves, Stephanie & Fernández, Guillén & Doeller, Christian. (2020). The Hippocampus Maps Concept Space, Not Feature Space. The Journal of Neuroscience. 40. JN-RM. 10.1523/JNEUROSCI.0494-20.2020.

Tolman, E. C. (1948). Cognitive maps in rats and men. Psychological Review, 55(4), 189–208. https://doi.org/10.1037/h0061626

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson, Organization of Memory. Academic Press.

Tulving, E. (1983) Elements of Episodic Memory. Oxford University Press, Oxford.

Tulving, E. (1995). Organization of memory: Quo vadis? In M. S. Gazzaniga (Ed.), The Cognitive Neurosciences (p. 839–853). The MIT Press.

Tulving, E. (2002). Episodic Memory: From Mind to Brain. Annual Review of Psychology, 53(1), 1-25.

Ulanovsky, N., & Moss, C. F. (2007). Hippocampal cellular and network activity in freely moving echolocating bats. Nature neuroscience, 10(2), 224–233. https://doi.org/10.1038/nn1829

van Strien, N. M., Cappaert, N. L., & Witter, M. P. (2009). The anatomy of memory: an interactive overview of the parahippocampal-hippocampal network. Nature reviews. Neuroscience, 10(4), 272–282. https://doi.org/10.1038/nrn2614

Viganò S, Rubino V, Di Soccio A, Buiatti M, Piazza M. (2021). Grid-like and distance codes for representing word meaning in the human brain, NeuroImage, Volume 232, 117876, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2021.117876.

Wang, F., & Diana, R. A. (2017). Temporal context in human fMRI. Current Opinion in Behavioral Sciences, 17, 57–64. https://doi.org/10.1016/j.cobeha.2017.06.004

Wang, X.-J., Hu, H., Huang, C., Kennedy, H., Li, C. T., Logothetis, N., Lu, Z.-L., Luo, Q., Poo, M.-m., Tsao, D., Wu, S., Wu, Z., Zhang, X., Zhou, D. (2020). Computational neuroscience: A frontier of the 21st century. National Science Review, 7(9), 1418–1422. https://doi.org/10.1093/nsr/nwaa129

Warren, et al. (2014). False Recall Is Reduced by Damage to the Ventromedial Prefrontal Cortex: Implications for Understanding the Neural Correlates of Schematic Memory. Journal of Neuroscience, 34(22), 7677–7682.

Weinberger DR, Berman KF, Suddath R, Torrey EF. (1992). Evidence of dysfunction of a prefrontal-limbic network in schizophrenia: a magnetic resonance imaging and regional cerebral blood flow study of discordant monozygotic twins. Am J Psychiatry.; 149(7): 890-897.

White, L. M. (1995). Temporal Difference Learning: Eligibility Traces and the Successor Representation for Actions. PhD thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

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

Wilkins LK, Girard TA, Christensen BK, King J, Kiang M, Bohbot VD. (2019). Spontaneous spatial navigation circuitry in schizophrenia spectrum disorders. Psychiatry Res.; 278: 125-128.

Wilkins LK, Girard TA, Konishi K, King M, Herdman KA, King J et al. (2013). Selective deficit in spatial memory strategies contrast to intact response strategies in patients with schizophrenia spectrum disorders tested in a virtual navigation task. Hippocampus.; 23(11): 1015-1024.

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. eLife, 8, e49547. https://doi.org/10.7554/eLife.49547

Winn, J., & Bishop, C. M. (2005). Variational message passing. Journal of Machine Learning Research, 6, 661-694.

Winter SS, Mehlman ML, Clark BJ, Taube JS. (2015). Passive Transport Disrupts Grid Signals in the Parahippocampal Cortex. Curr Biol.; 25(19): 2493-2502.

Winter, S. S., Clark, B. J., & Taube, J. S. (2015). Spatial navigation. Disruption of the head direction cell network impairs the parahippocampal grid cell signal. Science (New York, N.Y.), 347(6224), 870–874. https://doi.org/10.1126/science.1259591

Yartsev, M. M., & Ulanovsky, N. (2013). Representation of three-dimensional space in the hippocampus of flying bats. Science (New York, N.Y.), 340(6130), 367–372. https://doi.org/10.1126/science.1235338

Zeidman, P., Lutti, A., & Maguire, E. A. (2015). Investigating the functions of subregions within anterior hippocampus. Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 73, 240-256.

Zeithamova, D., et al. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron, 75(1), 168–179. DOI: 10.1016/j.neuron.2012.05.010

Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., Xu, W., Yang, N., Danko, T., Chen, L., Wernig, M., & Südhof, T. C. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron, 78(5), 785–798. https://doi.org/10.1016/j.neuron.2013.05.029

Zhou, C. Y., Talmi, D., Daw, N., & Mattar, M. G. (2023). Episodic retrieval for model-based evaluation in sequential decision tasks. https://doi.org/10.31234/osf.io/3sqjh.