Learning to Imagine: Generative Models of Memory Construction and Consolidation

Author: Eleanor Spens
Supervisor: Prof. Neil Burgess

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

Institute of Cognitive Neuroscience
University College London

Declaration of authorship

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

Abstract

Episodic memory is the (re)construction of an experience rather than the retrieval of a copy; memories involve schema-based predictions, show classic patterns of distortion, and share neural substrates with imagination. Brains need to make predictions to survive, and to achieve this must extract statistical structure from experience. Generative neural networks provide a mechanism for learning this by 'prediction error' minimisation. I explore how the brain develops generative models through memory consolidation, how these models reconstruct experiences during memory 'retrieval', and how they support other cognitive functions.

First I present a computational model in which episodic memories are initially encoded in the hippocampus (a modern Hopfield network), then replayed to train a neocortical generative network (variational autoencoder) to (re)create sensory experiences via latent variable representations. Using images, I simulate how this generative network supports episodic memory, semantic memory, imagination, and inference. The network can reconstruct scenes from partial inputs according to learned schemas (which produces gist-based distortions) and imagine novel scenes consistent with those schemas. I also show how unique and predictable elements of memories could be stored and reconstructed by efficiently combining both hippocampal and neocortical systems, optimising the use of limited hippocampal storage.

I then extend the model to sequential stimuli, with the generative networks trained not only to reconstruct their own inputs, but to predict the next input during replay. I apply this model to statistical learning, relational inference, and planning tasks, consider memory distortions in narratives, and explore 'retrieval augmented generation' as a model of hippocampal-neocortical interaction during recall. Finally, I address the question of continual learning, and suggest that generative replay may stabilise existing memories as new ones are assimilated into the generative model.

In conclusion, I explore how replayed memories update a generative, or predictive, model of the world, which supports multiple cognitive functions.

Impact statement

One might assume that our memories accurately record the details of our experiences, and that remembering is like retrieving a file from a mental 'filing cabinet'. But in reality, memories are active reconstructions of what happened, which show classic patterns of distortion, are influenced by our beliefs about the world, and undergo many changes after their encoding (through a process known as consolidation). Brains need to make predictions to survive, e.g. to predict that food can be found in a certain location, or to predict the presence of a predator from a distant sound. To learn these complex correlations between different stimuli, biological intelligence needs a way to extract statistical structure from experience. Generative neural networks, such as those used in modern machine learning, provide a mechanism for this.

In this thesis, I explore how the brain uses memories to develop generative models of experience through consolidation, and in turn how these generative models help to reconstruct experiences during memory 'retrieval'. After episodic memories are initially encoded in the hippocampus, they are replayed in fast-forward during rest and sleep, and I simulate how this process helps a larger neocortical network learn to make predictions. I then simulate how this large generative network supports episodic memory (for experiences), semantic memory (for facts), imagination, and inference.

This research is broadly relevant to the neuroscience of learning, memory, and imagination, providing a unified view of memory construction and consolidation, in which consolidation is a process of 'learning to imagine', and remembering involves imagining the past. It also demonstrates how certain recent advances in machine learning (e.g. large language models) can be applied to questions in cognitive neuroscience.

In addition, a better understanding of biological learning and memory can inspire improvements in artificial intelligence. For example, modelling how the brain achieves lifelong learning may shed light on how machine learning systems can avoid catastrophic forgetting. Similarly, understanding how hippocampal and neocortical

memory systems work together may contribute towards more sophisticated memory in artificial intelligence. This area of research is also relevant to understanding neurological and psychiatric symptoms and conditions. For example, I model the effects of damage to different regions of the network, touch on how rumination and other forms of maladaptive imagination may relate to the proposed model, and discuss the individual differences that may arise from hypo-priors and hyper-priors.

Acknowledgements

First of all, thank you to Neil - I can't imagine a better supervisor. Thanks to my second supervisor, Tim, and to my thesis committee members, Brad and Benedetto, for their advice and support too. I'm also grateful to everyone in the Space and Memory Lab, from whom I've learned so much over the course of the PhD.

Thanks to my colleagues in my previous job who had to endure a computational neuroscience update while getting coffee nearly every day for years, and to my former managers for their encouragement.

Thank you to Mum, Dad, Ralph, Jo, Thomas, Luke, Rebecca, Jason, Paul, Barbara, Ann, James, Kelsey, and Zorro. Particular thanks go to Mum and Dad, whose work triggered my interest in the mechanisms underlying imagination, and to Ralph for being my first example of academic life - chatting about your PhD project around the dinner table is a great childhood memory, and inspired me to do one myself. Most of all, thank you to Brian.

UCL Research Paper Declaration Form: Referencing the doctoral candidate's own published work(s)

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?**
   A generative model of memory construction and consolidation

   (b) **Please include a link to or DOI for the work:**
   https://www.nature.com/articles/s41562-023-01799-z

   (c) **Where was the work published?**
   Nature Human Behaviour

   (d) **Who published the work?**
   Springer Nature

   (e) **When was the work published?**
   19 January 2024

   (f) **List the manuscript's authors in the order they appear on the publication:**
   Eleanor Spens and Neil Burgess

   (g) **Was the work peer reviewd?**
   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.**
   https://www.biorxiv.org/content/10.1101/2023.01.19.524711v3

   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?**
   (b) **Has the manuscript been uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please please give a link or DOI:**
   (c) **Where is the work intended to be published?**
   (d) **List the manuscript's authors in the intended authorship order:**
   (e) **Stage of publication:**
3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Eleanor Spens and Neil Burgess designed the research and wrote the paper. Eleanor Spens did the computational modelling.
4. **In which chapter(s) of your thesis can this material be found?**
   The start of Chapter One and most of Chapter Two

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: Eleanor Spens
Date: 12 March 2024

Supervisor/Senior Author (where appropriate)**:** Neil Burgess
Date: 12 March 2024

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Episodic memory involves autobiographical experiences in their spatiotemporal context, whereas semantic memory involves factual knowledge (Tulving, 1985). The former is thought to rapidly encode events in the hippocampus, enabling the latter to learn statistical regularities in the neocortex (Marr, 1970, 1971; McClelland et al., 1995; Teyler & DiScenna, 1986). Crucially, episodic memory is constructive; recall is the (re)construction of a past experience, rather than the retrieval of a copy (Bartlett, 1932; Schacter, 2012). But the mechanisms behind episodic (re)construction, and the link to semantic memory, are not well understood.

Old memories can be preserved after hippocampal damage despite amnesia for recent ones (Scoville & Milner, 1957), suggesting that memories initially encoded in the hippocampus end up being stored in neocortical areas, an idea known as 'systems consolidation' (Squire & Alvarez, 1995). The standard model of systems consolidation involves transfer of information from the hippocampus to neocortex (Alvarez & Squire, 1994; Marr, 1970, 1971; McClelland et al., 1995), whereas other views suggest that episodic and semantic information from the same events can exist in parallel (Nadel & Moscovitch, 1997). Hippocampal 'replay' of patterns of neural activity during rest (Diba & Buzsáki, 2007; Wilson & McNaughton, 1994) is thought to play a role in consolidation (Ego-Stengel & Wilson, 2010; Girardeau et al., 2009). However,

consolidation does not just change which brain regions support memory traces; it also converts them into a more abstract representation (Norman et al., 2021; Winocur & Moscovitch, 2011).

Generative models capture the probability distributions underlying data, enabling the generation of realistic new items by sampling from these distributions. This thesis proposes that consolidated memory takes the form of a generative network, trained to capture the statistical structure of stored events by learning to reproduce them (see also Káli & Dayan, 2000, 2002). As consolidation proceeds, the generative network supports both the recall of 'facts' (semantic memory), and the reconstruction of experience from these 'facts' (episodic memory) via the hippocampal formation, in conjunction with additional information from the hippocampal trace that becomes less necessary as training progresses.

This builds on existing models of spatial cognition in which recall and imagination of scenes involve the same neural circuits (Becker & Burgess, 2000; Bicanski & Burgess, 2018; Byrne et al., 2007), and is supported by evidence from neuropsychology that damage to the hippocampal formation (HF) leads to deficits in imagination (Hassabis et al., 2007), episodic future thinking (Schacter et al., 2017), dreaming (Spanò et al., 2020), and daydreaming (McCormick et al., 2018), as well as by neuroimaging evidence that recall and imagination involve similar neural processes (Addis et al., 2007; Hassabis & Maguire, 2007).

I model consolidation as the training of a generative model by an initial autoassociative encoding of memory, through 'teacher-student learning' (Hinton et al., 2015) during hippocampal replay (see also Sun et al., 2021). Recall after consolidation has occurred is a generative process, mediated by schemas representing common structure across events, as are other forms of scene construction or imagination (Arbib, 2020). This model builds on research into the relationship between generative models and consolidation (Káli & Dayan, 2000, 2002), on the use of variational autoencoders to model the hippocampal formation (Nagy et al., 2020; Van de Ven et al., 2020; Whittington et al., 2020), and on the view that abstract allocentric latent variables are learned from egocentric sensory representations in spatial cognition (Bicanski &

Burgess, 2018).

More generally, this work builds on the idea that the memory system learns schemas which encode 'priors' for the reconstruction of input patterns (Fayyaz et al., 2022; Hemmer & Steyvers, 2009). Unpredictable aspects of experience need to be stored in detail for further learning, while fully predicted aspects do not, consistent with the idea that memory helps to predict the future (Bein et al., 2021; Biderman et al., 2020; Schacter et al., 2007; Sherman et al., 2022). I suggest that familiar components are encoded in the autoassociative network as concepts (relying on the generative network for reconstruction), whilst novel components are encoded in greater sensory detail. This is efficient in terms of memory storage (Barlow et al., 1961; Barlow, 1989; Benna & Fusi, 2021), and reflects the fact that consolidation can be a gradual transition, during which the autoassociative network supports aspects of memory not yet captured by the generative network. In other words, the generative network can reconstruct predictable aspects of an event from the outset based on existing schemas, but as consolidation progresses the network updates its schemas to reconstruct the event more accurately, until the formerly unpredictable details stored in HF are no longer required.

After presenting the static version of the model in Chapter Two, I extend it to sequential stimuli in Chapter Three, allowing the simulation of statistical learning, planning, memory for narratives, and more complex inference tasks. In Chapter Four, I demonstrate how the proposed system might avoid catastrophic forgetting by combining the consolidation of recent memories with the stabilisation of old ones through 'generative replay'. Beyond continual learning, generative replay can improve the ability to draw novel inferences, but under some conditions it can lead to a vicious cycle of model degeneration, in which statistical biases are reinforced over time.

The first section of this introduction outlines findings in neuroscience and psychology that the subsequent work attempts to explain, in particular the constructive nature of memory, its neural substrates, and its link to consolidation. The second part reviews mathematical and computational concepts relevant to these phenomena.

1.1 Memory construction

I begin by outlining the evidence for the constructive nature of episodic memory. This comes from two main sources: firstly, many findings from neuroimaging and neuropsychology support the idea that memory and imagination depend on similar neural circuits (or to take this a step further, that memory involves imagination to some extent). Secondly, the memory distortion literature demonstrates how beliefs shape memories.

1.1.1 Memory and imagination

There is plenty of evidence that recalling and imagining events are similar processes, as the constructive view of memory argues (Schacter, 2012), motivating a single computational model of these functions.

Firstly, evidence from neuropsychology shows that hippocampal damage leads to deficits in imagination. For example, Hassabis et al. (2007) found that patients with bilateral hippocampal damage struggle to imagine new experiences, and attributed this to difficulty integrating imagined elements into a coherent scene. Daydreaming is similarly affected; McCormick et al. (2018) found that mind-wandering after hippocampal damage changes 'from flexible, episodic, and scene based to abstract, semanticized, and verbal' (Discussion). Dreaming is another way in which the brain generates episodes. In a small study, Spanò et al. (2020) found that hippocampal amnesia patients reported fewer dreams than healthy controls when woken up at intervals during the night. When dreams were reported, they tended to be less rich in content. In a neuroimaging meta-analysis, Fox et al. (2013) found that the hippocampus (along with several other regions) was highly active during REM sleep, corroborating the findings of Spanò et al. (2020). This evidence implicates the hippocampus in several types of imagination.

Secondly, there is evidence from neuroimaging that recall and imagination involve similar neural processes. For example, Addis et al. (2007) scanned participants whilst they either recalled a past experience or imagined a future one. There were two

phases to the task: in response to a cue word, participants thought of a recalled or imagined event, and then produced further details about the event in an elaboration phase. The researchers found that brain activity was very similar in the remembering and imagining trials, most notably in the elaboration phase.

Furthermore, Hassabis and Maguire (2009) show that a very similar network is involved in recalling recent memories, recalling recent imagined experiences, and generating new imagined experiences (based on a conjunction analysis of fMRI data). This leads to the question of what common function underlies these three tasks. As Hassabis and Maguire (2007) discuss, two explanations had been proposed: mental time-travel and self-projection. However, they argue that scene construction is the missing link instead.

Contrary to the mental time-travel suggestion, Hassabis and Maguire (2009) note that the similarity to recall is observed whether the participant is told to imagine a future event (often referred to as episodic future thinking) or one without temporal context, such as a fictitious scenario. As a result, they suggest that the same mechanism underlies generating episodes whether they are in the past, future, or neither. Contrary to the self-projection suggestion, the researchers show that generating episodes which do not involve the self activates the common network too. Their view is similar to the constructive episodic simulation hypothesis put forward by Schacter and Addis (2007), but Hassabis and Maguire (2009) extend this to events with no temporal context or self-projection. This would be compatible with the idea that a single generative model could generate many types of episode, including recalled events.

Thinking of memory and imagination as functions of the same network builds on findings in the spatial cognition literature, e.g. in the Bicanski and Burgess (2018) model recall and imagination of scenes involve the same neural circuits. Specifically, place and head direction cells act as latent variables in a generative model (Becker & Burgess, 2000; Bicanski & Burgess, 2018; Byrne et al., 2007), so that a scene from a specific viewpoint can be generated. In addition to the imagination of novel scenes, the research shows how egocentric memories could be reconstructed from stored allocentric latent variables.

1.1.2 Memory distortions

Episodic memory in humans is prone to many errors. These distortions are not random noise, but show clear patterns, which provide support for a constructive view of memory.

Bartlett (1932) was the first to explore this topic experimentally. Students heard a story called 'The War of the Ghosts', and were asked to recall it after different time intervals. The story, a Native American myth, was deliberately chosen to be culturally unfamiliar to the students, making the memory distortions more pronounced. Bartlett found that the story was recalled in a way that was more consistent with the students' background knowledge of the world. For example, the word 'canoe' was often replaced by 'boat', and details were added to explain unusual elements of the story (i.e. confabulation and rationalisation were observed). Bergman and Roediger (1999) replicated the Bartlett (1932) experiment, as the initial findings were somewhat anecdotal, and confirmed that memory distortion increases over time after encoding.

In an experiment by Carmichael et al. (1932), participants were asked to reproduce ambiguous sketches. A context was established by telling the participants that they would see images from a certain category. It was found that when they tried to reproduce the image after a delay, their drawings were distorted to look more like members of the context class. There are several similar studies in which memories are distorted to better match their context, for example Ochsner et al. (1997) showed participants images of faces whilst they listened to a happy or angry voice. They remembered the voices seen whilst looking at smiling faces as happier, and voices seen whilst looking at frowning faces as angrier.

The Deese-Roediger-McDermott (DRM) task is a classic way to measure memory distortion (Deese, 1959; Roediger & McDermott, 1995). In brief, the DRM task involves showing participants a list of words that are semantically related to a 'lure word', which is not present in the list. There is a robust finding that false recognition of the lure word occurs, and the lure word is falsely recalled. The DRM results do

not require anything as elaborate as a generative model to explain them – simpler associative models of memory will do, as the researchers suggest themselves (Roediger et al., 1998). However, the type of memory distortions observed in the Carmichael and Bartlett experiments, especially the confabulation of details to rationalise aspects of the story in the latter, are harder to explain in this way.

Semantic influences at recall time can also produce episodic memory distortions. Loftus and Palmer (1974) designed a study in which participants watched a film of a car accident, and were then asked a series of questions, including how fast the car was going when it hit the other vehicle. The verb used to describe the collision (e.g. 'smashed', 'collided', or 'bumped') was manipulated and found to affect the recalled speed. These semantic influences can even produce entirely false memories. Loftus and Pickrell (1995) presented participants with four stories about their childhood. Three were true and one was false, but written (with the help of a relative) to contain plausible details consistent with the participant's childhood experiences. Participants were asked to recall as much as they could about each incident. After repeated interviews, many of the participants remembered the false event as having occurred, even providing additional details about their experience.

More recently, computational modelling has shed light on potential mechanisms underlying memory distortions. Nagy et al. (2020) propose that semantic memory is a generative latent variable model of experience. They argue that this is consistent with several types of gist-based distortions: firstly the semantic intrusions observed in the DRM task (Deese, 1959; Roediger & McDermott, 1995), and secondly the influence of contextual information on memory as in Carmichael et al. (1932). Specifically, Nagy et al. (2020) show that a beta-VAE (Higgins et al., 2016) trained on a class of images biases recall towards that class (however the authors used a separate model for each class of images, rather than a single model with context as an input, which seems more plausible).

1.2 Memory consolidation

Having outlined the evidence for the constructive view of memory, I provide a brief summary of the memory consolidation literature, including the role of hippocampal replay, how memories change over time, and the main theories of consolidation.

1.2.1 What is systems consolidation?

According to the standard view, systems consolidation is the process by which memories become less dependent on the hippocampus over time as they are transferred to neocortex for long-term storage (Marr, 1970, 1971; McClelland et al., 1995; Squire et al., 2015). While synaptic consolidation takes place within hours of learning, systems consolidation is much more prolonged (Frankland & Bontempi, 2005), although the timescale can be very variable (Tse et al., 2007).

Complementary learning systems (CLS; McClelland et al., 1995) is a model of consolidation consistent with the standard view. This hypothesis proposes that the hippocampus and neocortex play complementary roles in learning and memory; whilst the hippocampus is specialised for rapidly learning pattern-separated representations which minimise interference, the neocortex is specialised for gradually learning overlapping representations which support generalisation (Marr, 1970, 1971; McClelland et al., 1995). CLS proposes that episodic memories in the hippocampus are integrated into existing knowledge over time. See Figure 1.1c.

Retrograde amnesia provided early evidence for the standard view of systems consolidation, with recent memories more vulnerable to forgetting than remote ones following hippocampal damage (Squire et al., 2015), but the interpretation of these experiments remains controversial (Nadel & Moscovitch, 1997). Neuroimaging evidence has also been invoked to support this view. Squire et al. (2015) describe a common design in which participants learn similar information at several points in time prior to neuroimaging, e.g. learning pairs of associations both 24 hours and 15 minutes beforehand (Takashima et al., 2009). In general, hippocampal activity decreases and neocortical activity increases as a function of time since encoding.

However, McKenzie and Eichenbaum (2011) review evidence that 'the hippocampus is engaged during any memory processing that involves combinations of detailed associative and contextual information' (Consolidation section). This supports multiple trace theory (MTT; Nadel & Moscovitch, 1997), an alternative account of systems consolidation in which certain memories remain permanently dependent on the hippocampus. MTT suggests that the hippocampus and neocortex are specialised for different types of memory; memory traces bound together by the hippocampus are detailed, specific, and rich in context, i.e. more episodic in nature, while memory traces bound together by the cortex are more semantic in nature.

There is also evidence that detailed, rich spatial memory depends on the hippocampus indefinitely, analogous to the findings described above for episodic memory. Li et al. (2024) find that 'neither HPC nor MTL are critical for allocentric gross representations of large-scale environments' (Conclusion section), in fitting with the idea that semantic memory becomes HF-independent. They also found that 'the HPC appears critical for representing detailed spatial information ... regardless of the age of the memory' (Conclusion section), challenging the standard view of consolidation.

There could be two possible explanations for the continued reliance on the hippocampal formation for detailed, vivid episodic recall (e.g. Li et al., 2024; McKenzie & Eichenbaum, 2011; Nadel & Moscovitch, 1997): one is that certain details are stored in the memory trace in the hippocampus for a long time, and the other is that event construction as a *function* stays dependent on HF even without the retention of a hippocampal 'trace'. In other words, the hippocampus being required for recall does not necessarily imply the existence of a particular autoassociative 'engram' binding the memory together, but could instead reflect activity in a circuit for event (re)construction that passes through HF (Bicanski & Burgess, 2018; Hassabis & Maguire, 2009).

Episodic memories are in constant flux. McKenzie and Eichenbaum (2011) observe that 'the standard consolidation theories described above characterize consolidation as a one-time event, after which a memory is impermeable to subsequent disruption' (Reconsolidation section). But as they describe, this view is undermined by the finding

that a reminder cue can make a memory vulnerable to interference again, even if it has been consolidated completely – this is known as reconsolidation (Nader & Hardt, 2009). McKenzie and Eichenbaum (2011) make the stronger claim that reconsolidation is happening constantly; they suggest that consolidation and reconsolidation reflect the endless modification of schemas by new learning. Káli and Dayan (2004) also discuss the fact that representations in the cortex are 'ceaselessly plastic' (Introduction).

1.2.2 What is hippocampal replay?

Replay is a phenomenon in which hippocampal neurons reactivate memory traces in a temporally compressed form (Foster, 2017). It was first observed in rodent brains replaying the firing of a sequence of place cells during sharp wave ripples (Carr et al., 2011; Wilson & McNaughton, 1994), but has more recently been detected in human neuroimaging studies (e.g. Liu et al., 2019; Schapiro et al., 2018).

Replay involves synchronised activity in the hippocampus and neocortex (Preston & Eichenbaum, 2013; Rothschild et al., 2017). In particular, sharp wave ripples in the hippocampus – during which replay occurs most commonly - are synchronised with spindles in the medial prefrontal cortex (Frankland & Bontempi, 2005). As O'Neill et al. (2010) describe, this coordinated activity 'is consistent with the replay of a memory trace that is distributed across different brain regions, with each area contributing a component of the trace that reflects its role in waking processing' (Reactivation section). However there is some debate about whether this synchronized activity is initiated in the hippocampus or neocortex (Squire et al., 2015).

Replay of episodic memories during sleep and rest is thought to be a mechanism by which memory is 'consolidated' into neocortex (O'Neill et al., 2010). Interventions that interfere with replay, for example electrical stimulation which blocks sharp wave ripples, lead to increased forgetting (Girardeau et al., 2009), and replay after learning a task is correlated with subsequent performance (Peigneux et al., 2004). Note that replay also has several proposed functions other than consolidation, including retrieval (Carr et al., 2011), planning (Ólafsdóttir et al., 2018), and learning from rewards

(Michon et al., 2019).

1.2.3 Which memories get replayed?

There is ongoing debate about which memories are replayed, and how this influences consolidation.

One view is that replay emerges spontaneously from oscillatory activity in the brain. O'Neill et al. (2010) argue that in the recurrently connected CA3 region of the hippocampus, 'previously stored patterns could spontaneously recur when the coincident activation of some cells ... triggers the completion of an entire assembly pattern' (Reactivation section). Similarly, González et al. (2020) suggest that 'spontaneous activity during sleep combined with unsupervised plasticity can trigger reactivation of the previously learned memory patterns and modify synaptic weights reversing damage from the new learning' (Model Predictions section).

An alternative view is that memories are prioritised for replay by some measure of saliency. One idea is that rewarding events could be replayed preferentially (Kumaran et al., 2016), while other research suggests that replay may prioritise the memories most vulnerable to forgetting (Schapiro et al., 2018). Environmental stimuli can also have an effect, for example Bendor and Wilson (2012) show that stimuli during sleep can bias the content of hippocampal replay.

A newer view is that replay may be generative rather than, or as well as, veridical. In rodents, 'replayed' sequences can join together paths that were experienced on separate occasions (Gupta et al., 2010), traverse regions that have been seen but not visited (Ólafsdóttir et al., 2015; Pfeiffer & Foster, 2015), and even 'diffuse' throughout an open environment (Stella et al., 2019). In human neuroimaging studies, 'replayed' sequences do not always correspond to real memories either; Liu et al. (2019) found that, when participants learned a pattern according to which certain stimuli could be unscrambled, sequences experienced in a 'scrambled' form were 'unscrambled' in replay. One complexity of these findings is that observing a novel sequence being (re)activated in the hippocampus does not necessarily imply the sequence is stored

there, or that the activity is initiated there (as noted earlier in relation to hippocampal involvement in remote recall). It may be the case that 'standard' hippocampal replay represents a different process from generative replay, which could instead reflect the 'event construction' circuit discussed above.

Whilst offline replay occurs during sleep or rest, online replay occurs during an activity, or brief pauses in activity. Offline replay largely features remote environments, whereas online replay generally starts in the animal's current location (Foster, 2017). For example, 'preplay' is a replay-like phenomenon preceding an animal's behaviour (Dragoi & Tonegawa, 2011; Ólafsdóttir et al., 2018).

In summary, there are many varieties of replay. Chapters Two and Three focus on the effect of offline, remote replay of memories ('standard' hippocampal replay) on consolidation, but I return to 'generative replay' in Chapter Four.

1.2.4 Consolidation as transformation

Consolidation does not just change which brain regions are involved in an episodic memory; it also changes its properties. This transformation is sometimes referred to as semanticisation (Winocur & Moscovitch, 2011), and seen as a process by which episodic memories become semantic. But here we are more focused on episodic memories before and after consolidation than the extraction of semantic from episodic memory.

Consolidation increases memory distortion (Bartlett, 1932; Payne et al., 2009), but it can also improve performance on a range of tasks. This is especially true when the task involves making inferences based on multiple pieces of information; Ellenbogen et al. (2007) show that transitive inference performance improves over time following encoding, with sleep having more of an effect than waking rest. The ability to distinguish structured from unstructured sequences of tones also improves over time, suggesting that consolidation supports the extraction of statistical structure (Durrant et al., 2011).

More generally, experimental evidence supports the view that consolidation involves

learning patterns across a set of experiences. Squire et al. (2015) describe a study by Richards et al. (2014), in which one day after training, mice searched for a reward in the most recent location. But 'after 30 days, search was driven less by any single day's training, or even by the last day's training, but rather by the cumulative statistical distribution of training experience across days' (Interpreting section). This suggests that consolidation involves extracting the gist of an experience. (Note that the literature does not suggest that consolidation is required to perform these kinds of functions, just that it has a positive impact; clearly it is possibly to generalise, predict, and infer from episodic memories in real time, before consolidation occurs.)

Consolidation also makes memories more conceptual in nature. Lifanov et al. (2021) designed a study in which participants learned cue-object pairings, then at a range of intervals 'were asked to answer one conceptual and one perceptual question about the recalled object as fast as possible' (Introduction). The gap between the time to answer conceptual and perceptual questions about the cued object increased over time; after two days the reaction time to answer perceptual questions was significantly greater than for conceptual questions. They suggest that semantic content may be strengthened preferentially during the consolidation process relative to perceptual content. (But they concede that an alternative explanation might be that perceptual details were forgotten faster.) Lifanov et al. (2021) also compared active retrieval to the restudying of items, and found that active retrieval led to a greater advantage for the conceptual content. They hypothesise that each time a memory is recalled or replayed it is 'semanticised' a little more, in agreement with the claim that reconsolidation is a constant process (McKenzie & Eichenbaum, 2011).

In summary, consolidation extracts patterns and promotes inference, but increases distortion and the loss of perceptual detail from episodic memory. It at first seems paradoxical that memories can become more useful even as they become more distorted, but one might hypothesise that extracting conceptual structure from experience comes at the cost of detail and accuracy, reconciling these findings.

1.2.5 Consolidation and continual learning

Catastrophic forgetting refers to the overwriting of old knowledge by new knowledge when a neural network learns multiple tasks or distributions consecutively. Continual learning is the ability to learn, or memorise, a series of tasks, or items, sequentially, without the occurrence of catastrophic forgetting (Hadsell et al., 2020).

Of course gradual forgetting is expected, but catastrophic forgetting is a more dramatic interference of new with old knowledge, which is not typically observed in reality. If hippocampal traces are not retained in the hippocampus forever, as per CLS (McClelland et al., 1995), this poses a problem for our current understanding of consolidation: most connectionist models would predict catastrophic forgetting of old knowledge if there are no reminders of a category.

To clarify, this is a problem for the standard view of consolidation and the Complementary Learning Systems theory (CLS; McClelland et al., 1995), in which traces do not stay in the hippocampus forever. It is not a problem for Multiple Trace Theory (MTT; Nadel & Moscovitch, 1997) and its variants, in which traces, or even multiple overlapping traces, do. But episodic recall remaining dependent on HF as a function is *not* enough to avoid this problem, if traces (i.e. autoassociative engrams) are not preserved.

One suggestion, which does not require retaining relevant memories indefinitely, is that replay samples from a generative model (Van de Ven et al., 2020). However, this is not the only approach for avoiding catastrophic interference; other suggestions include expanding the network to capture each new task without interference (Rao et al., 2019), and freezing or penalising changes to important weights, e.g. elastic weight consolidation (Kirkpatrick et al., 2017). These questions are discussed further in Chapter Four.

1.3 Neural substrates of memory

In this section I review the neural substrates of memory, both its initial hippocampal encoding and the broader network that may be involved after consolidation.

1.3.1 The hippocampal formation

Situated in the medial temporal lobe, the hippocampal formation (HF) plays a crucial role in memory and spatial navigation. The hippocampal formation includes the hippocampus proper (consisting of the dentate gyrus, CA1, CA2, and CA3) and surrounding regions like the entorhinal cortex and subiculum.

The trisynaptic loop is a key pathway within the hippocampus, so called because it describes three sets of connections between different subregions (Andersen, 1975). Firstly, entorhinal cortex projects to granule cells in dentate gyrus via the perforant path. Secondly, granule cells in dentate gyrus project to pyramidal cells in CA3 via mossy fibres. Thirdly, these CA3 cells project to pyramidal cells in CA1 via Schaffer collaterals. Finally, projections from CA1 back to entorhinal cortex complete the circuit. The entorhinal cortex is the main route between HF and the rest of the brain; HF's extensive connections with other brain areas make it well suited to binding multimodal information together (Witter et al., 2000).

The DG and CA3 regions are thought to be specialised for pattern separation and completion respectively. In the Rolls-Treves model (Treves & Rolls, 1992), the dentate gyrus (DG) performs pattern separation. This is the process by which similar input patterns are transformed into more distinct, less overlapping output patterns before projecting to CA3, reducing interference between similar events. Meanwhile the CA3 region of the hippocampus, with its extensive recurrent connections, is thought to function as an autoassociative network, storing memory patterns such that the retrieval of a memory can be triggered by a partial or noisy version of that memory (known as pattern completion). In other words, DG and CA3 are thought to work together to transform representations to a suitable format, then store and retrieve these 'attractors' autoassociatively.

Neuropsychology is a rich source of evidence about hippocampal function. Individuals with hippocampal damage display severe impairments in declarative memory, the ability to consciously recall facts and events. These deficits include both anterograde amnesia (difficulty in forming new memories) and retrograde amnesia (difficulty in recalling past memories), although the latter is dependent on the age of the memory and the extent of the damage. Case studies like that of patient H.M., who had his medial temporal lobes surgically removed to treat his intractable epilepsy, have been key to understanding the role of the hippocampus in memory (Scoville & Milner, 1957). Whilst H.M. could remember events from his early life and learn new skills, he was unable to form new long-term memories of events or remember recent ones. This temporal gradient to patient H.M.'s retrograde memory deficits is characteristic of hippocampal amnesia.

Semantic memory is preserved when the hippocampus is lesioned (Manns et al., 2003; Squire et al., 2015; Vargha-Khadem et al., 1997), and hippocampal amnesics can describe the factual content of remote (i.e. consolidated) memories more successfully than recent ones (Scoville & Milner, 1957; Spiers et al., 2001). However they are often unable to recall these memories 'episodically', i.e. vivid, detailed 're-experiencing' of events appears to depend on HF (Nadel & Moscovitch, 1997). As described above, more recent evidence suggests that damage impairs not just episodic memory but the construction of mental events or scenes more broadly, affecting abilities such as imagination (Hassabis et al., 2007), episodic future thinking (Schacter et al., 2017), dreaming (Spanò et al., 2020), and daydreaming (McCormick et al., 2018). Conversely, hippocampal amnesics often retain their procedural memory, which is the memory for skills and tasks (e.g. for how to ride a bicycle). Similarly, short-term and working memory often remain intact in individuals with hippocampal damage.

1.3.2 Hippocampal indexing theory

How exactly does the hippocampus encode memories? Hippocampal indexing theory (Teyler & DiScenna, 1986) proposes that memory traces bind together neocortical areas that were active during the original experience (Figure 1.1d). In other words,

they are 'pointers' to the elements that make up a memory, so that activity in the hippocampal trace spreads to neocortex to reactivate the memory's components.

There is evidence that the hippocampus can bind both conceptual and sensory representations into an event. Quiroga (2012) observes that certain neurons in the hippocampus ('concept cells') respond selectively to concepts; in the classic example, a Jennifer Aniston neuron might fire in response to her picture, voice, and written name. Meanwhile, Wheeler et al. (2000) show that 'brain areas in visual and auditory cortex are transiently active during memories that involve vivid visual and auditory content, respectively'. Horner et al. (2016) corroborate the finding that late sensory areas are reactivated while recollecting a hippocampus-dependent memory of an event. (However, activity in sensory areas does not imply that the hippocampus connects directly to these areas, as it could be mediated by schematic representations.) Note that talking of sensory versus conceptual representations is a simplification, as there are many levels of abstraction, e.g. ranging from early visual cortex cells that respond to lines at different orientations to face-selective inferotemporal cortex cells to concept cells (Quiroga, 2012).

There is some tentative evidence that the distribution of connectivity between sensory and conceptual representations changes over time. In their rodent electrophysiology study, Yu et al. (2018) suggest that as time passes, 'links that map common features shared across experiences with specific features of single experiences become enriched whereas links between representations for specific features in the hippocampus and cortex diminish' (Discussion). This may be linked to lateralisation of the hippocampus' role in memory. Researchers have suggested that the left hippocampus encodes more schematic representations and the right hippocampus more item-specific ones (Chiarello & Beeman, 1997). Maguire and Frith (2003) found that activity in the right hippocampus was lower for more remote memories, unlike in the left hippocampus. This would seem to support the idea that schematic representations play a greater role over time, but the literature on this topic is limited.

1.3.3 Anterior vs. posterior hippocampus

The posterior hippocampus (pHPC) may encode more fine-grained and perceptual aspects of episodic memories, and the anterior hippocampus (aHPC) may encode more coarse-grained and conceptual aspects (Moscovitch et al., 2016).

The differences between the aHPC and pHPC are most established in spatial cognition. As Zeidman and Maguire (2016) explain, place cells in the aHPC (or the ventral hippocampus in rodents) have 'firing fields that cover a larger area of space than posterior (dorsal) place fields' (Representing The Environment section). It has been proposed that these differences reflect a gradient in detail, with the pHPC encoding more detailed representations. However, the authors note that recent findings may suggest a more complex picture. Keinath et al. (2014) found that the location of an animal can be decoded as precisely from the population of place cells in aHPC as the population in pHPC, despite each cell representing a larger area in the former. In other words, this suggests the anterior hippocampus may implement a more distributed representation that still has the same spatial resolution, rather than a representation with lower spatial resolution. In any case, there is general agreement that the aHPC supports 'a spatially large-scale or generalisable representation of the environment' (Zeidman & Maguire, 2016, Representing The Environment section).

More recently, there is also evidence of functional specialisation along the hippocampus when it comes to episodic memory. As Moscovitch et al. (2016) describe, the pHPC and connected posterior neocortical regions may represent 'the local, spatio-perceptual aspects of the experience', whereas the aHPC and several connected regions (especially the anterior temporal lobe, prefrontal cortex, and amygdala) may 'represent conceptual and emotional aspects' (Component Processes section). Robin and Moscovitch (2017) propose that the aHPC may encode more coarse-grained global representations which support processing of gist. However, there are dissenting views (Bonnici et al., 2013; Dandolo & Schwabe, 2018).

Zeidman and Maguire (2016) propose that the aHPC supports scene construction, both offline to recall or imagine scenes and online during perception. In a related study,

McCormick et al. (2021) measured hippocampal activity in participants comparing the layout of scenes, and comparing colours in those scenes. They found that the layout condition, which they linked to scene construction, was associated with more aHPC activity, and the colour condition, which they linked to scene perception, was associated with more pHPC activity. This is potentially consistent with the Moscovitch et al. (2016) view, as the mental model of a scene could be understood as its conceptual representation.

1.3.4 Latent variable representations in the brain

The entorhinal cortex (EC) is the main route into and out of the hippocampus. It is also where grid cells, which display a distinctive grid-like firing pattern as an animal moves through a space, are most often observed (Moser et al., 2008). Grid cells may be a mechanism behind path integration and vector navigation (Bush et al., 2015). The EC has also been linked to structural inference - which could be seen as the non-spatial equivalent of path integration - and prior models suggest it encodes latent structures underlying spatial and non-spatial tasks (Whittington et al., 2020).

Other recent studies have also explored the role of EC in conceptual knowledge. Constantinescu et al. (2016) found that the navigation of a conceptual space produced grid-like activity in EC (and elsewhere), as observed in spatial navigation. In this human fMRI study, participants learnt an abstract space defined by two axes, the leg length and neck length of cartoon birds. A grid-like signal was observed as participants imagined an item's trajectory in abstract space. The authors conclude that 'conceptual knowledge may also be organized by grid-like codes' (Introduction), supporting the hypothesis that conceptual knowledge may be encoded spatially as a cognitive map.

There is ongoing debate about the purpose of lateral EC. Tsao et al. (2013) explore the activity of two types of neuron in lateral entorhinal cortex as a rodent explores an environment: object cells fire when an object is present, whereas object trace cells fire when a previously encountered object has been removed (and could therefore be

seen as representing prediction error). An alternative suggestion is that 'populations of lateral entorhinal cortex neurons represent time ... [which] may be integrated with spatial inputs from the medial entorhinal cortex in the hippocampus' (Tsao et al., 2018, Abstract).

In addition, the ventromedial prefrontal cortex (vmPFC) is highly connected to the hippocampal formation and plays a crucial role in episodic memory processing (Gilboa & Marlatte, 2017). Recall-related activity in the vmPFC increases over time (Takashima et al., 2006), and interaction between the hippocampus and vmPFC is observed following the encoding of episodic memories (Gais et al., 2007). This interplay is thought to be a mechanism behind consolidation and the resulting transformation of memory into a more semanticised form (Winocur & Moscovitch, 2011). In an MEG study, McCormick et al. (2020) observed the hippocampus and vmPFC while participants recalled autobiographical memories from a verbal cue (controlling for the level of detail, and varying the remoteness). They found that vmPFC activity preceded hippocampal activity, for all but the most recent memories; they conclude that the vmPFC is not required to direct recall for memories that are very recently encoded, and therefore still intact in the hippocampus, but for all other memories it is.

The vmPFC is also thought to encode schemas (Ghosh et al., 2014). Although definitions of this term vary, schemas can be seen as templates for scenes and episodes. For example, one's schema for 'a birthday party' might bind together concepts such as 'a birthday cake', 'presents', and 'balloons' in a certain relation. Other findings implicate the vmPFC in transitive inference (Koscik & Tranel, 2012) and the integration of memories (Spalding et al., 2018). Common to all these capabilities is the extraction of underlying structure from experience. Mack et al. (2020) suggest the vmPFC performs dimensionality reduction on incoming data, compressing representations to remove irrelevant features as learning progresses.

Other candidate regions for latent variable representations of memory include anterior and lateral (anterolateral) temporal cortices. These regions, which are associated with semantic memory (Chan et al., 2001; Lambon Ralph & Patterson, 2008; Patterson et

al., 2007), might be hypothesised to contain latent variable representations capturing semantic structure.

1.4 Computational models of memory

In this section I briefly review existing computational models of memory, beginning with models of associative memory in the hippocampal formation, and moving on to models of systems consolidation.

1.4.1 Computational models of associative memory

Whilst the fate of memories after encoding is debated, there is a general consensus that memories are first stored in autoassociative networks that bind together memory elements into a trace (Marr, 1971; McNaughton & Morris, 1987). Each trace is thought to 'index' neocortical elements of the memory (Teyler & DiScenna, 1986), so that pattern completion of a partial input can reactivate (some approximation to) the original experience.

Many computational models of associative memory stem from the Hopfield network (Hopfield, 1982). This was not specifically designed as a model of the hippocampus, but as a more general model of how pattern completion, or 'content-addressable' memory, could be implemented in neural networks. A Hopfield network uses a simple Hebbian learning rule to memorise patterns after a single exposure. As a result, it is often considered more biologically plausible at a neural level than networks using backpropagation.

Consider a set of fully connected nodes, i.e. each node is connected to every other node. Let us assume that each node can take either 1 or -1 as its value. Whilst a Hopfield network can memorise many types of data, we will consider the memorisation of black and white images, where each pixel is represented by a node; black pixels are represented by 1 and white pixels by -1. The task is to memorise a set of images, such that the network can recall the most similar image when presented with a noisy or

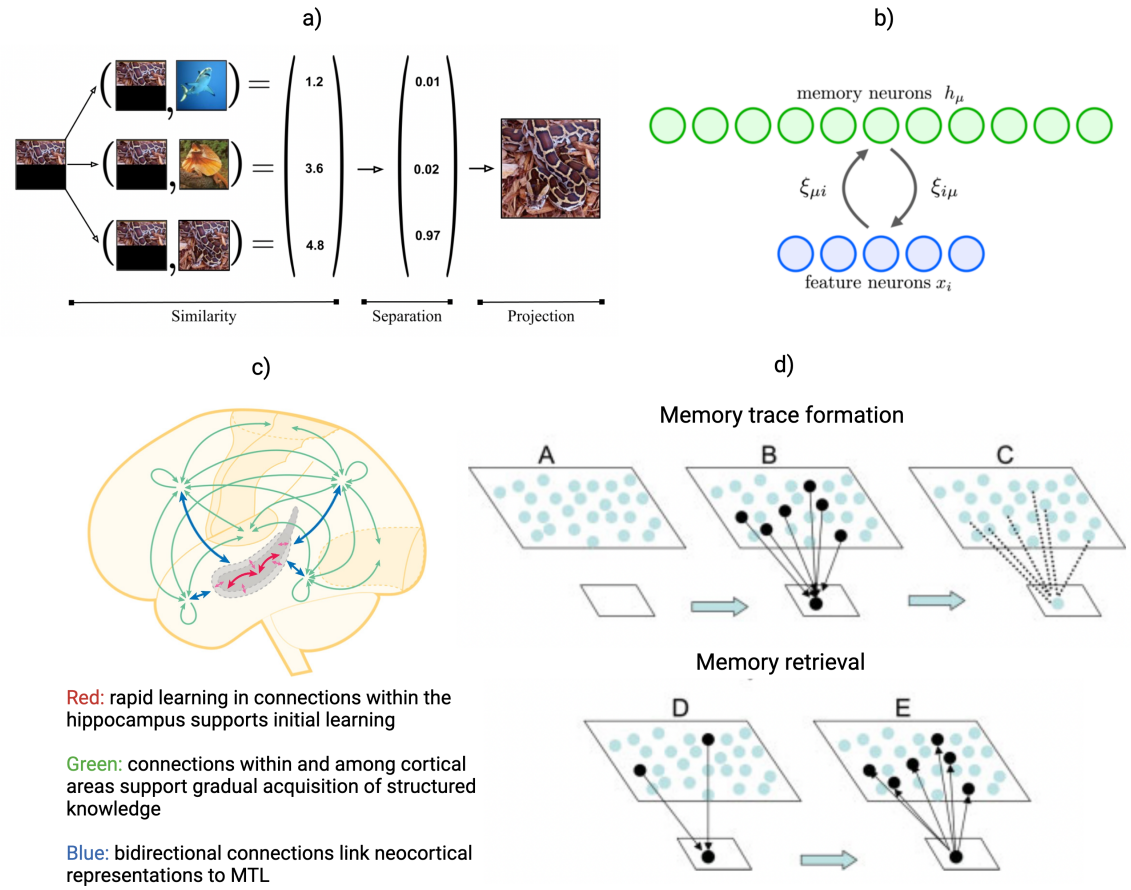

Figure 1.1: Relevant computational models. a) Figure 1 from Millidge et al. (2022), showing the Hopfield network and its successors as a series of three operations: the calculation of similarity, separation of the vector of similarities, and projection. b) Figure 2 from Krotov and Hopfield (2020), showing the formulation of Hopfield network variants as a two-layer network, with one layer of memory neurons and one layer of feature neurons. c) An adapted version of Figure 1 from Kumaran et al. (2016), summarising the CLS view. d) An adapted version of Figure 1 from Teyler and Rudy (2007), showing memory encoding and retrieval according to hippocampal indexing theory. Forming a memory trace involves the hippocampus 'binding' together neocortical features. A partial input reactivates the hippocampal trace, and thus 'pattern completes' the full memory.

incomplete input. The nodes are connected with symmetric weights. How is activation propagated through a Hopfield network? Similarly to other neural networks, the activation of a node is calculated from the weighted sum of its inputs. In a Hopfield network we simply take the sign of this weighted sum, i.e. it is rounded to whichever of -1 or 1 is closest. So as activation spreads through a Hopfield network, nodes change to the sign of the weighted sum of their inputs, until a stable state is reached in which nodes stop changing value. In more mathematical terms, this is a procedure for minimising an 'energy' function. The activations can be propagated synchronously – the activations of all the nodes are updated at once - or asynchronously. How are these weights learned? The change to a weight is the product of the activations at the nodes it connects. So if both nodes have the same activation, the weight increases. If the nodes have different activations, the weight decreases. To put this another way, if each pattern to be memorised is a vector with elements of 1 or -1, the weight update for that pattern is the outer product of the vector with itself (a matrix). The final set of weights is the sum of these matrices.

In this way Hopfield networks can memorise and retrieve patterns after a single exposure (Hopfield, 1982). However one issue is their limited capacity; a Hopfield network can only recall approximately 0.14d states, where d is the dimension of the input data (Ramsauer et al., 2020). It therefore seems unlikely that classical Hopfield networks are a good model of hippocampal memory encoding – even if we assume that only a temporary store is required until consolidation occurs. In addition, they frequently recall incorrect memories, as the energy function can get 'stuck' in a local minimum.

However, recent research has shown that the storage capacity of a Hopfield network can be increased by making the energy function more complex (Krotov & Hopfield, 2016). Demircigil et al. (2017) develop this idea further, increasing the capacity from approximately 0.14d to $2^{d/2}$ with the use of an exponential energy function. Ramsauer et al. (2020) further amend the energy function and enable the storage of patterns of continuous variables, allowing the recall of much more complex data. (For example, whilst classical Hopfield networks can only recall black and white images, the modern

variant can recall greyscale ones.)

However, understanding these new variants of Hopfield networks in terms of neural networks is less straightforward. The equations below (from Krotov & Hopfield, 2016) give the energy of a standard Hopfield Network in a state σ. To recap, a node's value is updated to the sign of the weighted sum of its inputs during recall; in other words, a node's value is flipped if it decreases the energy. The matrix T gives the weights of the network, and the calculation of T is simply Hebbian learning:

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i T_{ij} \sigma_j, \qquad T_{ij} = \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$

The equation below from Krotov and Hopfield (2020) gives the energy of a dense Hopfield network; in this example F(x) is x^3, but it can be any polynomial function. As above, at recall time a node's value flips if it decreases the energy. When F(x) is x^2, the equation reduces to the one above for a standard Hopfield network. In any other case, the tensor T has more than two indices, and can no longer be thought of a matrix produced by Hebbian learning. This means the energy is no longer a function of weights and activations in a neural network. Modern Hopfield networks (Krotov & Hopfield, 2020) suffer from the same problem.

$$E = -\sum_{\mu} F(\sum_{i} \xi_{\mu i} \sigma_i) = -\sum_{i,j,k} T_{ijk} \sigma_i \sigma_j \sigma_k$$

Krotov and Hopfield (2020) suggest a way to overcome this problem by using hidden units (which they call 'memory units') in addition to the 'feature units' which represent the input. With this change, the energy is again a function of weights and activations in a neural network, and so is the update rule that minimises it. In simple terms, with the addition of memory units, feature units can interact in more complex ways whilst remaining consistent with the constraints of standard neural networks (i.e. two-body synapses). As a result, a modern Hopfield network can be understood as a neural network, like its predecessor (see Figure 1.1b). Note that under certain conditions,

the Krotov and Hopfield (2020) formulation simplifies to dense associative memory (Krotov & Hopfield, 2016) rather than modern Hopfield networks (Ramsauer et al., 2020).

An important question is how the memories get encoded as the weights of a bipartite graph in the Krotov and Hopfield (2020) formulation of a modern Hopfield network. Each memory is bound together by a single node, which connects the features that comprise that memory. The weights between a given memory node and the feature nodes are simply the values of the features for that memory; these weights can be learned by Hebbian learning. Therefore encoding in a modern Hopfield network can be thought of as similar to previous models of the hippocampus as 'indexing', or binding together, a set of memory components (Teyler & DiScenna, 1986); the innovative aspect is the update rule, which is cleverly designed to guarantee the desired properties. The key point is that the Krotov and Hopfield (2020) formulation of a modern Hopfield network does not require a separate matrix of stored patterns – the patterns are encoded in the weights, and the energy is a function of weights and activations as explained above.

Millidge et al. (2022) give a unifying account of this family of models, building on the Krotov and Hopfield (2016) formulation to describe such networks as a sequence of similarity, separation, and projection operations. As Figure 1.1a shows, retrieval involves calculating the similarity between the query and stored patterns, passing the similarities through another function to separate them, and producing an output given the transformed similarities. This may repeat for multiple iterations. The Millidge et al. (2022) framework can help us to interpret the equation below, which gives the new state pattern ξ^{new} in a modern Hopfield network in terms of the previous state ξ, stored patterns X^T, and inverse temperature β:

$$\xi^{new} = Xsoftmax(\beta X^T \xi)$$

Firstly, the vector of similarities between the query ξ and memories X^T is obtained. Secondly, the softmax function transforms these so they add up to one, with β

determining whether the vector of similarities is 'flattened' (low β) or 'sharpened' (high β). Thirdly, ξ^{new} is calculated as the sum of stored patterns weighted by the transformed similarities. Low β makes it more likely that a superposition of stored attractors is retrieved, whereas high β makes it more likely that an individual attractor is.

1.4.2 Computational models of the hippocampal formation

There are several more detailed computational models of the hippocampal formation, building on the circuits described by Marr (1971) and Treves and Rolls (1992).

Some of these models focus on spatial cognition. For example, Bicanski and Burgess (2018) propose a model of visuospatial memory and imagination in which viewpoint-dependent imagery is (re)constructed from viewpoint-independent latent variables. To encode an event, parietal areas support an egocentric representation from the animal's point of view. Given the parietal inputs and the head direction of the animal, retrosplenial cortex then deduces an allocentric (viewpoint-independent) representation. The resulting allocentric representation involves boundary vector, object vector, and place cells, which can be thought of as latent variables for the scene (Becker & Burgess, 2000; Bicanski & Burgess, 2018; Byrne et al., 2007). These latent variables are then stored in the hippocampus. To recall a scene, egocentric representations are reconstructed from *stored* allocentric ones, while to imagine a scene, egocentric representations are constructed from *novel* allocentric ones.

Other models address more general functions. Gluck and Myers (1993) propose that the 'function of the hippocampal region is the recoding of internal stimulus representations to facilitate learning' (Discussion). These special representations are learned in the bottleneck layer of a 'predictive autoencoder' (which outputs a classification as well as the reconstructed input) in the hippocampal formation, and then used by other brain regions. The recoded representations have two key advantages compared to the original stimuli: improved ability to partition the data by predicted class, and compression of information that is irrelevant to prediction.

When the representations are visualised, stimuli from the same class are clustered while stimuli from different classes are separated.

Models exploring the role of different pathways within the hippocampal formation have been proposed in recent years. For example, C-HORSE (Schapiro et al., 2017) suggests how the hippocampal formation can learn both specific events and statistical regularities in 'complementary' pathways. The authors propose that a 'monosynaptic' pathway directly from EC to CA1 can rapidly learn patterns, whereas the trisynaptic loop via DG and CA3 encodes individual events (Treves & Rolls, 1992), and argue that this accounts for experimental findings which challenge CLS.

Several recent studies have proposed that the hippocampal formation is a generative model, with bottom-up connections to infer latent representations, and top-down connections to reconstruct sensory inputs. These studies focus on modelling spatial cognition at the neural level, as opposed to the cognitive level, so cover different ground to the work in this thesis.

Whittington et al. (2020) present the Tolman-Eichenbaum Machine (TEM) as a model of the hippocampal formation as a generative network supporting structural inference in both spatial and non-spatial tasks; this cognitive ability enables inference in new environments and tasks based on knowledge of a common underlying structure. A spatial example of structural inference is the finding of shortcuts (as this relies on the common structure of space), and a non-spatial example is inferring that A is the grandfather of C from the knowledge that A is the father of B, and B is the father of C (as this relies on the common structure of family trees). More simply, all transitive inference falls into this category. The relations in these tasks can be seen as edges in graphs. The TEM is a recurrent graph variational autoencoder that learns the hidden structure of input graphs, supporting structural inference. Grid cells are latent representations enabling the prediction of the next state. Place cells are the output of the variational autoencoder, which are then associated with sensory data via Hebbian learning. In other words, Whittington et al. (2020) propose that 'medial entorhinal cells form a basis describing structural knowledge, and hippocampal cells link this basis with sensory representations' (Summary section). They show that the

TEM is consistent with observed firing patterns of grid cells and place cells. See also Whittington et al. (2021), a more recent extension of the model.

Stoianov et al. (2022) suggest that the hippocampal formation is a hierarchical generative model supporting spatial cognition, and that as it learns new environments it generates samples of previous environments to enable continual learning. They point to the non-veridical sequences observed in the hippocampus (Dragoi & Tonegawa, 2011; Gupta et al., 2010; Liu et al., 2019; Ólafsdóttir et al., 2018; Stella et al., 2019) as evidence for this. Items, sequences, and maps correspond to different levels of abstraction in the hierarchy. The ability to infer which environment the agent is in, as represented by the final 'map' layer, is used to assess continual learning. The authors compare generative replay against experience replay as a baseline, and find that in some cases the former surpasses the latter. (See Chapter 4 for further discussion.) Similarly, George et al. (2023) simulate the hippocampal formation as a Helmholtz machine applied to a sequence of sensory inputs.

1.4.3 Computational models of systems consolidation

As described above, the complementary learning systems theory (CLS; Marr, 1970, 1971; McClelland et al., 1995) is consistent with the standard view of systems consolidation. It proposes that episodic memories are integrated into existing knowledge over time, with the hippocampus specialised for one-shot learning of episodic memory, and the neocortex for gradual learning of semantic memory.

Rival theories include Multiple Trace Theory (MTT; Nadel & Moscovitch, 1997) and Trace Transformation Theory (TTT; Winocur et al., 2010). In both MTT and TTT, episodic memory stays reliant on the hippocampus, while neocortical representations support semantic memory. MTT suggests that every time an episodic memory is retrieved, a new, slightly modified trace of that memory is created in the hippocampus (which also captures the context of the retrieval). This results in multiple overlapping traces for the same memory, such that older memories are more robust to partial hippocampal damage (an alternative explanation of the gradient

of retrograde amnesia). TTT agrees with MTT that hippocampal and neocortical traces exist in parallel, but focuses on the idea that memory traces are transformed as they are consolidated from the hippocampus to the neocortex. According to TTT, the initial episodic memories stored in the hippocampus are gradually abstracted into semantic memories in the neocortex, resulting in the loss of perceptual detail and context-specific information. In TTT either the episodic or semantic trace may be recalled depending on the situation, or the two traces may interact.

Extensive computational modelling has explored the implications of CLS. One recent development to the CLS framework (McClelland et al., 1995) focuses on the interplay between predictability and systems consolidation. The Go-CLS model (Sun et al., 2021) suggests that unpredictable experiences are less likely to be consolidated than predictable ones, as attempting to consolidate outliers might impair generalization. In other words, Sun et al. (2021) suggest that the brain regulates the amount of systems consolidation based on the predictability of experiences.

Previous work tends to focus on the extraction of semantic memory from episodic memory, but has less to say about the transformation of the episodic memories over time. For example, semantic memory is often simulated as the neocortex learning to classify samples replayed by the hippocampus, leaving the issue of post-consolidation episodic memory relatively unaddressed.

1.5 The Bayesian brain

1.5.1 Memory, novelty and prediction error

Novelty is thought to promote encoding within HF (Hasselmo et al., 1996), while more predictable events consistent with existing schemas are consolidated more rapidly (Tse et al., 2007). Furthermore, activity in the hippocampus can reflect prediction error or mismatch novelty (Chen et al., 2011; Kumaran & Maguire, 2006).

Previous computational modelling has explored how predictable and unpredictable memory elements might be processed differently. For example, the Tolman-

Eichenbaum Machine (Whittington et al., 2020) factorises shared structure (e.g. the structure of a family tree, or the common properties of Euclidean space) from specifics (e.g. the names of members of a particular family, or a reward being at a particular arbitrary location).

Other models share the intuition that the hippocampus encodes arbitrary specifics. Benna and Fusi (2021) propose that hippocampal storage capacity could be greatly increased if only the uncorrelated elements of memories are encoded in a Hopfield network. A set of realistic memories can be imagined as a set of dense, correlated vectors, and the capacity of a Hopfield network for such memories is fairly low. But if one makes each memory the sum of a sparse, uncorrelated vector per memory (capturing the arbitrary specifics), and a shared dense vector (capturing the common aspects), the capacity is much greater. The authors show that this can be modelled as the encoder component of an autoencoder, which projects the memories into a vector space that captures only the uncorrelated aspects, chained to a Hopfield network. (However this intriguing result may work less well if the system enters a completely novel environment. For one-shot encoding to take place, the stimuli must be encoded by an encoder that is unfamiliar with the environment, and therefore may perform poorly.) Other models also suggest that novel elements of memory should be encoded preferentially, e.g. Hedayati et al. (2022) propose a model in which novelty affects the degree of compression of representations in working memory.

Novelty has also been investigated as a factor affecting the duration of systems consolidation. Tse et al. (2007) describe evidence that 'systems consolidation can occur extremely quickly if an associative schema into which new information is incorporated has previously been created' (Abstract). In their study, rats were trained to learn multiple flavour-place associations, such that 'when cued with a specific flavor in start boxes at the side of the arena, the animals would be rewarded for going to the correct location by receiving more of that same food' (Experiments section). Once the rats were familiar with the task, they found that rats could consolidate a new flavour-place association after a single trial, suggesting that 'the rate at which systems consolidation occurs in the neocortex can be influenced by

what is already known' (Discussion). This challenges idea that the neocortex can only learn slowly, as CLS arguably implies (Kumaran et al., 2016).

But there is some debate about this in the literature. Whilst Tse et al. (2007) find that having a matching schema speeds up consolidation, Preston and Eichenbaum (2013) suggest that having fewer, simpler schemas related to the new information may do this instead; whilst humans take a long time to integrate new information into multiple complex schemas, rodents' schemas 'may be limited and relatively simplistic ... resulting in a decreased tendency to process new events in relation to existing memories and more rapid consolidation relative to other species' (Discussion). Perhaps what matters is the degree of inconsistency between the new memory and existing schemas. If so, both the lack of a relevant schema *or* a schema consistent with the new memory could lead to rapid systems consolidation. In contrast, consolidating a memory into a schema that is inconsistent would take longer. See also Van Kesteren et al. (2012) for a model of the relationship between schema congruency and consolidation.

The concept of mismatch between representations of a memory, and replay as a mechanism for resolving this mismatch, has been explored in the context of spatial cognition. Evans and Burgess (2019) propose a model in which information from the associative spatial system (i.e. hippocampal place cells) updates the metric spatial system (i.e. entorhinal grid cells) during replay, triggered by prediction error between the systems.

1.5.2 Bayesian accounts of perception and memory

Perception is now widely understood as a process of probabilistic inference of the state of the world from ambiguous sensory information, with important implications for memory. As Fiser et al. (2010) describe, perception was historically thought of as 'a series of classical signal processing operations, by which each sensory stimulus should give rise to a single perceptual interpretation' (Probabilistic Perception section). But this is challenged by the fact that sensory input may be consistent with multiple

interpretations, e.g. a two-dimensional view may be consistent with multiple three-dimensional objects (Kornmeier & Bach, 2005). Prior knowledge must therefore be drawn on in perception to deal with the noisy, partial, and ambiguous nature of sensory inputs.

More formally, it can be argued that this process approximates Bayesian inference. Cognition is thought to involve estimating the probability of certain hypotheses about the world based on observed data (Fiser et al., 2010). This can be achieved by combining a prior belief about the probability of the hypothesis with the likelihood of observed data given that hypothesis, and applying Bayes' theorem to obtain the posterior probability. For example, given an ambiguous silhouette moving in the distance, the posterior probability that the silhouette is that of a certain animal depends on a prior (i.e. the baseline probability of seeing the animal in the environment) and a likelihood (i.e. how well the silhouette matches the hypothesised animal).

There are a number of related findings about attribute estimation from memory. When people are shown a stimulus (e.g. an image of an object) then asked to remember its properties after a short delay, they tend to remember stimuli as being closer to the average than they actually were; this is known as the central tendency effect (Hollingworth, 1910). Petzschner et al. (2015) present a Bayesian analysis of magnitude estimation, in which the central tendency effect is 'the natural consequence of general principles underlying perceptual inference' (Theories section), arising from a predictive model of the world. In a Bayesian framework, the mean magnitude is a prior, which influences the judgements of magnitude, and a 'statistically optimal combination of prior knowledge and sensory input produces biased magnitude judgments whenever the prior differs from the current physical stimulus magnitude' (A Bayesian Framework section).

More generally, Lin et al. (2022) argue that memories of scenes are biased towards high-probability perspectives. This reflects the central tendency effect and is compatible with the Bayesian account. Similarly, in experiments where participants recall the location of an item in a circle, recalled stimuli are distorted towards a prototype, due to 'estimation processes that combine the remembered stimulus value with category

information' (Huttenlocher et al., 1991, Abstract).

Hemmer and Steyvers (2009) build on Huttenlocher et al. (1991) to give a broader account of reconstructive memory as Bayesian inference, in which the 'posterior probability $p(l|y)$ gives the likely stimulus values l given the noisy memory contents y' (Introduction). They note that in tasks like that of Huttenlocher et al. (1991), their model predicts a bias towards the mean of the category when the memory is reconstructed. These are just some of the studies giving a Bayesian perspective on memory, however it is unclear how these normative models might actually be implemented in the brain.

1.5.3 Early models of predictive coding

Predictive coding, a prominent account of perception as a Bayesian process, has its origins in information theory. As Millidge et al. (2021) describe, 'information theory tells us that information is inseparable from a lack of predictability. If something is predictable before observing it, it cannot give us much information' (Introduction). In contrast, if something is less predictable it is richer in information, so converting data to a 'minimally predictable' form maximises information transfer.

Predictive coding was originally applied to reduce redundancy in signal processing, thereby reducing the bandwidth required for data transmission (Millidge et al., 2021). Early techniques for video transmission involved subtracting each frame from the preceding one, or in other words predicting that consecutive frames would be identical and calculating the prediction error or 'residual' (this is especially effective when objects are moving on a static background, as only the moving objects need to be transmitted). More sophisticated techniques developed from this, such as linear predictive coding, where the prediction for each new frame is a linear combination of previous frames weighted by coefficients. When these coefficients are communicated at the start of the transmission, the receiver can reconstruct the compressed signals from the transmitted prediction errors. Barlow et al. (1961) and Barlow (1989) related this concept to signal processing in the nervous system, suggesting that

given the high energy consumption of neurons, evolutionary pressures encourage efficiency. Specifically, the minimum redundancy principle (Barlow et al., 1961; Barlow, 1989) proposes that the nervous system is optimised to reduce redundancy in sensory information, leading to sparse representations that transfer information efficiently.

Traditional views of sensory processing assumed information flow from low-level sensory to high-level conceptual representations. As Millidge et al. (2021) describe, the classical view of the visual system involved feature detectors arranged in a hierarchy, with more complex features consisting of combinations of simpler features. But several phenomena were observed that challenged this view. In particular, some low-level cells' responses depended on contextual information outside of their receptive field (for example, 'endstopping' refers to the reduction in certain neurons' responses when a stimulus extends beyond the receptive field). This is clearly consistent with 'top-down' as well as 'bottom-up' flow of information. (Centre-surround cells on the retina, which fire in response to either a light spot on a dark background or a dark spot on a light background, are another related example in the visual system. Srinivasan et al. (1982) suggest that they help minimise redundancy by indicating prediction error, relative to the prediction that nearby points in space are the same shade.)

To address these issues with the classical account, Mumford (1992) proposed an alternative account of the interaction of lower-level sensory and higher-level conceptual regions inspired by the minimum redundancy principle (Barlow et al., 1961; Barlow, 1989). In his view, 'the higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them ... a template reconstruction best fitting the lower-level view', while 'the lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back ... the features in its data which are not predicted by the higher area' (Abstract). Inspired by Mumford (1992), Rao and Ballard (1999) introduced the initial predictive coding network, showing that extra-classical receptive field effects are consistent with this model.

1.5.4 Predictive coding networks

In recent years the ideas described above have evolved into modern predictive coding networks. Predictive coding networks are a potential implementation of the free-energy principle, one version of the Bayesian brain view. Friston (2010) relates surprise to a mathematical quantity called free energy, and suggests that 'agents minimize free energy by changing their predictions (perception) or by changing the predicted sensory inputs (action)' (Key Points section). I now outline how predictive coding networks model perception. As Millidge et al. (2021) describe, the approach in Friston (2010) 'reformulates the mostly heuristic Rao and Ballard model in the language of variational Bayesian inference ... tying it the broader project of the Bayesian Brain' (Predictive Coding section).

The basic idea of predictive coding is that the brain involves representations at many different levels of abstraction, arranged in a hierarchy from lower-level sensory to higher-level conceptual features. Each layer in the hierarchy makes predictions about the layer below it, which are compared with the 'real' activity in the layer to calculate a prediction error. The prediction errors are then propagated upwards. During perception, the representations (i.e. patterns of activity across the nodes) are adjusted iteratively to minimise prediction error. During learning, the parameters (i.e. weights of the network) are adjusted to minimise prediction error.

Predictive coding networks are not unique in aiming to minimise prediction error – more established types of neural network such as autoencoders are also trained in this way. Furthermore, autoencoders and predictive coding networks can both be thought of as involving a hierarchy from lower-level sensory features to higher-level latent features, which then project back to sensory features. Both learn compressed representations through self-supervised learning, but there are several key differences.

Firstly, in a predictive coding network, prediction errors are transmitted upwards through the hierarchy, and predictions transmitted downwards. The fact that only errors or residuals are transmitted from lower to high levels distinguishes predictive coding networks from autoencoders, in which data is transmitted upwards. Secondly, a

predictive coding network uses only local learning rules, whereas autoencoders require backpropagation, which involves derivatives of the error being propagated to remote regions of the network to update the weights through gradient descent. As a result, in a predictive coding network the inference and generative weights are adjusted locally, whereas in an autoencoder they are trained from a single reconstruction loss.

See Section C.4 of the Appendix for more detail on predictive coding networks with the help of a simple toy example (Bogacz, 2017).

1.6 Generative models

Generative models are models which represent probability distributions across many variables, enabling them to generate outputs that resemble the data they were trained on. Generative models can be explicit or implicit density models: explicit density models provide a probability estimate for any data item, whereas implicit density models provide a procedure for generating data but not a probability estimate.

There is evidence that generative processing is central to many cognitive functions, including top-down processing, predicting future stimuli, imagination, and memory. These ideas are connected to the predictive brain hypothesis more broadly (Friston & Kiebel, 2009). This section describes some of the main varieties of generative model in more detail.

1.6.1 Generative models and cognition

Generative models have long been implicated in top-down processing, i.e. the effect of prior knowledge and contextual information on perception. Classic examples of top-down processing include pareidolia (Liu et al., 2014), the perception of phantom limbs (Ramachandran & Hirstein, 1998), and the McGurk effect (McGurk & MacDonald, 1976).

In one illustrative example, Leonard et al. (2016) used electrocorticography to explore neural activity when listening to ambiguous stimuli. Words were played to the

participants with some phonemes replaced by noise, in a sentence that provided context. The study found that when participants reported hearing the incomplete sound as a given word, the activity was very similar to the response to the unaltered word. That is, when a word was heard due to phonemic restoration, the cortical activity was very similar to that for the unaltered version, and occurred at the same time as for the unaltered word. Furthermore, stimulus spectrogram reconstruction of the brain activity showed that the spectrograms of the restored words closely matched the spectrogram of the perceived unaltered words. They indicate that when phonemic restoration occurs, the brain generates the missing sound in auditory cortex in real time. This illustrates why top-down processing is hard to explain as inference occurring after perception, but is instead an indication of generative models in the brain.

There are many other reasons why generative models might be useful for cognition. Firstly, generative models underlie the phenomenon of learning through imagination, e.g. in model-based reinforcement learning (Sutton, 1991). In a more recent example, Ha and Schmidhuber (2018) generate recurrent neural network 'world models' of several video game environments, and show that training a reinforcement learning agent in these generated environments gives good results.

Another advantage is that sampling from a generative model during category learning can result in better ability to generalise, as a result of exposing the classifier to more varied data (Barry & Love, 2021). For example, a generative model could extrapolate from a few images of an unfamiliar animal to imagine variants from the same category (e.g. views of the animal from different angles), and a classifier trained on both the original stimuli and generated variants might be more robust. Augmenting classifiers' training data with synthetic examples from generative models is now often used in the machine learning literature, and is especially beneficial if there is little real training data (Trabucco et al., 2023).

1.6.2 Early generative networks

Early generative networks include Boltzmann machines, Helmholtz machines, and deep belief networks. These predecessors of modern generative networks originated in neuroscience, have biologically plausible learning algorithms, and continue to be relevant today (e.g. George et al., 2023), so here I briefly review the main families of model.

Boltzmann machines

The Boltzmann machine, a simpler predecessor to the Helmholtz machine, consists of a set of hidden and a set of visible units (Ackley et al., 1985). These are binary and stochastic, as opposed to the continuous and deterministic activities in a standard feedforward network. Training consists of two phases: in the positive phase, the visible units are 'clamped' to a particular example from the training data. In the negative phase, this example data point is provided to the visible units but activity then propagates freely within the network, until it settles into a stable state. The probability that two units are both 'on' in the positive phase is compared to this probability in the negative phase, with the weight between the two units reduced proportionally to this difference (intuitively, one wants the probabilities to be the same).

In a restricted Boltzmann machine (RBM), visible units are only connected to hidden units and hidden units are only connected to visible units. In others words, an RBM is a fully-connected bipartite graph (or equivalently a network with one hidden layer, and no intra-layer connections). RBMs also compare activity in positive and negative phases, but their reduced connectivity allows the simpler contrastive divergence algorithm to be used (Hinton, 2012). Contrastive divergence works as follows:

1. Start with an example v from the training data and obtain the corresponding hidden layer activity h.
2. Calculate the positive gradient as the outer product of v and h.

3. Reconstruct the visible layer activity v' from the hidden layer activity h, and then obtain a new hidden layer activity h' from this reconstructed input.
4. Calculate the negative gradient as the outer product of v' and the new hidden layer activation h'.
5. Update the weights of the model proportionally to the positive gradient minus the negative gradient.

Deep belief networks (Hinton, 2009) essentially stack together multiple RBMs, allowing more complex data to be captured in a hierarchy of hidden representations.

Helmholtz machines

Helmholtz machines build on Boltzmann machines, and are often thought of as an ancestor of variational autoencoders.

Dayan et al. (1995) 'view the human perceptual system as a statistical inference engine whose function is to infer the probable causes of sensory input' (Introduction). They propose the Helmholtz machine, consisting of a recognition and a generative model that train each other via an unsupervised learning algorithm called the wake-sleep algorithm (Hinton et al., 1995). A recognition model allows the system to infer the causes of sensory data, and a generative model allows the system to infer sensory data from underlying causes, i.e. latent variables.

The Helmholtz machine works by minimising the Helmholtz free energy in the system (Dayan et al., 1995). In brief, we want to maximise the log probability of the observed data given some latent variables. With Bayes' theorem and variational inference, the problem is reformulated as minimising the Helmholtz free energy. (This quantity from statistical physics can very loosely be thought of a measure of 'surprise'.) The Helmholtz machine is designed to do this.

So how is this implemented? A Helmholtz machine has two sets of connections between each pair of layers: top-down connections form a generative model going from latent variables to observations, and bottom-up connections form a recognition

model going from observations to latent variables. The bottom layer of the network represents raw sensory data, but Helmholtz machines can have multiple hidden layers above this (capturing latent variables at increasing levels of abstraction). The recognition and generative models are used to train each other via the wake-sleep learning algorithm, with training using an entirely local delta rule. (As above, this is an advantage from a computational neuroscience perspective, as non-local learning rules have been criticised as being biologically implausible.)

One notable difference compared to a 'standard' neural network is that the neurons in a Helmholtz machine (as in a Boltzmann machine) are stochastic rather than deterministic, and have binary rather than continuous outputs. A binary stochastic neuron outputs one with a probability p and zero with a probability $1\text{-}p$. This gives the network the right properties to minimise the Helmholtz free energy.

During the wake phase, sensory data is passed up through the recognition network. Next, activations are propagated down through the generative network, and the generative weights are adjusted while the recognition weights stay fixed. During the sleep phase, a hidden representation is passed down through the generative network. Next, activations are propagated up through the recognition network, and the recognition weights are adjusted while the generative weights stay fixed. In summary, in the wake phase the generative weights 'are adapted to increase the probability that they would reconstruct the correct activity vector in the layer below', while in the sleep phase, the recognition weights 'are adapted to increase the probability that they would produce the correct activity vector in the layer above' (Hinton et al., 1995, Abstract).

1.6.3 Variational autoencoders

An autoencoder is a neural network which encodes an input into a more compressed representation (in a 'bottleneck' layer with fewer neurons than the input and output layers), and then decodes this back to the original. It learns by minimising the difference between the inputs and outputs. There is no guarantee that decoding

an arbitrary compressed representation produces a sensible output, so standard autoencoders do not perform well as generative models. In other words, there are many 'gaps' in the vector space of the compressed representations which do not correspond to anything meaningful. However, one can train an autoencoder with special properties, such that each latent variable is normally distributed for a given input, allowing one to sample realistic items. The result is called a variational autoencoder (Kingma & Welling, 2013, 2019).

The following description provides the intuition behind the model, rather than the underlying mathematics. Making the latent variables for a given input normally distributed requires the network to have an unusual architecture. Suppose we want our compressed representation to be of dimension n, i.e. we want n latent variables in our variational autoencoder. For these to be normally distributed, we need to produce a mean and standard deviation for each of the n variables for a given input. (Note that the mean and standard deviation vectors are a function of the input, not fixed values for the whole dataset.) Therefore two vectors constitute the compressed representation: one for the means and one for the standard deviations. A sampling step then samples a value for each of the n latent variables. This sampling gives us the desired properties locally. In other words, an input is encoded as a probability distribution, rather than a single vector of latent variables. This means that the model learns to decode not just a single encoding, but a region of the latent space.

As a result, the latent space has the properties we want in the vicinity of each data point. However, if we want to be able to randomly select a point from anywhere in the latent space and generate a realistic output, we need something else too. In addition to the usual reconstruction loss for autoencoders, a special loss function called the Kullback-Leibler divergence is used. The Kullback-Leibler divergence measures how different two probability distributions are; in this case, it measures how much the distribution of each latent variable differs from a normal distribution with a mean of zero and standard deviation of one. In other words, the Kullback-Leibler loss is at a minimum when each latent variable's mean is zero and standard deviation is one. Using this loss moves the distributions towards each other, so that the means are as

nearby as possible whilst still capturing the variation in the data. This means that there are fewer 'gaps' in the latent space, ensuring that variational autoencoders are able to generate realistic outputs.

To describe variational autoencoders more mathematically, their loss function can be derived through variational inference (as exact Bayesian inference is intractable). As Odaibo (2019) describes, let us consider a VAE with input vector x and latent variable vector z. The encoder of the VAE learns to infer latent variables from observed data, i.e. it implements a function $q_\theta(z|x)$, parameterised by the weights θ. The decoder of the VAE learns to infer observed data from latent variables, i.e. it implements a function $p_\phi(x|z)$, parameterised by the weights ϕ.

We start with the first equation below, giving the Kullback-Leibler (KL) divergence (a measure of the difference between two probability distributions) between the approximate and true posteriors. Crucially, the KL divergence is always greater than or equal to zero. By expanding this expression according to the definition of the KL divergence, applying Bayes' theorem, and rearranging the result, we end up with the second equation below, where the right hand side gives the 'evidence lower bound' (ELBO):

$$D_{\mathrm{KL}}(q_\theta(z|x_i)||p(z|x_i)) \geq 0$$

$$\log p(x_i) \geq \mathbb{E}_{q_\theta(z|x_i)}[\log p_\phi(x_i|z)] - D_{\mathrm{KL}}(q_\theta(z|x_i)||p(z))$$

The ELBO can be maximised in order to maximise the left hand side by proxy. Minimising its negative produces the following loss function, in which the first term (the reconstruction loss) promotes accurate reconstruction of input data, and the second term (the KL loss) makes the latent variables approximately normally distributed:

$$\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] + D_{\mathrm{KL}}(q_\theta(z|x) \,||\, p(z))$$

One appeal of variational autoencoders is that it is straightforward to control what is generated. Applications in machine learning include altering the age of a face, by calculating the direction in the latent space that corresponds to age, and moving the latent variables of the input in this direction (Yan et al., 2016). See Figure C.1, adapted from Hou et al. (2017), for an example of this. It is also possible to interpolate between examples by picking a point in the latent space between them, e.g. to blend two songs into a single song (Roberts et al., 2018).

There are a number of applications of variational autoencoders in the computational neuroscience literature. In Van de Ven et al. (2020), replayed events are sampled from a variational autoencoder, and this helps to avoid catastrophic interference. Nagy et al. (2020) propose that semantic memory provides a statistical model of the world, modelled as a variational autoencoder, and argue that this is consistent with gist-based memory distortions. Whittington et al. (2020) present the Tolman-Eichenbaum machine as a model of generalisation in the hippocampal-entorhinal system, in both spatial and non-spatial tasks in which inferences can be made based on structural regularities. As part of their model, the authors use a recurrent graph variational autoencoder.

1.6.4 Generative adversarial networks

The basic concept of generative adversarial networks (Goodfellow et al., 2014) is simple: one trains a generator and discriminator in parallel. The discriminator learns to tell real from generated examples, and the generator learns to generate items that trick the discriminator (given a random input in the first layer). As both models learn, these items become increasingly realistic. This is reminiscent of other machine learning approaches in which the verdict of one model trains another, such as actor-critic methods in reinforcement learning (Sutton, Barto et al., 1998).

This technique has been used to generate images with particular success. Deep convolutional GANs, using a deconvolutional generator combined with a convolutional discriminator, were state of the art for image generation prior to diffusion models

(Radford et al., 2015).

There are several ways to control the output of a GAN. One intuitive method is to do vector arithmetic with the latent vectors, as Radford et al. (2015) show. The authors demonstrate that in the latent vector space, 'man with glasses' minus 'man without glasses' plus 'woman without glasses' gives 'woman with glasses'. This is similar to the vector arithmetic with word vectors demonstrated by Mikolov et al. (2013). Another option for controlling the output is conditional GANs (Mirza & Osindero, 2014).

Gershman (2019) suggests that generative activity in the brain might work like a GAN, with the discriminator corresponding to the 'reality testing' function of the prefrontal cortex (although he describes this theory as speculative). He motivates this in several ways. Firstly, he notes that previous research has generally assumed that the brain learns to estimate specific probabilities for states, much like an explicit density model. But implicit models such as GANs, which provide a sampling method but do not estimate the probability of the resulting items, are easier to learn. Secondly, he suggests that the GAN view accounts better than alternatives for 'the phenomenology of illusion' (Introduction). Thirdly, he proposes that the GAN view might shed light on 'the origins of delusions, hallucinations, and confabulations that arise in certain mental disorders' (Introduction). Gershman (2019) identifies several empirical predictions this model would make, for example that an impaired discriminator would lead to systematic problems with statistical learning. He suggests that patients with schizophrenia or prefrontal damage impairing the proposed discriminator could be studied to see if this is true.

Some types of model have fused the idea of GANs and VAEs. Like VAEs, adversarial autoencoders (AAEs) try to make the latent variables match a desired distribution (Makhzani et al., 2015). The effect is much like a VAE but achieved in a different way. Rather than the VAE's approach of a special sampling architecture combined with a KL loss, the AEA takes inspiration from GANs. A discriminator learns to tell the difference between latent vectors produced by the current AAE, and samples drawn from a normal distribution for each latent variable. The encoder then learns to trick

the discriminator, by making its latent space more like the desired distribution.

1.6.5 Autoregressive sequence models

Another area in which generative models have been very successful is text generation. For example, GPT-2, an autoregressive model trained on the task of predicting the next item in a sequence, surpassed previous benchmarks by a long way (Radford et al., 2019). GPT-2 and subsequent models of this family are explicit density models, in which the probability of a generated sentence is a product of conditional probabilities for each token (chunk of characters) given all the preceding tokens in the sentence.

As an aside, this differs from the other main way that modern language models are trained, masked language modelling. Masked language modelling works by training the model to 'fill in the blanks' (Devlin et al., 2018). The training data is prepared by replacing some fraction of the words with 'mask' tokens, and the model learns to replace these mask tokens with the right words. (This design is more akin to a noise-reducing autoencoder than a generative model.) As a result, each sequence is not assigned a probability like in causal language modelling. The causal language modelling objective means that the prediction is unidirectional, rather than bidirectional like in masked language modelling, i.e. the model learns to predict the next word, rather than missing words based on the surrounding context.

Most modern language models involve transformers (Vaswani et al., 2017), a type of neural network architecture that relies on the attention mechanism to capture complex interdependencies between elements of the input. In recent years transformer-based networks have replaced recurrent and convolutional networks as an approach to this problem.

Autoregressive sequence models are central to the sequential model presented in Chapter Three, so are discussed in more detail there.

1.6.6 Diffusion models

The intuition behind diffusion models is to define a process for iteratively adding noise to an image (the forward diffusion process), then learn to reverse this step by step (the reverse diffusion process) (Ho et al., 2020). The forward diffusion process adds Gaussian noise at each step, and is designed to guarantee that the final result is an approximately isotropic Gaussian distribution. (Furthermore it provides an equation to predict the image at time step t of the diffusion process, without needing to perform forward diffusion at every intervening step.) At each step the reverse diffusion process predicts the noise with a U-Net architecture, in which a sequence of downsampling then upsampling layers are applied, then deduces the corresponding image. (In simple terms a U-Net can be thought of as an autoencoder with special features, as it compresses then decompresses the image data.) Note that some approaches predict the denoised image directly instead (Ramesh et al., 2022).

The training algorithm works as follows. Firstly, a time step t (between 0 and T) is randomly chosen, and a sample is taken from the normal distribution $N(0, 1)$. Using these inputs, the forward diffusion process is applied, producing the noisy image at time step t. The U-Net takes the noisy image and time step as inputs, and outputs the predicted noise. The loss function is then the difference between the predicted noise at time t and the true noise (the sample from $N(0, 1)$). The U-Net's weights are adjusted using this loss.

The sampling algorithm works as follows. Starting with pure Gaussian noise, an image is generated by reversing the forward diffusion process from time step T (the last step of the process) to 0. At each time step the trained U-Net is used to predict the noise given the image x_t and the time step. In addition, a new sample from the normal distribution is taken. Using these components (x_t, the predicted noise, the sample from $N(0, 1)$, and scheduling information), an equation predicts the image x_{t-1}, which is then the input for the next step of reverse diffusion.

1.6.7 Generating images from text

Advances in machine learning models for image generation from text may shed light on the interactions between semantic memory and event construction. DALL-E 2 (Ramesh et al., 2022) builds on two key components: diffusion models, which learn to reverse the process of adding noise to an image as described above, and CLIP.

CLIP (Radford et al., 2021) is a multimodal embedding model. Radford et al. (2021) present a network which learns to represent both language and imagery in the same vector space, through the process of 'Contrastive Language-Image Pretraining'. As with most embedding models, the training process gradually moves matching pairs closer and non-matching pairs further apart in the space. For CLIP, 'matching pairs' are text-image pairs which occurred together in a large dataset of web content, for example an image and its caption. The result is a multimodal model which supports text-to-image searching, amongst other things.

To generate an image from a text input with DALL-E 2 (Ramesh et al., 2022), a CLIP text embedding is first obtained for the caption, and is then converted into a CLIP image embedding with a 'prior model'. (This is necessary because CLIP's training objective encourages matching text to be closer to a given image than any non-matching text, but the image and text representations still differ, and require another model to 'translate' between them. This makes sense as there may be many equally valid images for one text input.)

A diffusion model, unCLIP, based on its predecessor GLIDE (Nichol et al., 2021), then generates an image conditioned on the CLIP image embedding. During training, unCLIP is provided with both a degraded form of the target image and the CLIP image embedding of the original, and its weights are adjusted to reconstruct the original as well as possible. Specifically, a lower-dimensional projection of the CLIP image embedding is added to the timestep embedding (this is used in unconditional diffusion models to tell the U-Net the current timestep, which indicates how much noise there is in the input image). Finally, another diffusion model increases the resolution of the image.

DALL-E 2 (Ramesh et al., 2022) was designed to allow unconditional image generation too, i.e. image generation from scratch without a text input. This was achieved by removing a subset of text inputs in the training data. Random regions of images were also removed to enable 'inpainting', i.e. filling in missing regions in an image.

Chapter 2

A generative model of memory construction and consolidation

2.1 Introduction

The model presented in this chapter draws together existing ideas in machine learning to suggest an explanation for the following key features of memory, only subsets of which are captured by previous models:

1. The initial encoding of memory requires only a single exposure to an event, and depends on the hippocampal formation (HF), while the consolidated form of memory is acquired more gradually (Alvarez & Squire, 1994; Marr, 1970, 1971), as in the complementary learning systems model (CLS; McClelland et al., 1995).

2. The semantic content of memories becomes independent of HF over time (Manns et al., 2003; Squire et al., 2015; Vargha-Khadem et al., 1997), consistent with CLS.

3. Vivid, detailed episodic memory remains dependent on HF (McKenzie & Eichenbaum, 2011), consistent with multiple trace theory (Nadel & Moscovitch, 1997)

(but not CLS).

4. Similar neural circuits are involved in recall, imagination, and episodic future thinking (Addis et al., 2007; Hassabis & Maguire, 2007), suggesting a common mechanism for event generation, as modelled in spatial cognition (Bicanski & Burgess, 2018).

5. Consolidation extracts statistical regularities from episodic memories to inform behaviour (Durrant et al., 2011; Richards et al., 2014), and supports relational inference and generalisation (Ellenbogen et al., 2007). The Tolman-Eichenbaum machine (TEM; Whittington et al., 2020) simulates this in the domain of multiple tasks with common transition structures (see also Kumaran et al., 2016), while Schapiro et al. (2017) model how both individual examples and statistical regularities could be learned within HF.

6. Post-consolidation episodic memories are more prone to schema-based distortions, in which semantic or contextual knowledge influences recall (Bartlett, 1932; Payne et al., 2009), consistent with the behaviour of generative models (Nagy et al., 2020).

7. Neural representations in entorhinal cortex (EC) such as grid cells (Hafting et al., 2005) are thought to encode latent structures underlying experiences (Constantinescu et al., 2016; Whittington et al., 2020), and other regions of association cortex, such as medial prefrontal cortex (mPFC), may compress stimuli to a minimal representation (Mack et al., 2020).

8. Novelty is thought to promote encoding within HF (Hasselmo et al., 1996), while more predictable events consistent with existing schemas are consolidated more rapidly (Tse et al., 2007). Activity in the hippocampus can reflect prediction error or mismatch novelty (Chen et al., 2011; Kumaran & Maguire, 2006), and novelty is believed to affect the degree of compression of representations in memory (Hedayati et al., 2022) to make efficient use of limited HF capacity (Benna & Fusi, 2021).

9. Memory traces in hippocampus appear to involve a mixture of sensory and conceptual features, with the latter encoded by concept cells (Quiroga, 2012), potentially bound together by episode-specific neurons (Kolibius et al., 2021). Few models explore how this could happen.

2.1.1 Consolidation as the training of a generative model

This chapter proposes that the initial representation of memories can be used to train a generative network, which learns to reconstruct memories by capturing the statistical structure of experienced events (or 'schemas'). First, the hippocampus rapidly encodes an event, then generative networks gradually take over, after being trained on replayed representations from the hippocampus. This makes the memory more abstracted, more supportive of generalisation and relational inference, but also more prone to gist-based distortion. The generative networks can be used to reconstruct (for memory) or construct (for imagination) sensory experience, or to support semantic memory and relational inference directly from their latent variable representations (see Figure 2.1).

Before consolidation, the hippocampal autoassociative network encodes the memory. A modern Hopfield network (Ramsauer et al., 2020) is used, which can be interpreted such that the feature units activated by an event are bound together by a memory unit (Krotov & Hopfield, 2020). Teacher-student learning (Hinton et al., 2015) allows transfer of memories from one neural network to another during consolidation (Sun et al., 2021). Accordingly, outputs from the autoassociative network are used to train the generative network: random inputs to the hippocampus result in the reactivation of memories, and this reactivation results in consolidation. After consolidation, generative networks encode the information contained in memories. Reliance on the generative networks increases over time as they learn to reconstruct a particular event.

Specifically, the generative networks are implemented as variational autoencoders (VAEs), which are autoencoders with special properties such that the most compressed

layer represents a set of latent variables, which can be sampled from to generate realistic new examples corresponding to the training dataset (Kingma & Welling, 2013, 2019). Latent variables can be thought of as hidden factors behind the observed data, and directions in the latent space can correspond to meaningful transformations. The VAE's encoder *encodes* sensory experience as latent variables, while its decoder *decodes* latent variables back to sensory experience. In psychological terms, after training on a class of stimuli VAEs can reconstruct such stimuli from a partial input, according to the schema for that class, and generate novel stimuli consistent with the schema. The use of VAEs is illustrative, and one would expect a range of other generative latent variable models, such as predictive coding networks (Dayan et al., 1995; Friston, 2010; Rao & Ballard, 1999), to show similar behaviour.

Generative networks capture probability distributions underlying events, or 'schemas'. In other words, here 'schemas' are rules or priors (expected probability distributions) for reconstructing a certain type of stimulus (e.g. the schema for an office predicts the presence of co-occurring objects like desks and chairs, facilitating episode generation), whereas concepts represent categories but not necessarily how to reconstruct them. However, schemas and concepts are closely related, and their meanings can overlap, with conflicting definitions in the psychology literature (Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017).

During perception, the generative model provides an ongoing estimate of novelty from its reconstruction error (a.k.a. 'prediction error', the difference between input and output representations). Aspects of an event that are consistent with previous experience (i.e. with low reconstruction error) do not need to be encoded in detail in the autoassociative 'teacher' network (Bein et al., 2021; Biderman et al., 2020; Schacter et al., 2007; Sherman et al., 2022). Once the generative network's reconstruction error is sufficiently low, the hippocampal trace is unnecessary, freeing up capacity for new encodings. However, I have not simulated decay, deletion or capacity constraints in the autoassociative memory part of the model.

2.1.2 Combining conceptual and sensory features in episodic memory

Consolidation is often considered in terms of fine-grained sensory representations updating coarse-grained conceptual representations, e.g. the sight of a particular dog updating the concept of a dog. Modelling hippocampal representations as sensory-like is a reasonable simplification, which I make in simulations of the 'basic' model in Figure 2.1. However, memories probably bind together representations along a spectrum from coarse-grained and conceptual to fine-grained and sensory. For example, the hippocampal encoding of a day at the beach is likely to bind together coarse-grained concepts like 'beach' and 'sea' along with sensory representations like the melody of an unfamiliar song, or sight of a particular sandcastle, consistent with the evidence for concept cells in hippocampus (Quiroga, 2012). (This also fits with the observation that ambiguous images 'flip' between interpretations in perception, but are stable when held in memory (Chambers & Reisberg, 1985), reflecting how the conceptual content of memories constrains recall.)

Furthermore, encoding every sensory detail in the hippocampus would be inefficient (elements already predicted by conceptual representations being redundant); an efficient system should take advantage of shared structure across memories to encode only what is necessary (Barlow et al., 1961; Barlow, 1989). Accordingly, I suggest that predictable elements are encoded as conceptual features linked to the generative latent variable representation, while unpredictable elements are encoded in a more detailed and veridical form as sensory features.

Suppose someone sees an unfamiliar animal in the forest (Figure 2.2b). Much of the event might be consistent with an existing forest schema, but the unfamiliar animal would be novel. In the extended model (Figure 2.2, Section 2.3.5) the reconstruction error per element of the experience is calculated by the generative model during perception, and elements with high reconstruction error are encoded in the autoassociative network as sensory features, along with conceptual features linked to the generative model's latent variable representation. In other words, each

pattern is split into a predictable component (approximating the generative network's prediction for the pattern), plus an unpredictable component (elements with high prediction error). This produces a sparser vector than storing every element in detail, increasing the capacity of the network (Benna & Fusi, 2021).

2.1.3 Neural substrates of the model

Which brain regions do the components of this model represent? The autoassociative network involves the hippocampus binding together the constituents of a memory in the neocortex, whereas the generative network involves neocortical inputs projecting to latent variable representations in higher association cortex, which then project back to neocortex via the HF. The entorhinal cortex (EC), medial prefrontal cortex (mPFC), and anterolateral temporal lobe (alTL) are all prime candidates for the site of latent variable representations.

Firstly, EC is the main route between the hippocampus and neocortex, and where grid cells are most often observed (Moser et al., 2008), which are thought to be a latent variable representation of spatial or relational structure (Constantinescu et al., 2016; Whittington et al., 2020). Secondly, mPFC and its connections to HF play a crucial role in episodic memory processing (Benchenane et al., 2010; Frankland & Bontempi, 2005; Gais et al., 2007; Gilboa & Marlatte, 2017; Takashima et al., 2006; Van Kesteren et al., 2010), are thought to encode schemas (Ghosh & Gilboa, 2014; Tse et al., 2007), are implicated in transitive inference (Koscik & Tranel, 2012) and the integration of memories (Spalding et al., 2018), and perform dimensionality reduction by compressing irrelevant features (Mack et al., 2020). Thirdly, the anterior and lateral temporal cortices associated with semantic memory (Chan et al., 2001) and retrograde amnesia (Bright et al., 2006) likely contain latent variable representations capturing semantic structure. This might correspond to the 'anterior temporal network' associated with semantic dementia (Ranganath & Ritchey, 2012), while the first network (between sensory and entorhinal cortices) might correspond to the 'posterior medial network' (Ranganath & Ritchey, 2012), and to the network mapping between visual scenes and allocentric spatial representations (Becker & Burgess, 2000;

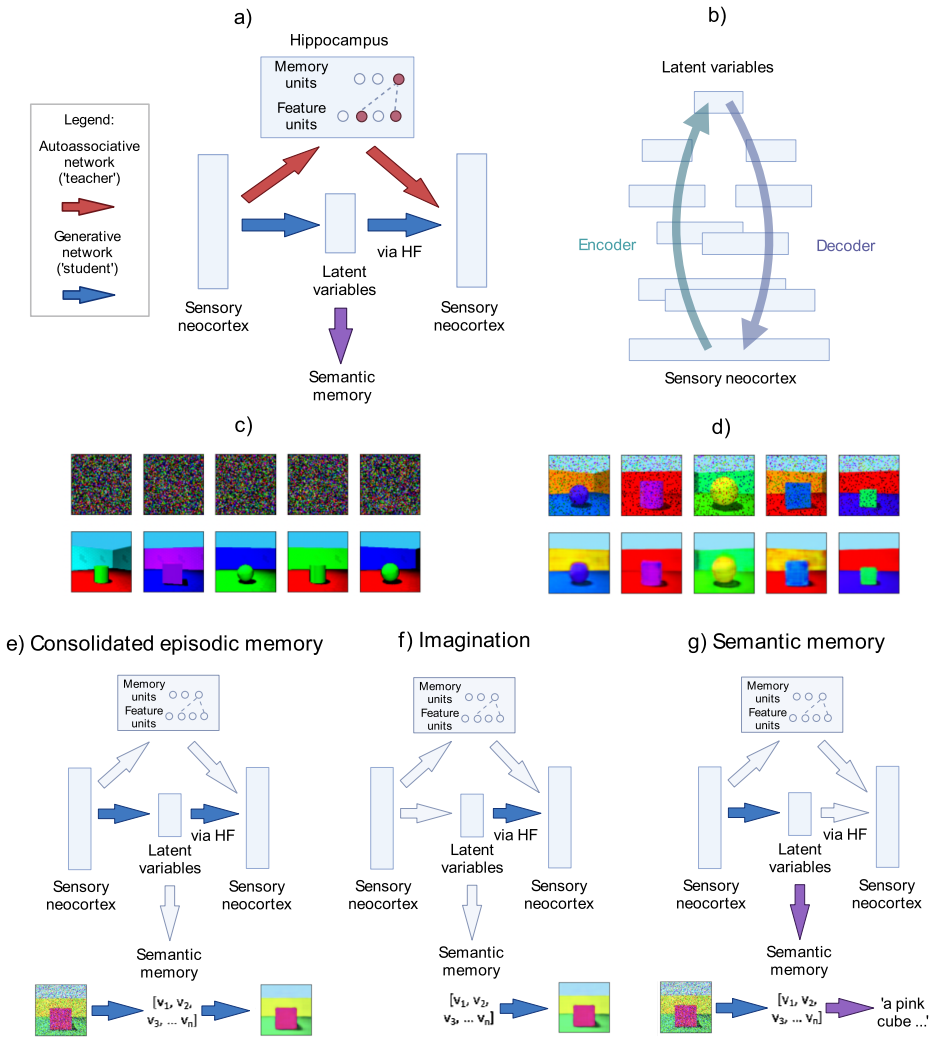

Figure 2.1: Architecture of the basic model. a) First the hippocampus rapidly encodes an event, modelled as one-shot memorisation in an autoassociative network (a modern Hopfield network). Then generative networks are trained on replayed representations from the autoassociative network, learning to reconstruct memories by capturing the statistical structure of experienced events. b) A more detailed schematic of the generative network to indicate the multiple layers of, and overlap between, the encoder and decoder (where layers closer to sensory neocortex overlap more). The generation of a sensory experience, e.g. visual imagery, requires the decoder to sensory neocortex via HF. c) Random noise inputs to the modern Hopfield network (upper row) reactivate its memories (lower row) after 10,000 items from the Shapes3D dataset are encoded, with five examples shown. d) The generative model (a variational autoencoder) can recall images (lower row) from a partial input (upper row), following training on 10,000 replayed memories sampled from the modern Hopfield network. e) Episodic memory after consolidation: a partial input is mapped to latent variables whose return projections to sensory neocortex via HF then decode these back into a sensory experience. f) Imagination: latent variables are decoded into an experience via HF and return projections to neocortex. g) Semantic memory: a partial input is mapped to latent variables, which capture the 'key facts' of the scene. The bottom rows of parts e-g) illustrate these functions in a model that has encoded the Shapes3D dataset into latent variables $[v_1, v_2, v_3 \ldots v_n]$. (Diagrams were created using BioRender.com.)

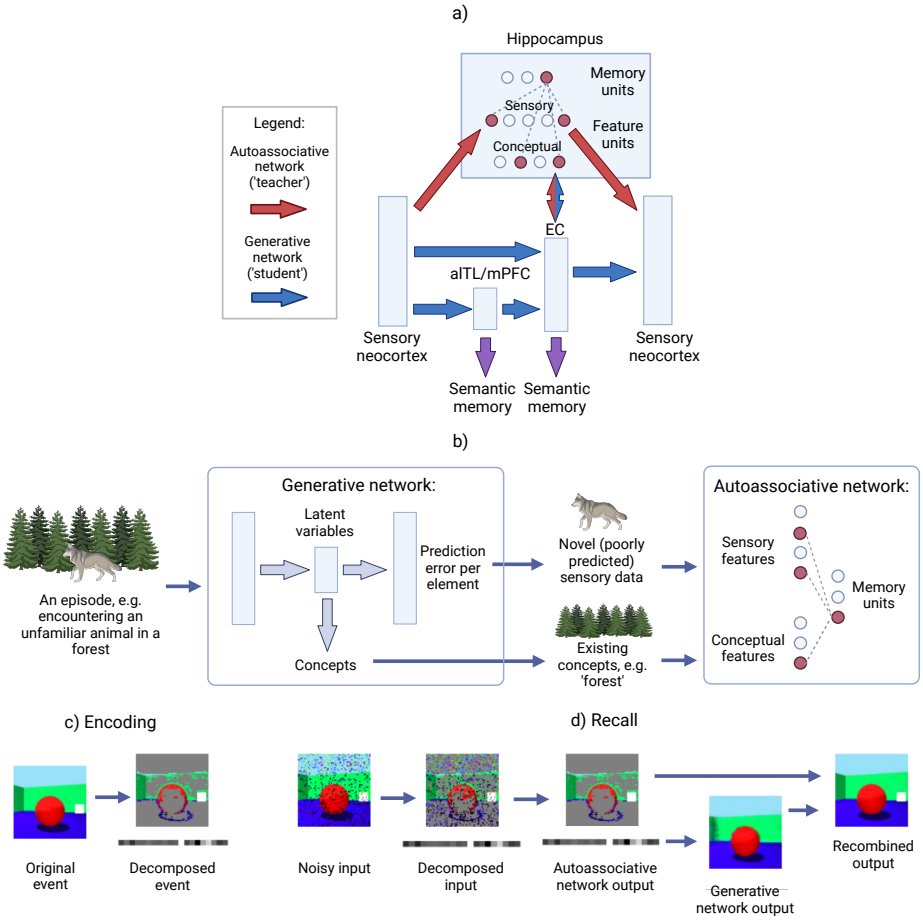

Figure 2.2: Architecture of the extended model. a) Each scene is initially encoded as a combination of predictable conceptual features related to the latent variables of the generative network and unpredictable sensory features that were poorly predicted by the generative network. A modern Hopfield network (in red) encodes both sensory and conceptual features (with connections to sensory neocortex and latent variables in EC respectively), binding them together via memory units. Memories may eventually be learned by the generative model (in blue), but consolidation can be a prolonged process, during which time the generative network provides schemas for reconstruction and the autoassociative network supports new or detailed information not yet captured by these schemas. Multiple generative networks can be trained concurrently, with different networks optimised for different tasks. This includes networks with latent variables in EC, medial prefrontal cortex (mPFC), and anterolateral temporal lobe (alTL), each with their own semantic projections. But in all cases, return projections to sensory neocortex are via HF. b) An illustration of encoding in the extended model. c) Encoding 'scenes' from the Shapes3D dataset, with each 'scene' decomposed into unpredicted sensory features (above) and conceptual features linked to the generative network's latent variables (below). d) Recalling 'scenes' from the Shapes3D dataset. First the input is decomposed, then the modern Hopfield network performs pattern completion on both sensory and conceptual features. The conceptual features (which in these simulations are simply the generative network's latent variables) are then decoded into a schema-based prediction, onto which any stored sensory features are overwritten. (Diagrams were created using BioRender.com.)

Bicanski & Burgess, 2018; Byrne et al., 2007).

Which regions constitute the generative network's decoder? The decoder converts latent variable representations in higher association cortex back to sensory neocortical representations via HF. Patients with damage to the hippocampus proper but not EC can generate simple scenes (or fragments thereof), but an intact hippocampus is required for more coherent imagery of complex ones (Hassabis et al., 2007). One might hypothesise that conceptual units in the hippocampus proper help to generate complex, conceptually coherent scenes (perhaps through a recurrent 'clean up' mechanism), but that an intact EC and its return pathway to sensory neocortex (the ventral visual stream for images) can still decode representations to some extent in their absence.

Multiple generative networks can be trained concurrently from a single autoassociative network through consolidation, with different networks optimised for different tasks. In other words, multiple networks could update their parameters to minimise prediction error based on the same replayed memories. This could consist of a primary generative network with latent variables in EC, plus additional parallel pathways from higher sensory cortex to EC via latent variables in medial prefrontal cortex (mPFC) or anterolateral temporal lobe (alTL). (Computationally, the shared connections could be fixed as the alternative pathways are trained.) Note that in all cases, return projections to sensory neocortex via HF are required to decode latent variables into sensory experiences.

2.2 Methods

2.2.1 Data

In the simulations, images represent events, with the exception of the Deese-Roediger-McDermott (Deese, 1959; Roediger & McDermott, 1995) task stimuli. The Shapes3D dataset (Burgess & Kim, 2018) is used throughout, except for the use of MNIST (LeCun et al., 2010) to explore certain distortions. Note that one modern Hopfield

network was used per dataset, and one generative model was trained per dataset from the corresponding modern Hopfield network's outputs.

2.2.2 Basic model

In this model, the hippocampus rapidly encodes an event, modelled as one-shot memorisation in an autoassociative network (a modern Hopfield network). Then generative networks are trained on replayed representations from the autoassociative network, learning to reconstruct memories by capturing the statistical structure of experienced events.

The generative networks used are variational autoencoders (VAEs). As described above, a VAE is a type of autoencoder designed in such a way that randomly sampling values for the latent variables in the model's 'bottleneck' layer generates valid stimuli (Kingma & Welling, 2013). Figure C.1, adapted from Hou et al. (2017), shows how directions in the latent space can correspond to meaningful transformations. Whilst most diagrams show the VAE's input and output layers in sensory neocortex as separated (in line with conventions for visualising neural networks), it is important to note that the input and output layers are in fact the same, as shown in Figure 2.1b. There may be considerable overlap between the encoder and decoder, especially closer to sensory neocortex, but this is not modelled explicitly. The autoassociative model is a modern Hopfield network, with the property that even random input values will retrieve one of the stored patterns via pattern completion. Specifically, I consider the biological interpretation of the modern Hopfield network as feature units and memory units suggested by Krotov and Hopfield (2020).

Consolidation is modelled as teacher-student learning, where the autoassociative network is the 'teacher' and the generative network is the 'student', trained on replayed representations from the 'teacher'. Random noise (consisting of uniformly sampled values in each channel for each pixel) is given as an input to the modern Hopfield network, then the outputs of the network are used to train the VAE. (These outputs represent the high-level sensory representations activated by hippocampal

pattern completion, via return projections to sensory cortex.) The noise input to the autoassociative network could potentially represent random activation during sleep (González et al., 2020; Pezzulo et al., 2021; Stella et al., 2019). Attributes such as reward salience may also influence which memories are replayed, but are not modelled here (Igata et al., 2021).

During the encoding state in the simulations, images are stored in a continuous modern Hopfield network with a high inverse temperature, β, set to 20 (higher values of β produce attractor states corresponding to individual memories, while lower values of β make metastable states more likely). Ramsauer et al. (2020) provide a Python implementation of modern Hopfield networks that I use in my code. During the 'rest' state, random noise is given as an input N times to the modern Hopfield network, retrieving N attractor states from the network. (The distribution of retrieved attractor states was not tested, but was approximately random, and very few spurious attractors were observed with sufficiently high inverse temperature). In the main simulations, 10,000 items from the Shapes3D dataset are encoded in the modern Hopfield network, and 10,000 replayed states are used to train the VAE (i.e. N is 10,000). Rather than replaying new samples from the MHN at each epoch of the VAE's training, a single set of samples is used for efficiency and simplicity.

A VAE is then trained on the 'replayed' images from the modern Hopfield network, using the Keras API for TensorFlow (Abadi et al., 2016). The loss function (i.e. the error minimised through training) is the sum of two terms, the reconstruction error and the Kullback-Leibler divergence (Kingma & Welling, 2013); the former encourages accurate reconstruction, while the latter (which measures the divergence between the latent variables and a Gaussian distribution) encourages a latent space one can sample from. Specifically, the reconstruction loss in the model is a mean absolute error loss. (Note that the terms reconstruction error and prediction error are used interchangeably throughout this chapter.)

The stochastic gradient descent method used is the AMSGrad variant of the Adam optimiser, with early stopping enabled, for a maximum of 50 epochs (where an epoch is a complete pass through the training set). A latent variable vector length of 20,

learning rate of 0.001 and Kullback-Leibler weighting of 1 were used in the main results. The VAEs were not optimised for performance, as their purpose is illustrative (more data and hyperparameter tuning would be likely to improve reconstruction accuracy). Architectural choices within the VAE are not principled, but are based on successful architectures for similar stimuli in the literature. See Section C.1 of the Appendix for details of the VAE's architecture. The VAEs are trained using gradient descent and back-propagation as usual.

Whilst this is not modelled explicitly, once the generative network's reconstruction error is sufficiently low, the hippocampal trace is unnecessary. As a result it could be 'marked for deletion' or overwritten in some way, freeing up capacity for new encodings. However, I have not simulated decay, deletion or capacity constraints in the autoassociative memory part of the model. In these simulations, the main cause of forgetting would be interference from new memories in the generative model.

Note that throughout the simulations, the input to recall is a noisy version of the encoded stimulus image. Specifically, noise is added by replacing a random fraction (0.1 unless stated otherwise) of values in the image array by zero.

Whilst I use only one modality at a time (imagery for the majority of simulations, text for the DRM task), the model is compatible with the multimodal nature of experience, as multimodal inputs to VAEs are possible, which result in a multimodal latent space (Khattar et al., 2019). This could reflect the multimodal nature of concept cells in the hippocampus (Quiroga, 2012).

2.2.3 Modelling semantic memory

Semantic memory is modelled as the ability to decode latent variables into semantic information, without the need to reconstruct the event episodically.

Decoding accuracy is measured by training a support vector machine to classify the central object's shape from the network's latent variables, using 200 examples at the end of each epoch, and measuring classification accuracy on a held-out test set.

2.2.4 Modelling imagination and inference

In the generative network, new items can either be generated from externally specified (or randomly sampled) latent variables (imagination), or by transforming the latent variable representations of specific events (relational inference).

Imagination is simulated by sampling from categories in the latent space then decoding the results. In Figure 2.3d, examples of the four different object shapes are generated by Monte Carlo sampling for simplicity, i.e. samples from the latent space are classified by the semantic decoding classifier, and examples that activate each category are displayed. (Note that there are many alternative ways to do this, e.g. by extracting the decision boundaries from the classifier and sampling within the region corresponding to each class.) Generating imagined scenes from more naturalistic inputs, e.g. natural language descriptions, would require a much more sophisticated text to latent space model, but recent machine learning advances suggest this is possible (Ramesh et al., 2022; Ramesh et al., 2021).

Inference is simulated by interpolating between the latent representations of events (Figure 2.3c), or by doing vector arithmetic in the latent space (Figure 2.3b). To demonstrate interpolation, each row of Figure 2.3c shows items generated from latent variables along a line in the latent space between two real items from the training data. To demonstrate vector arithmetic, each equation in Figure 2.3b shows $result = vector_A + (vector_B - vector_C)$ (reflecting relational inference problems of the form 'what is to A as B is to C?'), where the result is produced by taking the relation between $vector_B$ and $vector_C$, applying that to $vector_A$, and decoding the result. In other words, the three items on the right of each equation in Figure 2.3b are real items from the training data. Their latent variable representations are combined as vectors according to the equation shown, giving the latent variable representation from which the first item is generated. Thus the pair in brackets describes a relation which is applied to the first item on the right to produce the new item on the left of the equation.

2.2.5 Modelling schema-based distortions

The hypothesis that items recalled by the generative network become more prototypical can be tested with the basic model, but in this simulation the MNIST digits dataset (LeCun et al., 2010) is used to exemplify ten clearly defined classes of items (see Figure 2.4a). To measure this distortion quantitatively in Figure 2.4b, I calculated the intra-class variation, defined as the median variance per pixel within each MNIST class, before and after recall, for 5000 images from the test set. (See Section C.1 of the Appendix for details of the model architecture.)

To visualise the explanation for this, the pixel and latent spaces before and after recall (of 2000 images from the MNIST test set) were projected into 2D with UMAP (McInnes et al., 2018), a dimensionality reduction method, and colour-coded by class (see Figure 2.4c-d).

2.2.6 Modelling boundary extension and contraction

Boundary extension is the tendency to remember a wider field of view than was observed for certain stimuli (Intraub & Richardson, 1989), while boundary contraction is the tendency to remember a narrower one (Bainbridge & Baker, 2020a). Whether boundaries are extended or contracted seems to depend on the perceived distance of the central object, with unusually close-up (i.e. 'object-oriented') views causing boundary extension, and unusually far away (i.e. 'scene-oriented') views causing boundary contraction (Bainbridge & Baker, 2020a).

Boundary extension and contraction were tested in the basic model by giving it a range of artificially 'zoomed in' or 'zoomed out' images, adapted from Shapes3D scenes not seen during training, and observing the outputs. The 'zoomed in' view is produced by removing n pixels from the margin. The 'zoomed out' view is produced by extrapolating the pixels at the margin outwards by n additional pixels. (In both cases the new images were then resized to the standard size.) The zoom level is the ratio of the central object size in the output image to the size in the original image, given as a percentage, e.g. an image with a zoom level of 80%, or a ratio of 0.8, is

produced by adding a margin so that the object size is 80% of the original size. As the Shapes3D images are of width and height 64, the number of pixels to add or remove is given by $margin = (32/ratio) - 32$.

In Figure 2.5c, the change in object size between the noisy input and output is estimated as follows: first the image is converted to a few colours by k-means clustering of pixels. Then the colour of the central object is determined by finding the predominant colour in a particular central region of the image. A 1D array of pixels corresponding to a vertical line at the horizontal midpoint of the image is processed to identify the fraction of pixels of the central object colour. This enables the change in object size to be calculated, which is plotted against the degree of 'zoom'. (For this object size estimation approach to work, I filter the Shapes3D dataset to images where the object colour is different from both the wall and floor colour, and additionally to cubes to minimise shadow.)

Note that the measure of boundary extension vs. contraction displayed in Figure 2.5c, reproduced from Park et al. (2021), is not based on the degree of distortion, but is produced by averaging 'closer' vs. 'further' judgements of an identical stimulus image in comparison to the remembered image. This differs from the measure in Figure 2.5c, which instead corresponds to the drawing-based measure in Bainbridge and Baker (2020a), however these measures have been shown to be correlated (Bainbridge & Baker, 2020a).

2.2.7 Extended model

The extended model is designed to capture the fact that memory traces in hippocampus bind together a mixture of sensory and conceptual elements, with the latter encoded by concept cells (Quiroga, 2012), and the fact that schemas shape the reconstruction of memories even prior to consolidation, as shown by the rapid onset of schema-based distortions (Deese, 1959; Roediger & McDermott, 1995).

In the extended model, each scene is initially encoded as the combination of a predictable and an unpredictable component. The predictable component consists

of concepts captured by the latent variables of the generative network, and the unpredictable component consists of parts of the stimuli that were poorly predicted by the generative network. Thus the Modern Hopfield Network model has both conceptual and sensory feature units which store the predictable and unpredictable aspects of memory respectively. Whilst memories may eventually become fully dependent on the generative model, consolidation can be a prolonged process, during which the generative network provides schemas for reconstruction and the autoassociative network supports new or detailed information not yet captured by schemas. (The VAE trained in the basic model simulations was used in the extended model simulations described below.)

How does encoding work in the simulations? For a new image, the prediction error of each pixel is calculated by the VAE (simply the magnitude of the difference between the VAE's input and output). Those pixels with a reconstruction error above the threshold constitute the unpredictable component, while the VAE's latent variables constitute the predictable component, and these components are combined into a single vector and encoded in the modern Hopfield network. Note that when the threshold is zero, the reconstruction is guaranteed to be perfect, but as the threshold increases, the reconstruction decreases in accuracy.

How does recall work prior to full consolidation? After decomposing the input into its predictable (conceptual) and unpredictable (sensory) components, as described above, the autoassociative network can retrieve a memory. The image corresponding to the conceptual component must then be obtained by decoding the stored latent variables. Next, the predictable and unpredictable elements are recombined, simply by overwriting the initial schematic reconstruction in sensory neocortex with any stored (i.e. non-zero) sensory features in hippocampus. Figure 2.6a-b shows this process. The lower the error threshold for encoding sensory details, the more information is stored in the autoassociative network, reducing the reconstruction error of recall (see also Section 2.3.4).

How does replay work? When the autoassociative network is given random noise, both the unpredictable elements and the corresponding latent variables are retrieved.

In Figure 2.6d, the square images show the unpredictable elements of MNIST images and the rectangles below these display the vector of latent variables. As the generative model improves, the presence of hippocampal sensory features that no longer differ from the initial reconstruction indicates that the hippocampal representation is no longer needed.

Note that the latent variable representation is not stable as the generative network learns. If some latent variables are stored in the autoassociative network while the VAE continues to change, the quality of the VAE's reconstruction will gradually worsen; this is also a feature of previous models (Benna & Fusi, 2021). Some degree of degradation may reflect forgetting, but consolidation can be a prolonged process, and hippocampal representations can persist in this time. Therefore it seems more likely that concepts derived from latent variables are stored than the latent variables themselves, promoting the stability of hippocampal representations. (For example, in humans language provides a set of relatively persistent concepts, stabilised by the need to communicate.) Projections from the latent variables can classify attributes with only a small amount of training data (see Section 2.3.2); there could be a two-way mapping between latent variables and concepts, which supports categorisation of incoming experience as well as semantic memory. However, for simplicity the conceptual features are simply a one-to-one copy of latent variable representations in these simulations.

It may also be possible to stabilise the latent variable representations by reducing catastrophic forgetting in the generative network, e.g. by using generative as well as hippocampal replay (Káli & Dayan, 2004; Van de Ven et al., 2020; Van de Ven & Tolias, 2018), with the generative network trained on its own self-generated representations in addition to new memories. Chapter Four explores this idea further. This builds on previous research suggesting certain stages of sleep are optimised to preserve remote memories, while other consolidate new ones (Singh et al., 2022). This could reduce interference of new learning with remote memories in the generative network as well as making hippocampal representations in the extended model more stable.

2.2.8 Modelling schema-based distortions in the extended model

I simulate the contextual modulation of memory as in Carmichael et al. (1932) in the extended model by manipulating the conceptual component of an 'event'. To model an external conceptual context being encoded, the original image is stored in the autoassociative network along with activation of a given concept, represented as the latent variables for that class. Whilst in most simulations the latent variables stored in the modern Hopfield network are simply the output of the VAE's encoder, here an external context activates the conceptual representation, consistent with activity in EC, mPFC, or alTL driven by extrinsic factors.

During recall, a noisy input is processed by the generative network to produce a predicted conceptual feature and the sensory features not predicted by the prototype for that concept, for input to the autoassociative MHN. Pattern completion in the MHN produces the originally encoded sensory and conceptual features, and these are recombined to produce the final output.

The Deese-Roediger-McDermott (DRM) task is a classic way to measure gist-based memory distortion (Deese, 1959; Roediger & McDermott, 1995). Next I demonstrate the rapid onset of semantic intrusions in this task in the extended model, coming about as a consequence of learning the co-occurrence statistics of words in a text dataset representing 'background knowledge'. This follows on from previous work showing that VAEs produce semantic intrusions (Nagy et al., 2020).

In brief, the DRM task involves showing participants a list of words that are semantically related to a 'lure word', which is not present in the list. There is a tendency for both false recognition and false recall of the lure word. I focus on modelling the recall task, but the same model could be extended to recognition (with recognition memory measured by the reconstruction error of the network).

The generative network was pre-trained on a set of word lists extracted from simple stories (Mostafazadeh et al., 2016), representing learning from replayed memories

prior to the DRM stimuli (although replay was not simulated explicitly). Words occurring in fewer than 0.05% or more than 10% of documents were discarded, in order to keep the vocabulary to a manageable size of 4206 words (this meant that some rarer words in the DRM lists were removed). The word lists were converted to vectors of word counts of length 4206, in which the value at index i of the vector for a given list indicated the count of word i in the document. As these representations ignore word order, a sequential model is not required (however this prevents exploring the effect of list position on recall).

Specifically, the variational autoencoder used for this simulation consists of an input layer followed by a dropout layer projecting to 300 latent variables (sampled from representations of the mean and log variance vectors as usual), which then project to an output layer with a sigmoid activation, so that predictions are between zero and one, and L1 regularisation to promote sparsity in this layer. As above, this was implemented using the Keras API for the TensorFlow library (Abadi et al., 2016; Chollet et al., 2015), with the VAE trained to reconstruct input vectors in the usual way.

Following pre-training of the generative network, the system encodes the DRM stimuli, with each of the 20 word lists represented as vectors of word counts. One important detail was the addition of a term, given by 'id_n' for the nth document in the corpus, representing the unique spatiotemporal context of each word list. Note that this is a highly simplified representation of the spatiotemporal context (Howard & Kahana, 2002) for illustration. This enabled recall to be modelled by presenting the network with the 'id_n' term, and seeing which terms were retrieved.

In the extended model, the latent representation of the word list is encoded in the MHN as the conceptual component, while the unique 'id_n' terms are encoded veridically (as vectors of word counts of length 4226 - the original vocabulary size plus the 20 new 'id_n' terms - with one at 'id_n' and zero elsewhere). The sparse vector representing the unexpected 'id_n' term is analogous to the sparse arrays of poorly predicted pixels in the main simulations of the extended model.

When the MHN is given 'id_n' as an input, it retrieves the hippocampal trace, consisting of 'id_n' together with the latent representation of the word list. The latent representation is then decoded to produce the outputs shown in Figure 2.8a (a dotted line shows a threshold for recall, interpreting the output as a probability so that words with an output greater than 0.5 are recalled).

To test the effect of varying the number of associates, as in Robinson and Roediger (1997), subsets of the DRM lists were encoded in the way described above. Specifically, to test the probability of lure recall with n associates studied, n items from each DRM list were encoded. For each list, this was repeated for 20 randomly sampled combinations of n items. Once again recall was tested by giving the system 'id_n' as an input.

2.3 Results

2.3.1 Modelling encoding and recall

Each new event is encoded as an autoassociative trace in the hippocampus, modelled as a modern Hopfield network. Two properties of this network are particularly important: memorisation occurs with only one exposure, and random inputs to the network retrieve stored memories sampled from the whole set of memories (modelling replay).

Recall is modelled as (re)constructing a scene from a partial input. Firstly, I simulate encoding, recall, and replay in the autoassociative network. The network memorises a set of scenes, representing events, as described above. When the network is given a partial input, it recalls the closest stored memory. When the network is given random noise, it retrieves stored memories, corresponding to hippocampal replay (Figure 2.1c). Secondly, I simulate recall in the generative network, following training on reactivated memories from the autoassociative network. As Figure 2.1d shows, the generative network is able to reconstruct the original image when presented with a partial version of an item from the training data.

In the basic model (Figure 2.1a), the prediction error could be calculated for each event, so that only the unpredictable events are stored in the hippocampus, as the predictable ones can already be retrieved by the generative network (however this is not simulated explicitly). In the extended model (Figure 2.2, Section 2.3.5), prediction error is calculated for each element of an event, determining which sensory details are stored.

2.3.2 Modelling semantic memory

Existing semantic memory survives when the hippocampus is lesioned (Manns et al., 2003; Squire et al., 2015; Vargha-Khadem et al., 1997), and hippocampal amnesics can describe remote memories more successfully than recent ones (Scoville & Milner, 1957; Spiers et al., 2001), even if they might not recall them 'episodically' (Nadel & Moscovitch, 1997). This temporal gradient indicates that the semantic component of memories becomes HF-independent. In the model, EC lesions impair all truly episodic recollection, since the return projections from HF are required for generation of sensory experiences, but here I describe how remote memories could be retrieved *in semantic form* despite lesions including hippocampus and EC.

The latent variable representation of an event in the generative network encodes the key facts about the event, and can drive semantic memory directly, without decoding the representation back into a sensory experience (Figure 2.1g). The output route via HF is necessary for turning latent variable representations in mPFC or alTL into a sensory experience, but the latent variables themselves could support semantic retrieval. Thus, when the HF (including EC) is removed, the model can still support retrieval of semantic information (see Section 2.3.7 for details). To show this, I trained models to predict attributes of each image from its latent vector. Figure 2.3a shows that semantic 'decoding accuracy' increases as training progresses, reflecting the learning of semantic structure as a byproduct of learning to reconstruct the sensory input patterns. Whilst semantic memory is much more complex than simple classification, richer 'semantic' outputs such as verbal descriptions can also be decoded from latent variable representations of images (Mokady et al., 2021; Vinyals

et al., 2015).

Notably, there is good performance with only a small amount of training data when decoding the latent variables, compared to decoding alternative representations such as the sensory input or intermediate layer activations, i.e. few-shot learning is possible by making use of compressed 'semantic' representations. See Figure B.2.

2.3.3 Imagination, episodic future thinking, and relational inference

Next let us consider the generation of events that have *not* been experienced from the generative network's latent variables. Events can either be generated by external specification of latent variables (imagination), or by transforming the latent variable representations of other events (relational inference). The former is simulated by sampling from categories in the latent space then decoding the results; Figure 2.3d shows that the generative network can 'imagine' new instances of each shape category following consolidation. The latter is simulated by interpolating between the latent representations of events (Figure 2.3c), or by doing vector arithmetic in the latent space (Figure 2.3b). Figure 2.3c shows that a spectrum of variants between two items can be inferred, while Figure 2.3b shows that the model can infer a variant of an input stimulus by applying a given transformation in the latent space. This demonstrates that the model has learnt some conceptual structure to the data, supporting reasoning tasks of the form 'what is to A as B is to C?', and provides a model for the flexible recombination of memories thought to underlie episodic future thinking (Schacter et al., 2017).

2.3.4 Modelling schema-based distortions

The schema-based distortions observed in human episodic memory increase over time (Bartlett, 1932) and with sleep (Payne et al., 2009), suggesting an association with consolidation. Recall by the generative network distorts memories towards prototypical representations. Figure 2.4a-d shows that MNIST digits (LeCun et

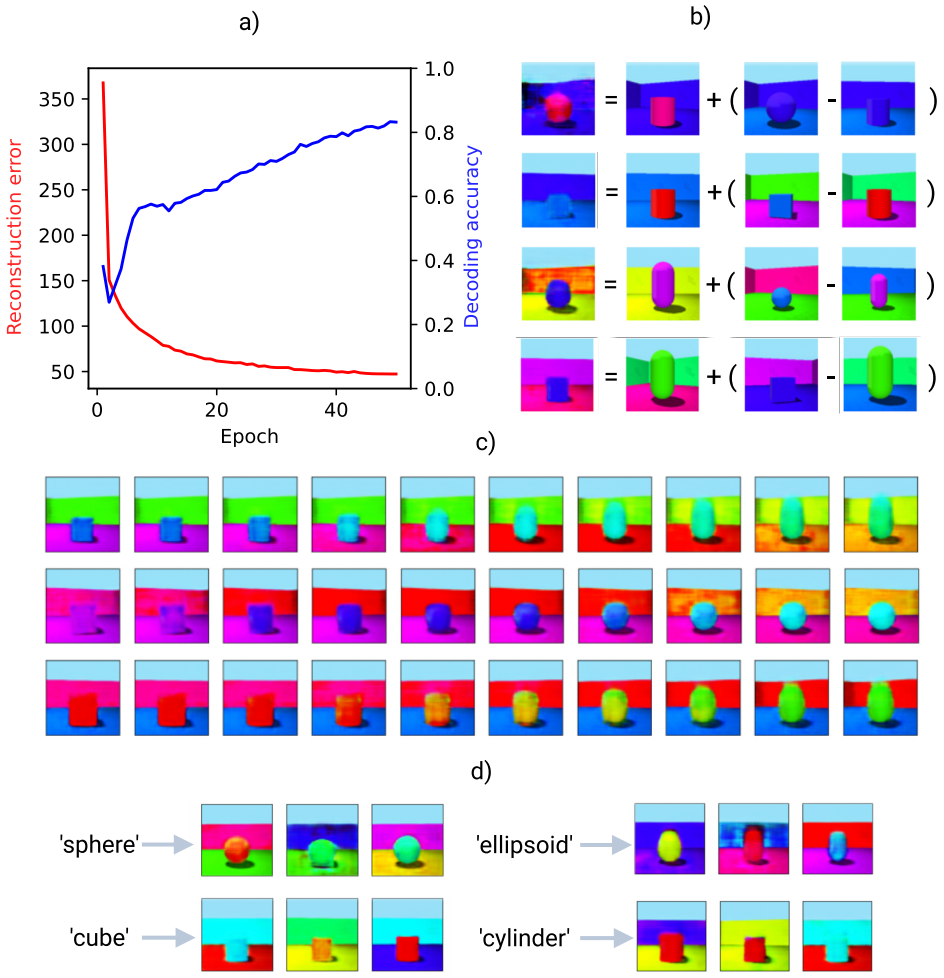

Figure 2.3: Learning, relational inference and imagination in the generative model. a) Reconstruction error (red) and decoding accuracy (blue) improve during training of the generative model. Decoding accuracy refers to performance of a support vector classifier trained to output the central object's shape from the latent variables, using 200 examples at the end of each epoch of generative model training. An epoch is one presentation of the training set of 10,000 samples from the hippocampus. b) Relational inference as vector arithmetic in the latent space. The three items on the right of each equation are items from the training data. Their latent variable representations are combined as vectors according to the equation, giving the latent variable representation from which the first item is generated. The pair in brackets describes a relation which is applied to the second item to produce the first. In the top row, the object shape changes from cylinder to sphere. In the second, the object shape changes from a cylinder to a cube, and the object colour from red to blue. In the third and fourth, more complex transitions change the object colour and shape, wall colour, and angle. c) Imagining new items via interpolation in latent space. Each row shows points along a line in the latent space between two items from the training data, decoded into images by the generative network's decoder. d) Imagining new items from a category. Samples from each of the shape categories of the support vector classifier in part a) are shown.

al., 2010) 'recalled' by a VAE become more prototypical (MNIST is used for this because each image has a single category). Recalled pairs from the same class become more similar, i.e. intra-class variation decreases ($t(7839) = 60.523, p < 0.001, d = -0.684, 95\%CI = [0.021, 0.022]$). The pixel space of MNIST digits before and after recall, and the latent space of their encodings, also show this effect. In summary, recall with a generative network distorts stimuli towards more prototypical representations, even when no class information is given during training. As reliance on the generative model increases, so does the level of distortion.

Boundary extension and contraction exemplify this phenomenon. Boundary extension is the tendency to remember a wider field of view than was observed (Intraub & Richardson, 1989), while boundary contraction is the opposite (Bainbridge & Baker, 2020a). Unusually close-up views appear to cause boundary extension, and unusually far away ones boundary contraction (Bainbridge & Baker, 2020a), although this is debated (Bainbridge & Baker, 2020b; Intraub, 2020). I modelled this by giving the generative network a range of new scenes which were artificially 'zoomed in' or 'zoomed out' compared to those in its training set; its reconstructions are distorted towards the 'typical view' (Figure 2.5a), as in human data. Figure 2.5c shows change to the object size in memory quantitatively, mirroring the findings in Park et al. (2021) (Figure 2.5b). Note that the measure of boundary extension vs. contraction used by Park et al. (2021) is produced by averaging 'closer' vs. 'further' judgements of an identical stimulus image in comparison to the remembered image, rather than the drawing-based measure used here, but the two measures are significantly correlated (Bainbridge & Baker, 2020a).

2.3.5 Combining conceptual and unpredictable sensory features

In the extended model, memories stored in the hippocampal autoassociative network combine conceptual features (derived from the generative network's latent variables) and unpredictable sensory features (those with a high reconstruction error during encoding), see Figure 2.2. In these simulations, the conceptual features are simply

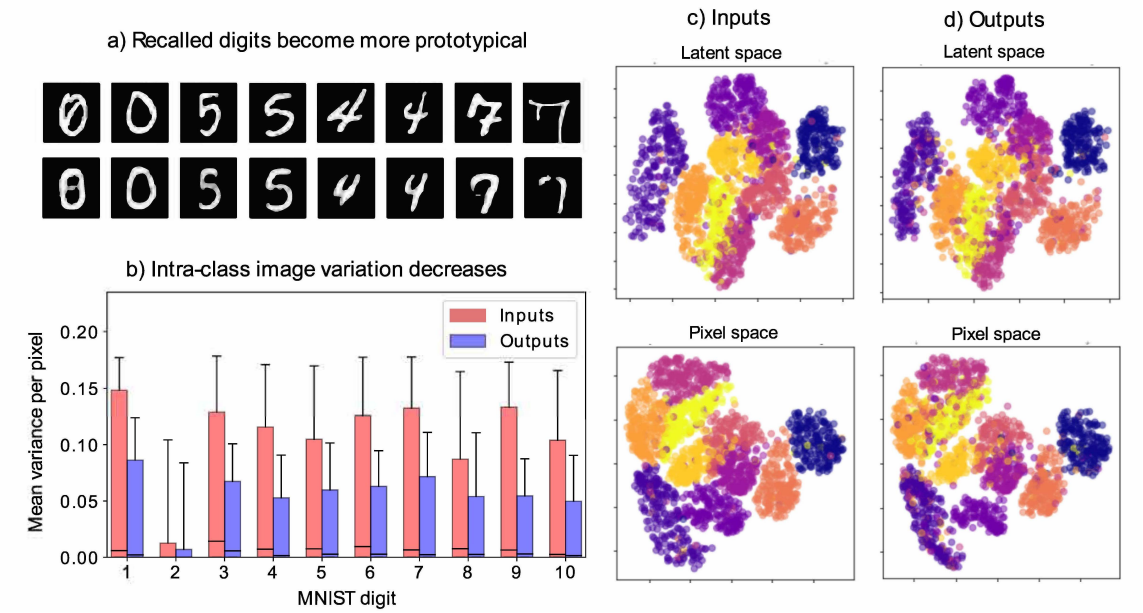

Figure 2.4: Generative network shows schema-based distortions. a) MNIST digits (above) and the VAE's output for each (below). Recalled pairs from the same class become more similar. 10,000 items from the MNIST dataset were encoded in the modern Hopfield network, and 10,000 replayed samples were used to train the VAE. b) The variation within each MNIST class is smaller for the recalled items than for the original inputs. For each of the ten classes, the variance per pixel is calculated across 500 images, and the 784 pixel variances are then plotted for each class, before and after recall. In each box plot, the box gives the interquartile range, its central line gives the median, and its whiskers extend to the 10^{th} and 90^{th} percentiles of the data. c-d) The pixel space of MNIST digits (lower row) and the latent space of their encodings (upper row) show more compact clusters for the generative network's outputs (d) than for its inputs (c). Pixel and latent spaces are shown projected into 2D with UMAP (McInnes et al., 2018) and colour-coded by class.

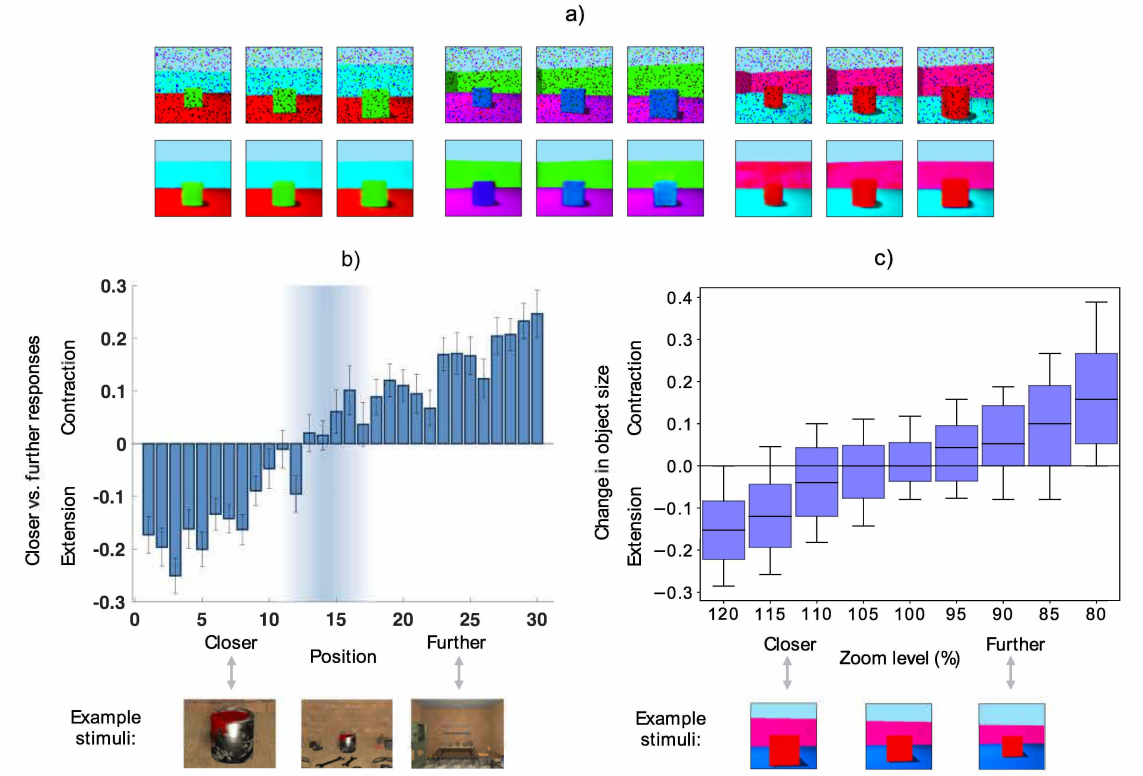

Figure 2.5: Boundary extension and contraction. a) Examples of boundary extension and contraction. The upper row shows the noisy input images (from a held-out test set), with an atypically 'zoomed out' or 'zoomed in' view (by 80% and 120% on the left and right respectively) for three original images. The lower row shows the predicted images for each input image, which are distorted towards the 'typical view' in each case. b) Adapted figure from Park et al. (2021), showing the distribution of boundary extension vs. contraction as a function of the viewpoint of an image (with 900 trials per position). Example stimuli are shown below. c) In the model the VAE increases the estimated size of the central object in atypically 'zoomed out' views compared to the training data, and decreases it in atypically 'zoomed in' views, as in Park et al. (2021). 200 images are used at each 'zoom level'. See Figure 2.4b for a description of box plot elements.

a one-to-one copy of latent variable representations. (Since latent variable representations are not stable as the generative network learns, it seems more likely that concepts *derived* from latent variables are stored than the latent variables themselves, so this is a simplification - see Section 2.2.7 for further details.)

Figure 2.6a-b shows the stages of recall in the extended model, after encoding with a lower or higher prediction error threshold. After decomposing the input into its predictable (conceptual) and unpredictable (sensory) features, the autoassociative network performs pattern completion on the combined representation. The prototypical (i.e. predicted) image corresponding to the retrieved conceptual features must then be obtained by decoding the associated latent variable representation into an experience, via the return projections to sensory neocortex. Next, the predictable and unpredictable elements are recombined, simply by overwriting the prototypical prediction with any unpredictable elements, via the connections from sensory features to sensory neocortex. The extended model is therefore able to exploit the generative network to reconstruct the predictable aspects of the event from its latent variables, storing only those sensory details that were poorly predicted in the autoassociative network. Equally, as the generative network improves, sensory features stored in hippocampus may no longer differ significantly from the initial schematic reconstruction in sensory neocortex, signalling that the hippocampal representation is no longer needed.

2.3.6 Schema-based distortions in the extended model

The schema-based distortions shown in the basic model result from the generative network and increase with dependence on it, but memory distortions can also have a rapid onset (Deese, 1959; Roediger & McDermott, 1995). In the extended model, even immediate recall involves a combination of conceptual and sensory features, and the presence of conceptual features induces distortions prior to consolidation of that specific memory.

In general, recall is biased towards the 'mean' of the class soon after encoding, due

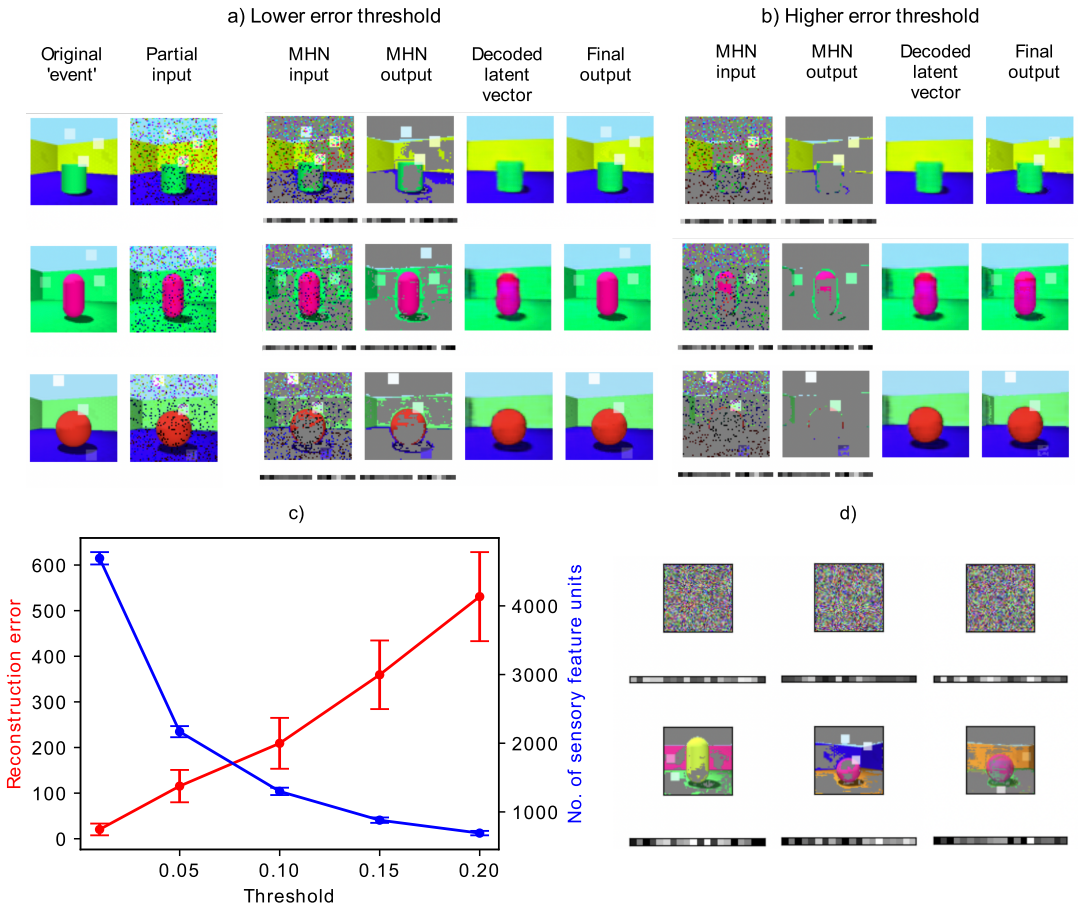

Figure 2.6: Retrieval dependence on reconstruction error threshold, and replay in the extended model. a) The stages of recall are shown from left to right in each row (see Figure 2.2d). Each scene consists of a standard Shapes3D image with the addition of novel features (several white squares overlaid on the image with varying opacity). b) Repeating this process with a higher error threshold for encoding (with the same events and partial inputs) means fewer poorly predicted sensory features are stored in the autoassociative modern Hopfield network (MHN), leading to more prototypical recall with increased reconstruction error. c) Average reconstruction error and number of sensory features (i.e. pixels) stored in the autoassociative MHN against the error threshold for encoding. 100 images are tested and error bars give the standard error of the mean. d) Replay in the extended model. The autoassociative network retrieves memories when random noise is given as input. As above, the square images show the poorly predicted sensory features and the rectangles below these display the latent variable representations.

to the influence of the conceptual representations (Figure 2.6a-b). This is more pronounced when the error threshold for encoding is high, as there is more reliance on the 'prototypical' representations, resulting in the recall of fewer novel features. At a lower error threshold, more sensory detail is encoded, i.e. the dimension of the memory trace is higher. This results in a lower reconstruction error, indicating lower distortion, but at the expense of efficiency.

External context further distorts memory. Carmichael et al. (1932) asked participants to reproduce ambiguous sketches. A context was established by telling the participants that they would see images from a certain category. After a delay, drawings from memory were distorted to look more like members of the context category. Figure 2.7b shows the result of encoding the same ambiguous image with two different externally provided concepts (a cube in the upper row, a sphere in the lower row), represented by the latent variables for each concept, as opposed to the latent variables predicted by the image itself as in Figures 2.6a-b. During recall, the encoded concept is retrieved in the autoassociative network, determining the prototypical scene reconstructed by the generative network. This biases recall towards the class provided as context, mirroring Figure 2.7a.

I also simulate the Deese-Roediger-McDermott (DRM) task (Deese, 1959; Roediger & McDermott, 1995) in the extended model to demonstrate its applicability to non-image stimuli. In the DRM task participants are shown lists of words that are semantically related to 'lure words' not present in the list; there is a robust finding that false recognition and recall of the lure words occur (Deese, 1959; Roediger & McDermott, 1995). In the extended model, gist-based semantic intrusions arise as a consequence of learning the co-occurrence statistics of words. First the VAE is trained to reconstruct the sets of words in simple stories (Mostafazadeh et al., 2016) converted to vectors of word counts, representing background knowledge. The system then encodes the experimental lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent representation of each word list (respectively analogous to the stimulus-unique pixels and the VAE's latent representation of each image in Figure 2.6). As in the human data, lure

words are often but not always recalled when the system is presented with 'id_n' (Figure 2.8a), since the latent variable representations which generate the words in the list also tend to generate the lure word. The system also forgets some words, and produces additional semantic intrusions, e.g. 'vet' in the case of the 'doctor' list. In addition the chance of recalling the lure word is higher for longer lists ($r_s(10) = 0.998, p < 0.001, 95\%CI = [0.982, 1.000]$), as in human data from Robinson and Roediger (1997), as more related words provide a stronger 'prior' for the lure (Figure 2.8b).

2.3.7 Modelling brain damage

Recent episodic memory is impaired following damage to the hippocampal formation (HF), whereas semantic memory – including the semantic content of remote episodes - appears relatively spared. In the model the semantic form of a consolidated memory survives damage to HF thanks to latent variable representations in mPFC or alTL (even if those in EC are lesioned); Figure 2.3a demonstrates how semantic recall performance improves with the age of a memory over the course of consolidation, reflecting the temporal gradient of retrograde amnesia (see Section 2.3.2). However these semantic 'facts' cannot be used to generate an experience *episodically* without the generative network's decoder, in agreement with multiple trace theory (Nadel & Moscovitch, 1997).

The extent of retrograde amnesia can vary greatly depending, in part, on which regions of the HF are damaged (Cipolotti et al., 2001; Zola-Morgan et al., 1986). The dissociation of retrograde and anterograde amnesia in some cases suggests that the circuits for encoding memories and the circuits for recalling them via the HF only overlap partially (Zola-Morgan et al., 1986). For example, if the autoassociative network is damaged but not the generative network's decoder, the generative network can still perform reconstruction of fully consolidated memories. This could explain varying reports of the gradient of retrograde amnesia when assessing episodic recollection (as opposed to semantic memory), if the generative network's decoder is intact in patients showing spared episodic recollection of early memories

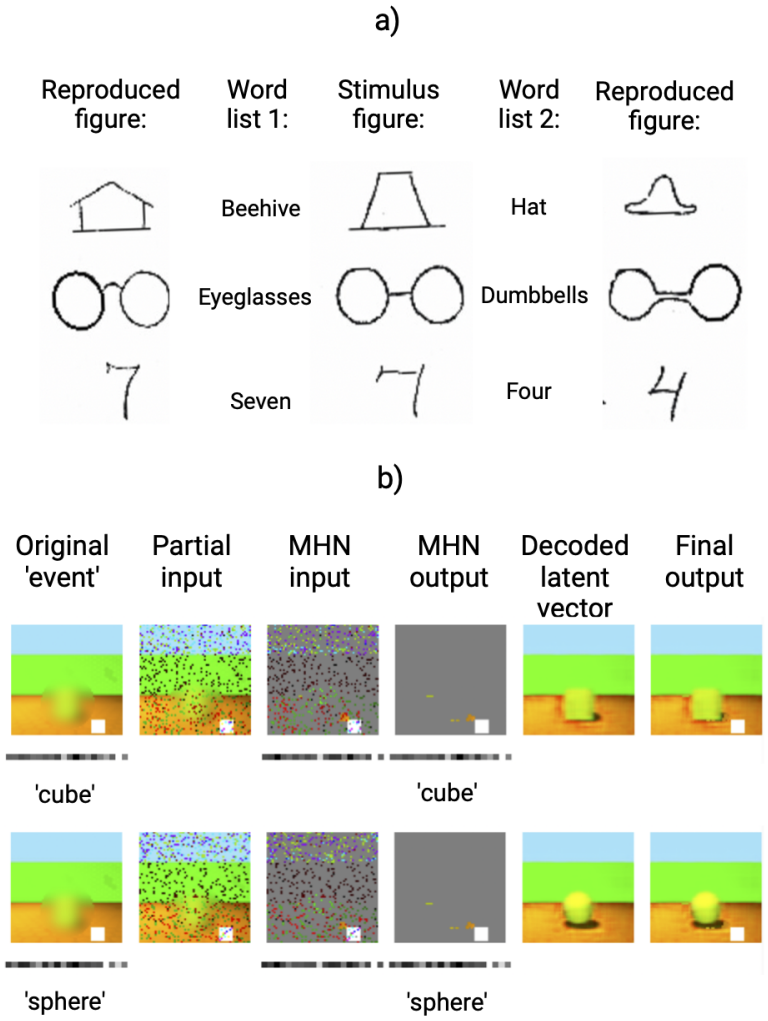

Figure 2.7: Schema-based distortions: effects of conceptual context in the extended model. a) Adapted figure from Carmichael et al. (1932) showing that recall of an ambiguous item (Stimulus figure, centre) depends on its context at encoding (Word from list 1, left; or list 2, right), as shown by drawing from memory (Reproduced figure, far left and far right). b) Memory distortions in the extended model, when the original scene (containing an ambiguous blurred shape) is encoded with a given concept (cube, above; sphere, below), represented by the latent variables for that class. Then a partial input is processed by the generative network to produce predicted conceptual features and the sensory features not predicted by the prototype for that concept (in this case a white square), for input to the autoassociative modern Hopfield network (MHN). However, pattern completion in the MHN reproduces the originally encoded sensory and conceptual features (cube, above; sphere, below), and these are recombined to produce the final output, which is distorted towards the encoded conceptual context.

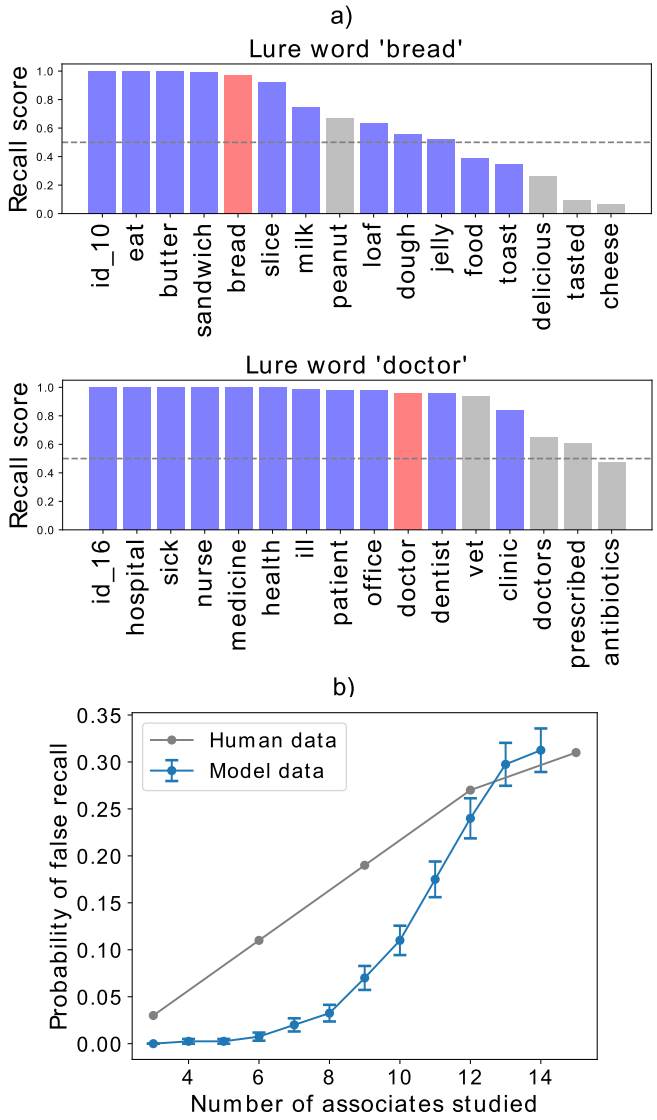

Figure 2.8: Modelling the Deese-Roediger-McDermott task. a) First the VAE is trained to reconstruct simple stories (Mostafazadeh et al., 2016) converted to vectors of word counts, representing background knowledge. The system then encodes the lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent variable representation of the word list. In each plot, recalled stimuli when the system is presented with 'id_n' are shown, with output scores treated as probabilities so that words with a score of above 0.5 are recalled. Words from the stimulus list are shown in blue, and lures in red. See Figure B.1 for results for the remaining DRM lists. b) The chance of recalling the lure word is higher when longer lists are encoded (blue). Each measurement is averaged across 400 trials (20 random subsets of each of the 20 DRM lists), and error bars give the standard error of the mean. This qualitatively resembles human data from Robinson and Roediger (1997) (grey).

(Squire et al., 2015). Note that the location of damage within the generative network's decoder also affects the resulting deficit in the model. In particular, patients with damage restricted to the hippocampus proper can (re)construct simple scenes but not more complex ones (Hassabis et al., 2007).

The model also shows the characteristic anterograde amnesia after hippocampal damage, as the hippocampus is required to initially bind features together and support off-line training of the generative model. Anterograde semantic learning would also be impaired by hippocampal damage (as the generative network is trained by hippocampal replay). Whilst hippocampal replay need not be the only mechanism for schema acquisition, it would likely be much slower without the benefit of replay. However, semantic learning over short timescales may be relatively unimpaired, as it is less dependent on extracting regularities from long-term memory (Knowlton et al., 1994).

In semantic dementia, semantic memory is impaired, and remote episodic memory is impaired more than recent episodic memory (Hodges & Graham, 2001). This would be consistent with lesions to the generative network, as recent memories can rely more on the hippocampal autoassociative network. However the exact effects would depend on the distribution of damage across the various potential generative networks in EC, mPFC and alTL. Of these, the alTL network is associated with semantic dementia, and the posterior medial network (corresponding to the generative network between sensory areas and EC) with Alzheimer's disease (Ranganath & Ritchey, 2012).

Finally, neuropsychological evidence suggests a distinction between familiarity and recollection, and furthermore a partial dissociation between different tests of familiarity; patients with selective hippocampal damage can exhibit recognition memory deficits in a simple 'yes/no' task with similar foils, but not in a 'forced choice' variant involving choosing the more familiar stimulus from a set (Migo et al., 2009). This is consistent with the idea that lower prediction error in the neocortical generative network indicates familiarity, but retrieval of unique details from the hippocampus is required for more definitive recognition memory.

2.4 Discussion

This chapter proposes a model of systems consolidation as the training of a generative neural network, which learns to support episodic memory, and also imagination, semantic memory and inference. This occurs through teacher-student learning. The hippocampal 'teacher' rapidly encodes an event, which may combine unpredictable sensory elements (with connections to and from sensory cortex) and predictable conceptual elements (with connections to and from latent variable representations in the generative network). After exposure to replayed representations from the 'teacher', the generative 'student' network supports reconstruction of events (in conjunction with stored hippocampal details until memories are fully consolidated).

In contrast to the relatively veridical initial encoding, the generative model learns to capture the probability distributions underlying experiences, or 'schemas'. This enables not just efficient recall, reconstructing memories without the need to store them individually, but also imagination (by sampling from the latent variable distributions) and inference (by using the learned statistics of experience to predict the values of unseen variables). In addition, semantic memory, i.e. factual knowledge, develops as a byproduct of learning to predict sensory experience. As the generative model becomes more accurate, the need to store and retrieve unpredicted details in hippocampus reduces (producing a gradient of retrograde amnesia in cases of hippocampal damage). However, the generative network necessarily introduces distortion compared to the initial memory system. Multiple generative networks can be trained in parallel, and this may include networks with latent variables in EC, mPFC, and alTL.

We can now compare the model's performance to the list of key findings from the introduction:

1. *Gradual consolidation follows one-shot encoding:* A memory is encoded in the hippocampal 'teacher' network after a single exposure, and transferred to the generative 'student' network after being replayed repeatedly (see Figure 2.1c-d).

2. *Semantic memory becomes hippocampus-independent*: The latent variable rep-

resentations learned by the generative networks constitute the 'key facts' of an episode, supporting semantic memory (see Figure 2.3a).

3. *Episodic memory remains hippocampus-dependent:* Return projections via HF to sensory neocortex are required to decode the latent variable representations into a sensory experience (see Figure 2.1). (EC is required for even simple (re)construction, while the hippocampus proper helps to generate complex, conceptually coherent scenes, and retrieves unpredictable details which are not yet consolidated into the generative network - see Section 2.1.3.)

4. *Shared substrate for episode generation:* Generative models are a common mechanism for episode generation. Familiar scenes can be reconstructed and new ones can be generated by sampling or transforming existing latent variable representations (Figure 2.3b-d), providing a model for imagination, scene construction and episodic future thinking.

5. *Consolidation promotes inference and generalisation*: Relational inference corresponds to vector arithmetic applied to the generative network's latent variables (Figure 2.3b).

6. *Episodic memories are distorted*: I show how memory distortions arise from the generative network (Figures 2.5, 2.4, 2.7, and 2.8). This extends the model of Nagy et al. (2020) to relate memory distortion to consolidation.

7. *Association cortex encodes latent structure:* Latent variable representations in EC, mPFC, and alTL provide schemas for episodic recollection and imagination (via HF) and for semantic retrieval and inference.

8. *Prediction error affects memory processing:* The generative network is constantly calculating the reconstruction error of experiences (Chen et al., 2011; Kumaran & Maguire, 2006). Events that are consistent with the existing generative model require less encoding in the autoassociative hippocampal network (see Figure 2.6).

9. *Episodic memories include conceptual features:* When an experience combines a

mixture of familiar and unfamiliar elements, both concepts and poorly-predicted sensory elements are stored in hippocampus via association to a specific memory unit.

This model can be seen as an update to the complementary learning systems (CLS; McClelland et al., 1995) framework to better account for points 3-9 above, reconciling the development of semantic representations in neocortex (as per CLS) with the continued dependence on the hippocampal formation for episodic recall (as per multiple trace theory; Nadel & Moscovitch, 1997). Furthermore, it provides a unified view of episode generation, of how episodic memories change over time and exhibit distortions, and of how semantic and episodic information are combined in memory. I build on previous work exploring the role of generative networks in consolidation (Káli & Dayan, 2000, 2002), as models of the hippocampal formation (Nagy et al., 2020; Van de Ven et al., 2020; Whittington et al., 2020), as priors for episodic memory (Fayyaz et al., 2022), and as models of spatial cognition (Bicanski & Burgess, 2018).

A key aspect of the model is that multiple generative networks can be trained concurrently from a single autoassociative network (Figure 2.2a), and may be optimised for different tasks. Thus, the latent representations in mPFC and alTL may be more closely linked to value or language than those in EC (Lin et al., 2016; Moscovitch & Melo, 1997). These differences may arise from differences in network structure (e.g. the degree of compression), or from additional training objectives that shape their representations (Gluck & Myers, 1993). (For instance, the generative network with latent variables in mPFC might be trained to predict task-relevant value in addition to the EC representations.) The generative networks might be expected to overlap more closer to their sensory inputs/outputs, where general-purpose features are more useful, and diverge as the representations become more abstract, or task-specific if there are additional training objectives (Yosinski et al., 2015). This may involve a primary VAE with latent variables in EC, with additional pathways from higher sensory cortex to EC routed via latent variables in mPFC or alTL.

This model raises some fundamental questions: Does true episodic memory require

event-unique detail, and does this require the hippocampus? Or can prototypical predictions qualify as memory rather than imagination? In the model, event-unique details are initially provided by the hippocampus, but can also be provided by the generative network. For example, if you know that someone attended your 8th birthday party and gave you a particular gift, these personal semantic facts need not be hippocampal-dependent, but could generate a scene with the right event-specific details, which would seem like episodic memory. The increasingly sophisticated generation of images from text using generative models (Ramesh et al., 2022) suggests that episode construction from semantic facts is computationally plausible.

Episodic memories are defined by their unique spatiotemporal context (Tulving, 1985). In the model, spatial and temporal context correspond to conceptual features captured by place (Ekstrom et al., 2005; O'Keefe & Dostrovsky, 1971) or time (Eichenbaum, 2014; Umbach et al., 2020) cells in hippocampus and might be linked to latent variable representations formed in EC, such as grid cells in medial EC, which form an efficient basis for locations in real (Dordek et al., 2016; Stachenfeld et al., 2017; Whittington et al., 2020) or cognitive spaces (Constantinescu et al., 2016; Whittington et al., 2020), or temporal context representations in lateral EC (Bright et al., 2020; Tsao et al., 2018). Events with specific spatial and temporal context can be generated from these latent variable representations, as has been modelled in detail for space (Becker & Burgess, 2000; Bicanski & Burgess, 2018; Byrne et al., 2007).

More generally, this work builds on the spatial cognition literature, in which place and head direction cells act as latent variables in a generative model (Becker & Burgess, 2000; Bicanski & Burgess, 2018; Byrne et al., 2007), allowing the generation of a scene from a specific viewpoint. Becker and Burgess (2000), Bicanski and Burgess (2018) and Byrne et al. (2007) explore how egocentric sensory representations could be transformed into allocentric latent variables prior to storage in the medial temporal lobe, and conversely how egocentric representations could be reconstructed from allocentric ones to support imagery. The latent representations learned through consolidation in the model correspond loosely to the allocentric representations, and the sensory representations produced by HF to the egocentric ones; only egocentric and

sensory representations are directly experienced, whereas allocentric and semantic representations are useful abstractions which can also be exploited for efficient hippocampal encoding.

The model simplifies the true nature of mnemonic processing in several ways. Firstly, episodic memories contain important sequential structure, not modelled by the encoding and reconstruction of simple scenes. Chapter Three expands the model's scope from 'snapshots' to 'sequences' of experience. Secondly, considering consolidation as a continual lifelong process, rather than during encoding of a single dataset, introduces new complexities, in particular the prevention of catastrophic forgetting of already consolidated memories as new memories are assimilated into the generative network. Chapter Four extends the model to address this. Thirdly, the interaction of sensory and conceptual features in hippocampus and latent variables in EC during retrieval could be more complex, with each type of representation contributing to pattern completion of the other as per interactions between items and contextual representations in the Temporal Context Model (Howard & Kahana, 2002), and might iterate over retrievals from both hippocampal and generative networks (Kumaran et al., 2016). Fourthly, the model distinguishes between 'sensory' and 'conceptual' representations in hippocampus, respectively linked to the sensory neocortex at the input/output of the generative network and to the latent variable layer in the middle. In reality a gradient of levels of representation in hippocampus is more likely, from detailed sensory representations to coarse-grained conceptual ones, respectively linked to lower or higher neocortical areas (Moscovitch et al., 2016), and might map onto the observed functional gradients along the longitudinal axis of the hippocampus (Strange et al., 2014). Fifthly, the generative network uses back-propagation of the prediction error between output and input patterns to learn. Generative networks with more plausible (if less efficient) learning rules exist (Dayan et al., 1995; Friston, 2010; Rao & Ballard, 1999), which have the advantage of producing a prediction error signal at each layer (between top-down prediction and bottom-up recognition), potentially allowing learning of concepts and exceptions at all levels of description. Finally, I model semantic memory as prediction of categorical information for an 'event', but

future work should model more complex semantic knowledge, e.g. by decoding language from latent representations of multimodal stimuli (Mokady et al., 2021; Vinyals et al., 2015). In particular the relationship between semantic memory for specific 'events' and the broader 'web' of general knowledge should be considered.

This model makes testable predictions. Firstly, if participants learn stimuli generated from known latent variables, it predicts that these specific latent variable representations should develop in association cortex over time (and that this representation would support, e.g., vector arithmetic and interpolation). This could be tested by representational similarity analysis, which should reveal a more conceptual similarity structure developing in association cortex through consolidation, as opposed to a similarity structure reflecting the sensory stimuli in earlier sensory cortices. If the stimuli also contained slight variation, i.e. they were not entirely described by the latent variables, the development of a latent variable representation should be correlated with gist-based distortions in memory, and anti-correlated with hippocampal processing of unpredictable elements.

Secondly, the model makes multiple predictions about the effects of brain damage. Just as boundary extension is reduced in patients with damage to HF (Mullally et al., 2012) or vmPFC (De Luca et al., 2018), the model predicts that other biases towards the 'canonical view' would be attenuated in such patients; for example, healthy controls would distort images with an atypical viewing angle towards a more typical angle in memory, but this would be reduced in, e.g., hippocampal patients. Similarly, ambiguous images such as the duck/rabbit drawing 'flip' between interpretations in perception, but are stable when held in imagery (Chambers & Reisberg, 1985), presumably due to maintained hippocampal conceptual representations. The model predicts that this conceptual stability in imagery would also be reduced in such patients. This could also extend to non-scene stimuli: if the Carmichael et al. (1932) task were tested with both healthy controls and patients with damage to the generative decoder, one would expect reduced contextual distortion in the latter. Furthermore, patients with an inaccurate generative model, e.g. due to semantic dementia, might rely more on sensory features to compensate. (Note that the pattern of deficits would

depend on both the nature of the priors encoded in the generative network *and* the error threshold for encoding. In some cases damage to the generative network could produce atypical 'priors' rather than suppressing them. Thus, if the generative network is inaccurate but the error threshold for encoding is high, atypical distortions will be observed, rather than a reduction in conceptual distortions.)

Thirdly, the model suggests that the error threshold for encoding could vary depending on the importance of the stimuli, or the amount of attentional resource available. For example, emotional salience could lower this threshold, with traumatic memories being encoded in greater sensory detail and with less contextual coherence (Bisby et al., 2020; Van Der Kolk et al., 1997). Equally, conditions such as autism spectrum disorder, which are potentially attributable to hypo-priors (Pellicano & Burr, 2012), might be associated with a lower prediction error threshold for veridical storage (and thus reduced conceptual influence on memory, and increased sensory detail). In addition, reality monitoring deficits would change the perceived prediction error relative to reality, leading to atypical memory storage (e.g. a reduced ability to compensate for prediction errors by storing sensory details).

Fourthly, biological intelligence excels at generalising from only a small number of examples. The model predicts that learning to generalise rapidly benefits from having a generative model that can create new examples, e.g. by inferring variants as in Figure 2.3b (see also Barry & Love, 2021). Finally, the model suggests a link between latent spaces and cognitive maps (Behrens et al., 2018). For example, one might predict that the position of a memory in latent space is reflected in place and grid cell firing, as observed for other conceptual representations (Behrens et al., 2018; Constantinescu et al., 2016; Nieh et al., 2021).

In summary, the proposed model takes inspiration from recent advances in machine learning to capture many of the intriguing phenomena associated with episodic memory, its (re)constructive nature, its relationship to schemas, and consolidation, as well as aspects of imagination, inference and semantic memory.

Chapter 3

Learning to construct sequential events

3.1 Introduction

Memories are not instantaneous snapshots, but are sequential in nature. Here I extend the account of consolidation as teacher-student training of generative networks to sequential stimuli. As in the 'static model' discussed so far, the hippocampal network is the 'teacher' training neocortical generative 'students' through replay. But here the generative networks are trained not only to reconstruct their own inputs, but to predict the next input in a sequence (Radford et al., 2019).

The 'sequential model' allows us to explore a broader range of phenomena than the static model. Whilst simple inference (inferring novel combinations of learned concepts) was demonstrated with the static model, more sophisticated structural inference involves sequences of relationships. In addition, planning involves sequences of states, actions, and rewards, and navigation involves spatiotemporal sequences. The sequential model therefore allows us to investigate how generative networks trained through consolidation might support structural inference, planning, and navigation.

Furthermore, in humans language and memory are inextricably linked, with many studies of memory involving narratives (Bartlett, 1932; Raykov et al., 2023; Zwaan & Radvansky, 1998). Until recently, the tools to simulate memory for narratives have not been mature enough to explore these experimental data computationally. However the arrival of transformer-based neural networks like GPT-2 (Radford et al., 2019), which is used to represent the neocortical generative network, enables the sequential model to be applied to text.

By representing a range of stimuli as sequences, I show that the sequential model is capable of classic statistical learning (Durrant et al., 2011) and spatial / relational inference (Whittington et al., 2020) tasks, displays similar gist-based memory distortions to those observed in human data (Bartlett, 1932; Raykov et al., 2023), and learns to support model-based planning over time (as in Vikbladh et al., 2024). Finally, I sketch out how the hippocampal 'memory bank' and neocortical generative network could work together to enable problem-solving from memory, inspired by the 'retrieval augmented generation' approach for combining large language models with non-parametric memory (Lewis et al., 2020).

3.1.1 Generative models for sequences

I begin by describing the generative network component of the sequential model. GPT-2 (Radford et al., 2019) is a deep neural network which can be trained on arbitrary linguistic or non-linguistic sequences; the objective is simply to predict the next item in sequences drawn from the training data. As with the VAEs in Chapter Two, training therefore involves learning to reconstruct inputs in a self-supervised manner.

GPT-2 is built on the transformer architecture, which uses the attention mechanism to weight the importance of items in a sequence relative to a given item, allowing the model to capture complex interdependencies (Vaswani et al., 2017). In short, the attention mechanism works as follows: a query, a key, and a value vector are produced for each element in the sequence by learned weights. For a given element, the model

computes a set of attention scores which determine how much attention should be paid to each element of the input sequence when representing this element. This is achieved by taking the dot products of the given element's query vector with the key vectors for each element. The final output is the sum of the value vectors for each element weighted by the corresponding attention scores (with some normalisation), allowing the model to aggregate information from the most relevant parts of the input. This describes a single 'attention head', but GPT-2's architecture features multiple attention blocks, each with multiple attention heads, which tend to develop different 'specialisms' through training. See Section C.2 of the Appendix for further mathematical details on attention.

The primary goal during training is to adjust the model's parameters through maximum likelihood estimation, so that the probability it predicts for the true next item in each sequence, based on the items so far, is as high as possible. In other words, the network's weights are updated to predict the probability distribution of the next item as accurately as possible. The training data for the original GPT-2 model is WebText, a dataset of online content scraped from outbound Reddit links and further processed to ensure quality. See Section 3.2.1 for more detail on the training procedure.

Once the model is trained, it can continue from an input sequence, or generate a new sequence from scratch, by iteratively predicting the next item from the items so far. (The input sequence on which the output is conditioned is known as a 'prompt'.) That is, it predicts the probability distribution across all items given the items so far, and one can either sample from this distribution or simply take the most probable item at each step. The equation below gives the probability of a sequence x as a product of conditional probabilities of its items:

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

When sampling from the learned probability distribution, a key determinant of the model's behaviour is the temperature. The temperature parameter modifies

the probability distribution for the next token: a lower temperature makes the model behave more like greedy decoding, favouring high probability items, while a higher temperature encourages more diverse outputs by flattening the probability distribution. This gives lower probability items a better chance of being selected, producing more 'imaginative' outputs, while for models trained on language a high enough temperature produces nonsensical text. (See Section 3.2.2 for discussion of more complex approaches to generating sequences from a trained model.)

To be more precise, the equation below describes the 'softmax with temperature' function that is applied to the vector of scores for each token. The softmax function transforms this vector of scores into a vector of *probabilities*. As the numerator is an exponential, a large temperature T flattens the distribution, whereas T close to zero approximates a 'one-hot' vector, with a probability of one for the most likely token. The denominator normalises each element in the vector by the sum of all the exponentials, ensuring the probabilities add up to one:

$$\sigma(s_i) = \frac{e^{\frac{s_i}{T}}}{\sum_{j=1}^{n} e^{\frac{s_j}{T}}}$$

An obvious question is whether models like GPT-2 can really capture specifics as well as generalities. A great deal of research (predominantly looking at this question from the perspective of information security) has demonstrated the tendency of language models to memorise their training data. Carlini et al. (2022) shown that the more parameters a model has, the more likely it is to memorise sequences it was trained on. The likelihood of memorisation also increases with the number of occurrences of the sequence in the training dataset, and the amount of training. In addition, memorised data can be retrieved more easily when the model is 'prompted' with (i.e. conditioned on) a longer cue. The propensity of these models to memorise their training data hints that they may be able to capture episodic specifics as well as semantic generalities.

3.1.2 Modelling sequence memory in the hippocampus

The mechanisms by which sequences are stored as traces in the initial hippocampal network are not the focus of this chapter, and the hippocampus is not simulated explicitly in the subsequent results. However, I now outline how the hippocampus might store sequences autoassociatively (or heteroassociatively) in the sequential model, in a way that supports replay and recall prior to consolidation.

In the static model, a modern Hopfield network (MHN; Ramsauer et al., 2020) represents the hippocampus, interpreted such that the feature units activated by an event are bound together by a memory unit (Krotov & Hopfield, 2020). In the sequential model, this could be adapted to store sequences as follows. Let us consider sequences represented as strings of characters or symbols, which can capture language, spatial trajectories, transitions in a graph, stimuli in sequential learning tasks, and more.

As discussed in Chapter One, Millidge et al. (2022) give a unifying account of how neural network models of associative memory such as MHNs operate, which helps to explain the extension to sequences. They develop the biologically plausible modern Hopfield network (Krotov & Hopfield, 2016) into 'a general framework for understanding the operation of ... memory networks as a sequence of three operations: similarity, separation, and projection', which they term 'universal Hopfield networks' (Millidge et al., 2022, Abstract). They observe that any of these models can be made asymmetric or heteroassociative rather than autoassociative, if the projections from feature units to memory units capture the current state, but the projections from memory units back to feature units are different (for sequences, these projection weights would correspond to the 'next state'). Then one state retrieves the next, rather than each state retrieving a pattern-completed version of itself.

The modern continuous asymmetric Hopfield network (MCAHN) converts the MHN to work for sequences, consistent with the framework above (Chaudhry et al., 2023). However there is a problem when you try to apply this naively to certain sequences, as the MCAHN is by default a 'Markov chain' model of sequential memory. Whilst it

works for sequences in which items are unique, it doesn't work for, e.g., sentences without modification, as it cannot accurately remember sequences with repeated items; see Figure 5a of Tang et al. (2023). When an item is repeated in a sequence, Tang et al. (2023) show that a MCAHN has trouble recalling the next item, instead producing a composite.

Tang et al. (2023) introduce the temporal predictive coding network (tPC), a more complex network learnt by gradient descent, based on the intuition that higher layers minimize prediction error at lower layers (Salvatori et al., 2021). (A drawback of the tPC is that the model must be 'presented with the sequence for multiple epochs' until convergence, so it is no longer capable of one-shot learning, which arguably undermines its plausibility as a model of hippocampal encoding.) They show that the one-layer tPC also has the MCAHN's issue with repeated items, but suggest the two-layer tPC to resolve this. As Tang et al. (2023) show in Figure 5d of their paper, the hidden layer represents an item's position in the sequence so the next item can be recalled correctly.

Alternatively, one could combine the concept of the modern asymmetric Hopfield network (with return projections from memory to feature units representing the next state) with a state representation that captures the history. This is based on previous work involving the role of temporal context in memory (Burgess & Hitch, 1999; Howard & Kahana, 2002). Consider the example of encoding sentences, i.e. sequences of characters. Suppose each state is a vector of length equal to the number of symbols, consisting of one at the index of the current symbol, plus the previous state multiplied by some decay factor. Then if we want to remember the string 'abc', the letter 'a' is represented as (1, 0, 0 ...), 'b' as (0.5, 1, 0 ...), 'c' as (0.25, 0.5, 1 ...), and so on (with a decay factor of 0.5, and ignoring normalisation). See Figure 3.1 for further details. This has the benefit of still being compatible with one-shot learning, although the memory capacity may not scale as well as the more complex predictive coding model approach. Initial testing suggests that this model can encode and retrieve sentences (although retrieval performance goes down with sentence length). See Section C.3 of the Appendix for further details.

This simple model clearly has some limitations. For example, it suggests that when a state from a certain point in an encoded sequence triggers recall, states preceding that point cannot be recalled (only the subsequent states). This is clearly not the case, so a model of sequence memory that can only retrieve states in one temporal direction is unsatisfactory. Also, if states involve vectors of continuous rather than binary variables, the decaying activity from previous states could be hard to distinguish from the current state, so a more complex solution could be required.

In addition, the sequences consist of a single symbol at any moment in time, but other sequences like frames in a video consist of complex representations at each time step. The Hopfield network variant could perhaps have a combination of autoassociative and heteroassociative connectivity to account for this, with autoassociative connections reconstructing the current time step, and heteroassociative connections activating the next (see Sompolinsky & Kanter, 1986).

3.1.3 Event perception and segmentation

Whilst time is continuous, humans experience it as discretised into events (Newtson, 1973; Newtson & Engquist, 1976), a process known as event segmentation. Event segmentation occurs at multiple levels of granularity, with coarser-grained segments made up of finer-grained ones in a hierarchical structure (Zacks, Tversky et al., 2001). Distinctive neural activity is observed at event boundaries, which is predictive of subsequent memory performance (Baldassano et al., 2017; Ben-Yakov & Dudai, 2011). In an fMRI study involving videos of everyday activities, Zacks, Braver et al. (2001) observed cortical activity at event boundaries even in passive viewing, suggesting that event segmentation is an automatic aspect of perception. Activity began to 'ramp up' some time before the boundary, and was stronger for more coarse-grained boundaries.

Zacks et al. (2007) propose a theory of how event segmentation occurs, in which event models in working memory represent the current situation, based on both bottom-up inputs from sensory data and top-down inputs from schemas. These event models

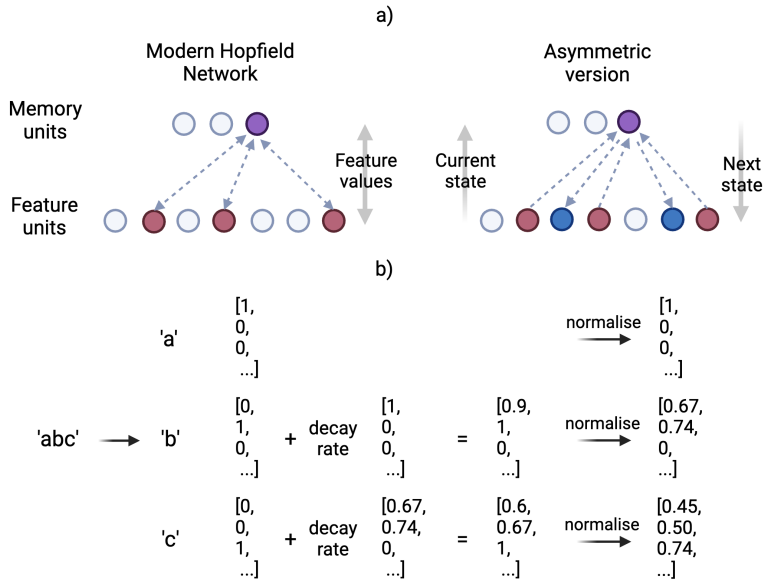

Figure 3.1: Using asymmetric modern Hopfield networks to model the hippocampus. a) In asymmetric versions of modern Hopfield networks, the weights between feature and memory units differ, as visualised from the perspective of Krotov and Hopfield (2020). b) A sequence of arbitrary symbols, e.g. 'abc', is represented as a sequence of vectors over the feature units, where each vector is the sum of the current state plus the decay rate (here set to 0.9) times the previous state. For each state, the return connections from memory units to feature units encode the next state.

'bias processing in the perceptual stream' and are 'robust to transient variability in the sensory input'. According to this theory, 'an error detection mechanism ... compares the perceptual processing stream's predictions to what actually happens in the world' (Architecture and Principles section). The event model is updated when the prediction error is above some threshold. Importantly, high prediction error leads to the flow of sensory inputs into the event model, while low prediction error suppresses sensory inputs (i.e. an 'error-based gating' mechanism). Zacks et al. (2007) suggest that event segmentation occurs 'simultaneously on a range of timescales, spanning from a few seconds to tens of minutes' (Multiple Timescales section). The Structured Event Memory model (Franklin et al., 2020) agrees on the importance of prediction error for segmenting sequences of scenes into events.

Event segmentation affects memory, with the order of items within an event (or context) remembered better than items spanning multiple events (DuBrow & Davachi, 2013). In addition, it has been shown that passing through a doorway can cause forgetting (the 'doorway effect'), even in virtual environments (Horner et al., 2016; Radvansky & Copeland, 2006). One explanation is that crossing spatial boundaries triggers a shift in the event / situation model, causing people to forget information associated with a previous spatial context (Radvansky & Copeland, 2006). In other words, this phenomenon suggests that spatial boundaries can serve as event boundaries.

Whilst the sequences used in the subsequent simulations are 'pre-segmented' for simplicity, the role of prediction error in event segmentation theory is potentially consistent with the sequential model. The generative network can provide an ongoing measure of 'surprise' during perception (Franklin et al., 2020), e.g. as quantified by perplexity (Radford et al., 2019).

3.1.4 Planning and memory

I also explore how the generative network trained on memories can simulate future events to support behaviour, such as when planning how to achieve a goal.

Planning involves coming up with a goal-oriented sequence of actions, e.g. planning a route to a certain location. As achieving a goal can be described in terms of maximising some reward, the study of planning overlaps with the study of reinforcement learning, a subfield of machine learning concerned with how agents take actions in an environment to maximise cumulative reward. Planning requires modelling of states, actions, and rewards as sequences, so the sequential model allows us to consider how the generative network might support planning.

Two concepts from reinforcement learning (RL) commonly applied to the study of planning are 'model-free' and 'model-based' learning (Daw et al., 2011). As the name suggests, model-free RL involves learning how to act without constructing a model of the task or environment. Rather than learning the 'transition structure' of the world, i.e. the probabilities of states given preceding states / actions, it learns associations between actions and rewards more directly. Q-learning (Watkins & Dayan, 1992) is one example of model-free RL; after learning a 'table' of Q-values for combinations of states and actions, one simply picks the next action with the highest Q-value for that state.

In contrast, model-based RL involves creating a model of the environment, allowing the simulation of events to guide decision-making. In a neuroscience context, the model of transition structure learned through model-based RL is often described in terms of a 'cognitive map' (Tolman, 1948). This enables greater behavioural flexibility, particularly when extrapolating behaviour to new situations, but comes at the expense of efficiency, as model-based planning requires a new 'rollout' of states each time.

There are alternative algorithms, e.g. the successor representation (SR) approach involves a compromise between model-based and model-free learning which is fast but also semi-flexible (Dayan, 1993). Specifically the SR approach caches the expected future occupancy of states, i.e. rather than storing a state / action and the eventual reward, the SR approach stores a state / action and the states that follow it. Model-based planning can deal with changes to rewards or transition statistics by simulating what might happen, whereas the SR approach can deal with changes to rewards but

not transition statistics.

The hippocampal formation is implicated in planning, particularly of the model-based variety, with impairments observed in hippocampal patients (Vikbladh et al., 2019). This is unsurprising given the evidence that HF is required to generate events, as simulations of future events fall into this category. Recent research suggests that planning becomes more model-based through consolidation (Vikbladh et al., 2024). Generative sequence models trained through consolidation could correspond to the model in model-based planning, and I explore this further in Section 3.3.3.

3.1.5 Combining parametric and non-parametric memory

When considering how hippocampal and neocortical networks interact in memory, one can draw inspiration from how 'parametric' and 'non-parametric' memory are combined when using large language models (LLMs) in the recent machine learning literature. Retrieval augmented generation (Lewis et al., 2020) refers to an approach for combining LLMs with a dataset of other information. To perform a task, relevant data is retrieved from the dataset (often based on the similarity of feature representations), and used to prompt the LLM. That is, the LLM's generation is conditioned on data from the dataset.

Lewis et al. (2020) introduce the concept of retrieval augmented generation (RAG), combining a retrieval algorithm with a sequence-to-sequence model to perform an abstractive question answering task. Their system retrieves documents relevant to a query and then generates a response based on the content of both the query and the retrieved documents, improving the factual accuracy of the answers. More recent papers extend this to large language models (LLMs). This mitigates the well-known issue of LLMs 'hallucinating' facts, and enables outputs to be generated based on information that was not in the LLM's training data.

To briefly illustrate this with an example, a typical task for RAG might be question answering based on a set of documents, e.g. neuroscience papers (as in Luo et al., 2024). If the papers are recent, they may not be in the training data for the chosen

LLM, and even if they are the LLM's memory for specific details of the papers may be poor (see Kandpal et al., 2023). Luckily RAG can improve the quality of generated answers by combining parametric memory (in the LLM's weights) with non-parametric memory (in an external store). Firstly, the background information is split into 'chunks' and then a text embedding model is used to produce a vector representation of each 'chunk'. These are typically transformer-based models that learn to embed texts with similar meanings nearby in a vector space (e.g. Karpukhin et al., 2020). As a result, vectors nearby to a query, e.g. 'What are grid cells for?' are typically the most relevant 'chunks' of text to that query, e.g. papers about grid cells. These vectors are stored, often in some way that supports approximate nearest neighbour algorithms for efficiency (e.g. Malkov & Yashunin, 2018). To ask the system a question, first a vector representation of the question is obtained, and then the nearest vectors are found in the external memory. A 'prompt' (the text input from which the LLM continues) is then constructed, in which the retrieved documents are given to the LLM together with the question (and any other instructions for how to answer it).

One might hypothesise that neocortical generative models and more veridical hippocampal representations could be combined in a similar way, with neocortical generations conditioned on hippocampal representations. (Of course hippocampal memory is parametric too, if parameters correspond to synaptic weights, so it would perhaps be more accurate to refer to 'learned' vs. 'encoded' memory, but I use the parametric vs. non-parametric terminology to correspond to the machine learning literature.)

3.2 Methods

3.2.1 Modelling sequence learning

The following simulations use autoregressive sequence models to represent the generative networks trained through hippocampal replay. Specifically GPT-2 (Radford et al., 2019), a transformer-based deep neural network for text generation, is used

(see Section 3.1.1).

The first stage of using GPT-2 (and similar models) is to prepare the inputs with tokenisation. A tokeniser maps commonly occurring chunks of characters to IDs (in order to look up the right token embedding in a learned embedding matrix); in the case of language tokens are often words or parts of words. In some simulations, a custom tokeniser is fitted to the dataset, and thus segments sequences into tokens based on the statistics of that dataset, whereas in others the default GPT-2 tokeniser is used. The concept of tokenisation is applicable to arbitrary sequences, but for simplicity and consistency across the simulations all stimuli are converted to strings of characters (if they are not already text-based).

As described in Section 3.1.1, the objective for training is causal language modelling, the task of predicting the next token ('chunk' of characters) in sequences from the training data. This is achieved with the Transformers Python library (Wolf et al., 2019).

What exactly does causal language modelling with a custom dataset involve? First the training data is split into blocks, and then for every block the cross-entropy loss is aggregated across all the next token prediction tasks within the block. For GPT-2, the block size (which is also the context size of the trained model) is 1024 tokens. This means that the model is trained to consider up to 1024 tokens of context when predicting the next token in a sequence. So for each block the model tries to predict the second token based on the first token, then the third token based on the first two, and so on, until it predicts the final token based on the preceding 1023.

The loss measures the difference between two probability distributions: the distribution predicted by the model and the actual distribution in the data. For each token prediction task, the actual distribution is a 'one-hot' vector with a one for the real next token and zeros elsewhere. Specifically, the cross-entropy loss for a single prediction task is calculated as the negative log probability assigned by the model to the actual next token. For a block of tokens, the total loss is the sum of the cross-entropy losses for each token prediction task within the block, and the weights

of the model are updated based on this total loss. This procedure is the same whether the model is fine-tuned or trained from scratch.

3.2.2 Sampling options

There are many ways to generate sequences given a trained sequence model like GPT-2. As a reminder, a token is a group of commonly co-occurring characters. Except for the simulation in Section 3.3.1, the same tokeniser is used as in the pre-trained GPT-2 model (Radford et al., 2019).

Greedy decoding, where the model selects the token with the highest probability as the next token in the sequence, is the simplest way to generate sequences. However this can lead to repetitive and predictable sequences, as greedy decoding always opts for the most likely option without exploring potential alternatives. Sampling from the learned probability distribution with a given temperature introduces randomness into the selection of the next item, and provides a way to control the model's 'imaginativeness'. As described in Section 3.1.1, the temperature parameter determines the 'sharpness' of the distribution from which output tokens are selected, so that sequences at a higher temperature are more 'imaginative', but more likely to be nonsensical.

Top-K sampling limits the model's choice to the K most likely next words and samples from this subset according to their probability distribution. This prevents the model from picking highly improbable words, reducing the risk of generating nonsensical text. Unlike top-K sampling, top-p (nucleus) sampling uses a cumulative probability threshold (p) and then selects from the smallest set of items whose combined probability is below this threshold. This method allows the model to consider a broader or narrower set of options depending on the certainty of its predictions, which can lead to more coherent outputs.

Beam search is not a sampling method but a search strategy that expands on greedy decoding by considering multiple potential paths through the model's probability 'landscape'. At each step, it keeps a fixed number (the beam width) of the most probable sequences generated so far and extends them, eventually choosing the

sequence with the highest overall probability. Beam search is particularly useful for tasks requiring high-quality outputs, such as translation or summarisation, but it can be computationally intensive.

Unless stated otherwise, sequences generated below are sampled from the probability distribution of tokens with a given temperature. However some of the variations described above are also used.

3.2.3 Training procedure

The hippocampus is not modelled explicitly in the subsequent simulations. However, the training data for the generative networks is intended to represent replayed sequences from the hippocampus, as in Chapter Two.

In some simulations, existing GPT-2 weights (Radford et al., 2019) are used as the starting point for further training, and in others the GPT-2 architecture is trained from scratch with randomly initialised weights. The 'further training' option is used for the simulations involving language and the model-based planning simulation because of the complexity of the stimuli. The 'from scratch' option is used for the statistical learning and structural inference simulations because the stimuli are relatively simple.

Two different sized variants of GPT-2 are used. The 'small' GPT-2 model has 117 million parameters, including 12 transformer blocks. The 'medium' GPT-2 model has 345 million parameters, including 24 transformer blocks. For both variants, the training data used by Radford et al. (2019) is WebText, a dataset of online content scraped from outbound Reddit links and further processed to ensure quality. (The default tokeniser for GPT-2 is also fitted to this dataset.)

The small variant of GPT-2 (Radford et al., 2019) was used in the statistical learning task because of its simplicity, and in the model-based planning task because many iterations of training were required, so the smaller model minimised cost. The other simulations (structural inference on spatial and family tree graphs, gist-based

distortions, and event extension / contraction) used the medium variant of GPT-2.

The simulations in Section 3.3.3 and in Chapter Four involve tasks where the stimuli are limited, but where it is reasonable to assume that the model has relevant 'background knowledge'. In these cases, the model is first 'pre-trained' before the task itself begins, typically on sequences which mirror the structure but *not* the content of the task stimuli. (For example, in the Vikbladh et al. (2024) task, the model after pre-training corresponds to the neocortical network at the point the participant passes a 'rules quiz' but before exposure to the task stimuli.)

Simulation	**Section**	**Training method**	**Model size**	**Training data**
Statistical learning	3.3.1	From scratch	Small	Sequences of tones from Durrant et al. (2011) represented by '2,2,3,1,. . . '
Structural inference (spatial)	3.3.2	From scratch	Medium	Walks on a graph of form 'mv SOUTH sz WEST li EAST sz . . . '
Structural inference (family trees)	3.3.2	From scratch	Medium	Walks on a graph of form 'yu GRANDPARENT_OF mi SIBLING_OF vb . . . '
Model-based planning	3.3.3	Further training	Small	Tasks from Vikbladh et al. (2024) represented by 'START: yellow vehicle, STOP: green, REWARD: animal, SEQUENCE: red animal (2), green vehicle (-1)'
Gist-based distortions	3.3.4	Further training	Medium	English language text
Event extension and contraction	3.3.5	Further training	Medium	English language text

Table 3.1: Summary of simulations and their training details. See Section 3.2.3 for further details of the training methods and model sizes, and Section 3.3 for further details of the tasks and training data.

3.3 Results

3.3.1 Statistical learning

I model the learning of sequential structure by training GPT-2 on sequences of arbitrary symbols with a statistical pattern, as in Durrant et al. (2011). To explore the effect of sleep on statistical learning, Durrant et al. (2011) constructed two types of sequence, both made up of regular tones at differing frequencies. One type had a structure in which the preceding two tones determined the next - i.e. each sequence was a second order Markov chain - except for 10% of transitions which were random to avoid repetition. The other type was unstructured, with random transitions between tones. After listening to a structured sequence for several minutes, participants were tested on their ability to distinguish short structured and unstructured sequences. Delayed recall was then tested, after a night's sleep for one group, and after a waking rest for the other. The authors found that sleep improved performance more than waking rest, suggesting that consolidation during sleep promotes learning of sequential structure.

This simulation aims to test the hypothesis that statistical learning of sequential structure through consolidation in Durrant et al. (2011) is consistent with the sequential model. I produced a set of sequences of 'tones' using the transition structure in Durrant et al. (2011), and represented them as comma-separated strings (e.g. '2,2,3,1,. . . ',). These are intended to mimic the stimuli encoded in the hippocampus and replayed during rest. After fitting a tokeniser with a vocabulary of ten tokens to the data, the small GPT-2 architecture was trained from scratch for three epochs on 2000 such sequences, each made up of 50 'tones'.

At the end of each epoch of the training, the perplexity was calculated for test sets of 100 structured and 100 unstructured sequences. Perplexity represents the unexpectedness or schema incongruency of a sequence given the learned statistical structure; the greater the perplexity of the unstructured compared to structured test sequences, the greater the ability to distinguish them. As Figure 3.2c shows, the difference in perplexity between structured and unstructured sequences increases

over time as the generative model is trained, in agreement with the finding that consolidation improves ability to distinguish between structured and unstructured sequences.

At the end of training, the model generated sequences (with a temperature of 0.1), conditioned on a single instance of each of the five tones to ensure some variation. The average transition probabilities given the preceding two tones were calculated, and the resulting transition structure plotted. As Figure 3.2b shows, training the generative network on sequences from the Durrant et al. (2011) task leads to the network learning the transition structure of the stimuli, so that it can predict the next item in structured sequences.

3.3.2 Relational inference

Consolidation not only extracts statistical regularities from episodic memories (Durrant et al., 2011), but also supports relational inference (Ellenbogen et al., 2007; Kumaran, 2012). A spatial example of relational inference is the finding of shortcuts, as this relies on the common structure of space, and a non-spatial example is inferring that A is the grandfather of C from the knowledge that A is the father of B, and B is the father of C, as this relies on the common structure of family trees. The relations in these tasks can be seen as edges in graphs, as simulated by the Tolman-Eichenbaum machine (TEM; Whittington et al., 2020) in the domain of multiple tasks with common transition structures.

This simulation aims to test the hypothesis that consolidation enables relational inference in the sequential model. As in Whittington et al. (2020), this type of inference is framed in terms of graph transitions. I consider inference in two types of graph: a spatial graph and a simple family tree graph. In the spatial graph, a three-by-three grid represents a simple 2D environment, where the nine nodes are locations and the edges between them ('NORTH', 'EAST', 'SOUTH' and 'WEST') are possible transitions (Figure 3.3a). Whilst each graph's structure is the same, nodes are labelled with names to represent arbitrary features at a particular location (random pairs

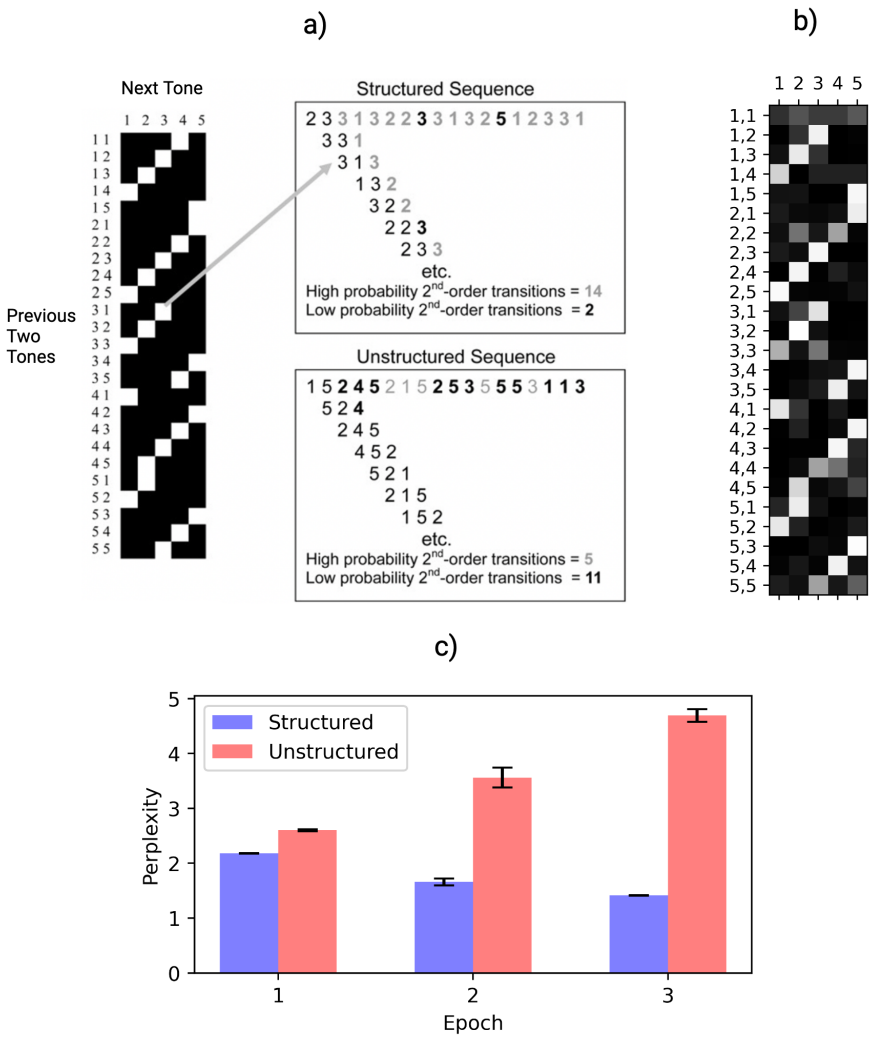

Figure 3.2: Statistical learning of sequential structure. a) Figure One from Durrant et al. (2011), showing the transition structure for structured sequences of tones in the task. In these sequences, the preceding two tones determine the next according to this matrix, except for 10% of transitions which are random to avoid repetition. That is, white squares indicate a 90% probability of a certain tone coming next in the sequence. b) The learned transition structure when GPT-2 is trained from scratch on data from Durrant et al. (2011), with transition probabilities extracted from data generated by the trained model (with a temperature of 0.1). c) Perplexity over time for structured vs. unstructured sequences, for a model trained on structured sequences. The ability to distinguish the two types based on their perplexity increases over time. Error bars give the SEM.

of letters are used to increase the possible number of names). Trajectories through the environment are walks on the resulting directed graph, which are represented as strings such as 'ab EAST wd SOUTH ea WEST hn'. (Note that Figure 3.3 represents the grid as a graph for consistency with the family tree model, but this is equivalent to the grids used in Chapter Four.)

The family tree graph has a simple structure for illustrative purposes, consisting of two children, their parents, and two sets of grandparents. See Figure 3.3b. I model this as a directed graph with edges 'PARENT_OF', 'CHILD_OF', 'SPOUSE_OF', 'SIBLING_OF', 'GRANDPARENT_OF', and 'GRANDCHILD_OF'. As in the spatial graph case, all graphs have the same structure, but each graph has different names assigned to its nodes, with each name representing a particular individual. Walks on the graph are represented by strings such as 'lk PARENT_OF nd SIBLING_OF re'.

GPT-2 models were then trained in order to explore how training generative 'world models' through consolidation might give rise to structural inference abilities. In each case, I created 100,000 graphs with the same structure but randomly chosen values (pairs of letters) for the nodes. A random walk of 50 transitions was sampled from each graph to create the training data. This was intended to represent - in a very abstract way - sequences of observations that might be experienced, encoded in the hippocampus, then replayed offline. GPT-2's medium-sized architecture was then trained from scratch for five epochs.

After training the models, I simulated the spatial task of predicting the next location in a sequence as the prediction of the next node in a graph, and the non-spatial task of inferring relationships between family members the same way. To test inference, I defined a set of cycles in the graph for which the final destination could be inferred given the sequence so far. For example, the next item after 'uq NORTH sx EAST tp SOUTH ec WEST' can be inferred to be 'uq' given the structure of spatial graphs, and the next item after 'qk PARENT_OF xm PARENT_OF vw GRANDCHILD_OF zq SPOUSE_OF' can be inferred to be 'qk' given the structure of family tree graphs. (Only a subset of these tasks were tested as there are a very large number of possible

loops, particularly for the family tree task.) These templates were then populated with random pairs of letters, so that none of the sequences used for testing featured in the training data. Beam search with five beams was used to generate predictions. On some of the family tree inference problems there are multiple possible answers which are consistent with, but cannot be inferred from, the sequence so far (such as the imagined family member 'ef' in 'ab CHILD_OF cd PARENT_OF ef'), and these are counted as incorrect, therefore this is quite a harsh performance metric.

Figure 3.4a and b show the 'loss' (aggregated error on the training data) of the spatial model and family tree model respectively. In both cases the loss gradually decreases, indicating improved ability to predict the next node on the set of graphs used for training, which corresponds to the consolidation of previously experienced environments.

Figure 3.4c shows good performance on a range of *novel* structural inference tasks. Simpler inferences include inferring that going one step west then east takes one back to the original location, e.g. the correct continuation of 'ab EAST cd WEST' is 'ab'. Similarly, a correct continuation of 'ab PARENT_OF cd CHILD_OF' is 'ab'. (As mentioned above, any other 'name' would be potentially correct too, but cannot be inferred given the information given, so is treated as incorrect in this task.) But surprisingly complex inferences are also possible, e.g. that 'kh CHILD_OF oi SPOUSE_OF tv CHILD_OF fh SPOUSE_OF xr GRANDPARENT_OF gq SIBLING_OF' is followed by 'kh'. Table 3.2 gives the average score for each 'template', while Figure 3.4c aggregates these results by the number of graph transitions (or 'hops') in the sequence.

The results are consistent with the claim that consolidation supports relational inference and generalisation. Furthermore they suggest that models trained on a simple prediction error minimisation objective can learn an abstract transition structure. Unlike in TEM (Whittington et al., 2020), in which structural regularities and arbitrary specifics are factorised by design, the model learns to separate structure and content (i.e. roles in the graph and the entities that fill them).

Many inference problems can be framed in terms of graphs or transition structures, so this approach could be more generally applicable.

3.3.3 Model-based planning

Vikbladh et al. (2024) design a task to assess the contribution of different strategies (i.e. model-based, model-free, or successor representation approaches) to planning, and explore how this changes with consolidation. As Figure 3.5, reproduced from Vikbladh et al. (2024), summarises, the task is as follows: nine items ('red animal', 'green animal', 'yellow animal', 'red vehicle', 'green vehicle', 'yellow vehicle', red fruit', 'green fruit', and 'yellow fruit') are arranged in a loop. The participants observe clockwise subsequences from the loop, determined by start, stop, and reward conditions. The start condition gives the item the sequence starts at. The stop condition gives the colour the sequence terminates at, e.g. if this is 'green' the sequence would stop upon reaching the first green object. The reward condition determines the object that receives a reward of 2 (while other objects receive -1).

Participants first learn the rules of the task, proceeding to the next stage after passing a quiz. They then experience sequences made up of the stimuli above in a random order, with one sequence per start state, and a random stop and reward condition. Crucially, each participant sees only one stop and one reward condition during training, requiring them to do 'revaluation' when deciding whether to accept or reject a trial with a new stop or reward condition in the subsequent testing. Performance is tested before and after a seven day delay to assess the effect of consolidation. (Note that participants only get feedback as to the sequence of states and reward values during the training phase. Thus there is no externally-driven learning during or between the two revaluation phases.)

The task allows model-based, model-free and successor representation strategies to be disentangled, since only model-based planning allows revaluation of both reward and transition statistics, whereas the successor representation allows only reward revaluation, and model-free planning allows neither. Vikbladh et al. (2024) found

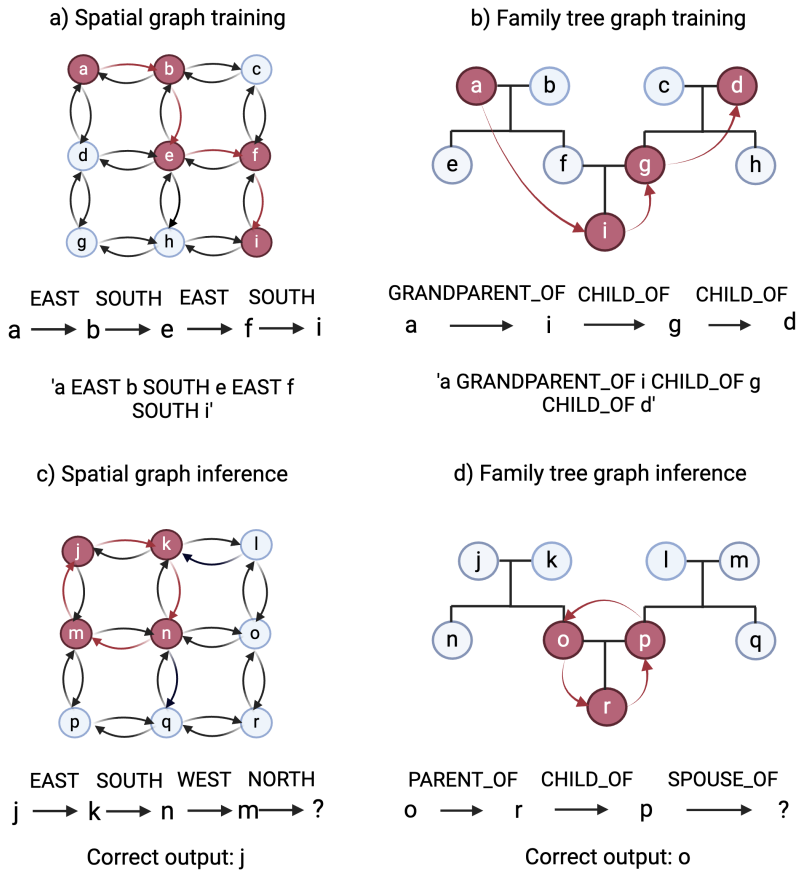

Figure 3.3: Learning structural regularities in graphs. a) The spatial graph task. Spatial trajectories are modelled as walks on the graph, and represented as strings as shown. b) The family tree graph task. Relationships between family members are modelled as walks on the graph, and represented as strings as shown. c) Testing structural inference, modelled as the ability to complete sequences on unseen graphs based on structural regularities, in the spatial graph. For example, it is possible to infer that the next location in the sequence shown is 'j', given the structural regularities of spatial graphs. d) Testing structural inference in the family tree graph. For example, it is possible to infer that the next person in the sequence shown is 'o', given the structural regularities of family tree graphs.

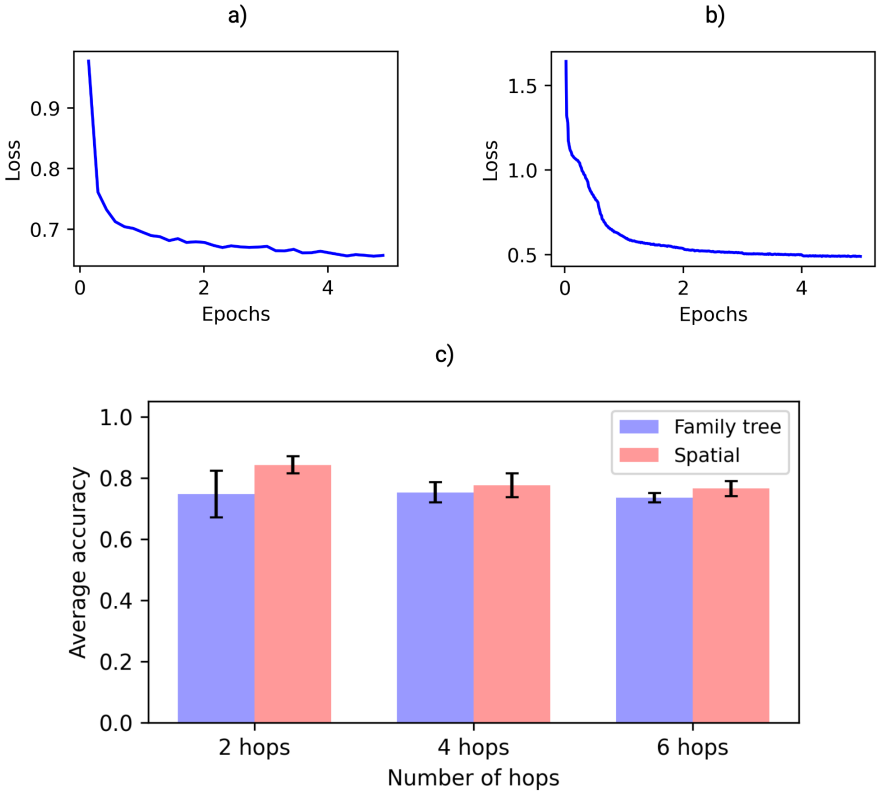

Figure 3.4: Structural inference in spatial and family tree graphs. a) The loss of GPT-2 trained for five epochs on sequences from spatial graphs. b) The loss of GPT-2 trained for five epochs on sequences from family tree graphs. c) Loop completion performance for the spatial and family tree models, grouped by the number of edges ('hops') in the template. Error bars give the SEM. See Table 3.2 for the accuracies for each template.

(a) Spatial task inference performance

Inference template	**Mean accuracy**
{} EAST {} WEST {}	0.87
{} WEST {} EAST {}	0.82
{} NORTH {} SOUTH {}	0.81
{} SOUTH {} NORTH {}	0.87
{} EAST {} SOUTH {} WEST {} NORTH {}	0.71
{} SOUTH {} WEST {} NORTH {} EAST {}	0.78
{} WEST {} NORTH {} EAST {} SOUTH {}	0.8
{} NORTH {} EAST {} SOUTH {} WEST {}	0.81
{} EAST {} EAST {} NORTH {} WEST {} WEST {} SOUTH {}	0.79
{} NORTH {} NORTH {} WEST {} SOUTH {} SOUTH {} EAST {}	0.74

(b) Family tree task inference performance

Inference template	**Mean accuracy**
{} CHILD_OF {} PARENT_OF {}	0.81
{} PARENT_OF {} CHILD_OF {}	0.76
{} GRANDCHILD_OF {} GRANDPARENT_OF {}	0.8
{} GRANDPARENT_OF {} GRANDCHILD_OF {}	0.62
{} CHILD_OF {} CHILD_OF {} GRANDPARENT_OF {} SIBLING_OF {}	0.7
{} CHILD_OF {} SPOUSE_OF {} PARENT_OF {} SIBLING_OF {}	0.75
{} PARENT_OF {} SIBLING_OF {} CHILD_OF {} SPOUSE_OF {}	0.79
{} PARENT_OF {} PARENT_OF {} GRANDCHILD_OF {} SPOUSE_OF {}	0.77
{} CHILD_OF {} SPOUSE_OF {} CHILD_OF {} SPOUSE_OF {} GRANDPARENT_OF {} SIBLING_OF {}	0.75
{} GRANDPARENT_OF {} SIBLING_OF {} CHILD_OF {} SPOUSE_OF {} CHILD_OF {} SPOUSE_OF {}	0.72

Table 3.2: Relational inference performance

that there was a mixture of successor representation and model-based strategies, with the model-based approach increasing with consolidation between day one and day seven. MEG decoding and analysis of response times indicated that model-based planning was associated with sequential 'rollouts' involving the medial temporal lobe on both days and prefrontal cortex on day seven.

The aim of this simulation is to test the hypothesis that model-based planning, based on the 'rollout' of sequences by the generative network, increases with consolidation in the sequential model. The stimuli for Vikbladh et al. (2024) can be represented as sequences of the form 'START: yellow fruit, STOP: red, REWARD: animal, SEQUENCE: green fruit (-1), red animal (2)', which makes it straightforward to train the small GPT-2 model. The task is simulated as follows: i) pre-train the model so that it learns the rules of the task, ii) train on stimuli for a particular task, representing consolidation through hippocampal replay, and iii) compare the generative model's accept / reject predictions to the true values over the course of consolidation.

To describe this in more detail, first I pre-train the model on arbitrary stimuli, to simulate learning the rules of the task. The pre-training data consists of the 81 possible start / stop / reward combinations for each of 1000 sets of nine items, giving a total of 81,000 sequences. Each item is a random pairing of one of three random adjectives and one of three random nouns; the start condition is one of the nine items, the reward condition is one of the three nouns, and the stop condition is one of the three adjectives. For example, one sequence might be 'START: stripy gerbil, STOP: angry, REWARD: cloud, SEQUENCE: busy cloud (2), angry plate (-1)'.

The GPT-2 model is fine-tuned on this shuffled dataset for three epochs. After this stage of training, the model can be given a prompt with randomly chosen stimuli / conditions, e.g. 'START: blue chair, STOP: sad, REWARD: bug, SEQUENCE:', and generate a sequence consistent with the rules, e.g. 'blue table (-1), sad bug (2)'. But the model knows nothing about the task stimuli or their order, i.e. none of the stimuli from the Vikbladh et al. (2024) task were used in fine-tuning. This is supposed to be equivalent to the participant at the point they pass the rules quiz,

prior to experiencing the task stimuli.

To simulate the task, the stimuli from Vikbladh et al. (2024) are used ('red animal', 'green animal', 'yellow animal', 'red vehicle', 'green vehicle', 'yellow vehicle', red fruit', 'green fruit', and 'yellow fruit'), shuffled into a random order. A single random stop and reward condition are selected and combined with each start state, as for the human participants. The training stimuli are then just nine sequences with these stop and reward criteria (one per start position). These nine sequences are oversampled to 1000 items (i.e. 1000 'replayed' samples are taken, so that 9 sequences are presented 1000 times in random order) to prevent overfitting to the order of the sequences. The model is further fine-tuned for 20 epochs on this dataset (i.e. there are 20 iterations of training on the dataset of 1000 samples). This simulation is rerun five times, with a random ordering of the stimuli, reward condition, and stop condition in each trial.

For each trial, three test sets are created: one for sequences requiring transition revaluation (i.e. with a new stop condition), one for sequences requiring reward revaluation (i.e. with a new reward condition), and one for sequences requiring both kinds of revaluation (i.e. with a new stop *and* reward condition). Performance on the accept / reject decision-making task is tested as follows:

1. Obtain the predicted sequence from the model given an input of the form 'START: yellow animal, STOP: red, REWARD: animal, SEQUENCE:'. Greedy decoding, i.e. simply taking the most probable next token, is used.

2. Extract the predicted rewards (any numbers in round brackets) from the sequence.

3. If the sum of the predicted rewards is greater than zero, the decision is to 'accept' the trial, and if the sum of the predicted rewards is less than or equal to zero, the decision is to 'reject' it. (This involves a slight simplification, as a trial with a predicted reward of zero could be accepted or rejected with the same result for the participant.)

4. Calculate the accuracy of the accept / reject decisions by comparing the predictions with the true values. As above, the true value is 'accept' if the sum of rewards extracted from the sequence is greater than zero, and is otherwise 'reject'.

Figure 3.6 shows that the performance on all three types of task (those requiring reward revaluation, those requiring transition revaluation, and those requiring both) improves over time. This suggests that the generative model learns the order of the items through 'consolidation' such that it can predict the sequence of items and rewards given novel reward and/or stop conditions.

In order to explore whether planning becomes more model-based, Figure 3.7 compares the network's decisions to the decisions that would be made given four strategies: model-free (taking the accept / reject decision for the same start item from the training data), reward revaluation (the decision given the ability to perform reward revaluation but *not* transition revaluation), transition revaluation (the decision given the ability to perform transition revaluation but *not* reward revaluation), and model-based (the correct decision, as would be deduced from a mental model of transition statistics).

Two statistical approaches are used to the compare the observed decisions to the decisions for a given strategy. Firstly, linear regression is used to predict the observed accept / reject decisions, with accept / reject decisions given each strategy as the regressors. The regression coefficients then indicate to what extent the decisions depend on each strategy. Secondly, the Pearson correlation coefficients between observed accept / reject decisions and the accept / reject decisions given each strategy are calculated.

The results in Figure 3.7 show that the trained network's behaviour is most consistent with model-based planning, and that the generative network becomes more 'model-based' over the course of 'consolidation'. This is a potential explanation of the effect of consolidation in Vikbladh et al. (2024), as the generative model can learn from hippocampal replay without any externally-driven learning.

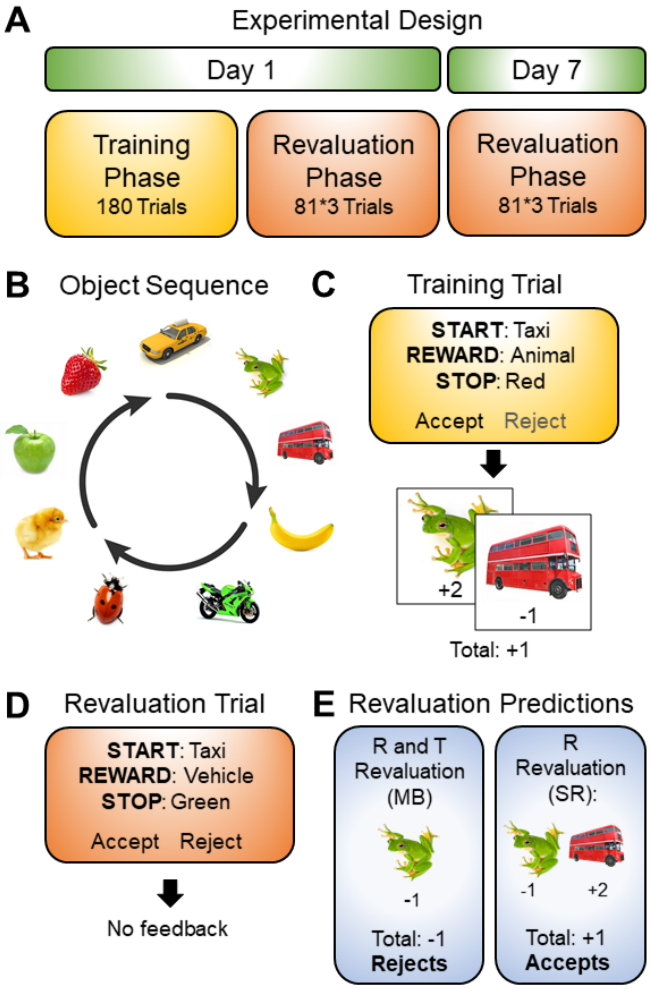

Figure 3.5: Figure reproduced from Vikbladh et al. (2024), showing the task design. a) Training takes place on day one, with tests of reward and transition revaluation on days one and seven. b) The stimuli are nine items arranged in a clockwise 'loop'. c) The task is to decide whether to accept or reject a trial given a start item and reward and stop conditions, based on the expected rewards obtained from the sequence. d) The participant sees only a single reward and stop condition during training, and revaluation tests performance when the reward and/or stop conditions change. e) Model-based planning (left) and the successor representation approach (right) allow different kinds of revaluation.

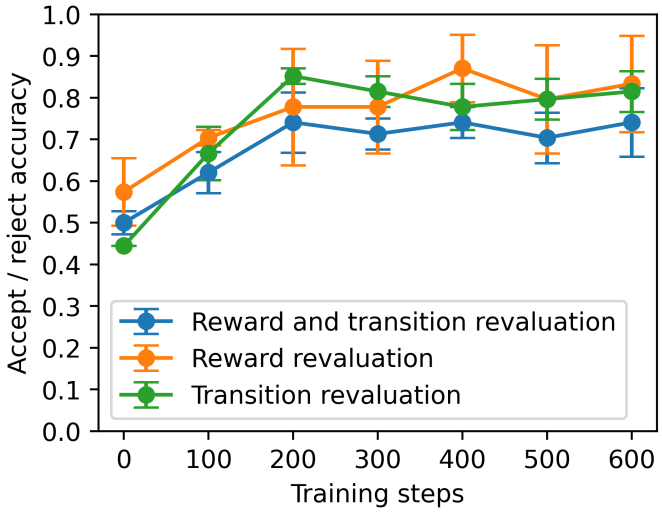

Figure 3.6: Planning over time. Accuracy of accept vs. reject choices over the course of training on a particular trial, for three types of problem. In each case the model pre-trained on the rules of the task was further trained on just nine sequences for 20 epochs. Reward revaluation problems are those where a new reward condition is used at test time which was not included in the training data, transition revaluation problems are those where a new stop condition is used at test time, and reward and transition revaluation problems are those where both the reward and stop conditions are new. The mean across five trials is taken. Errors bars give the SEM.

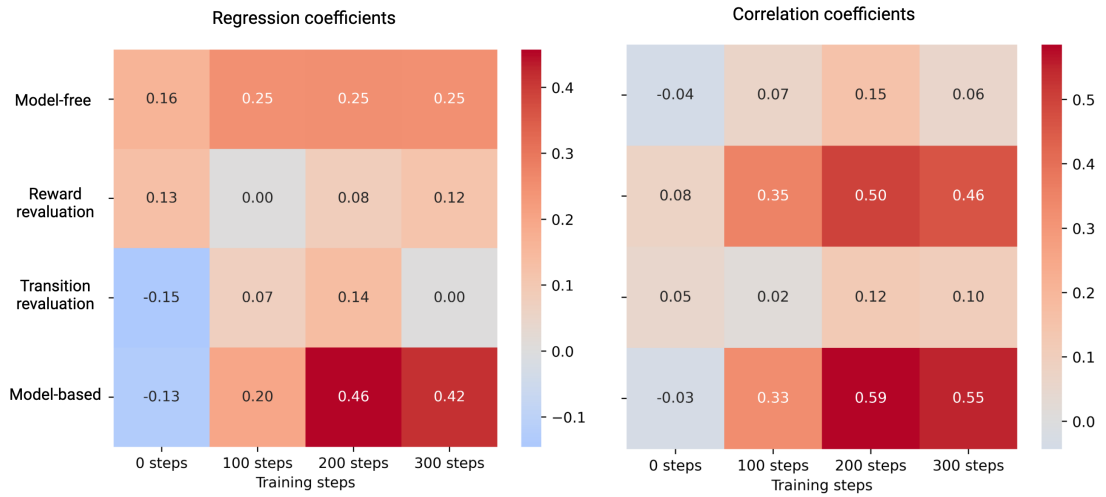

Figure 3.7: Heatmaps showing regression coefficients (left) and Pearson correlation coefficients (right), capturing which 'strategy' best explains the results over the course of training. The network's decisions after simulating 'consolidation' of the nine sequences are compared to the decisions that would be made given four strategies: model-free (taking the accept / reject decision for the same start item from the training data), reward revaluation (the decision given the ability to perform reward revaluation but not transition revaluation), transition revaluation (the decision given the ability to perform transition revaluation but not reward revaluation), and model-based (the correct decision, as deduced from a learned model of transition statistics for the task). Left: linear regression is used to predict the observed accept / reject decisions, with accept / reject decisions given each strategy as the regressors. The corresponding regression coefficients are shown. Right: the Pearson correlation coefficients between observed accept / reject decisions and the accept / reject decisions given each strategy are shown. The results show that the generative network becomes more 'model-based' over the course of 'consolidation'.

3.3.4 Gist-based distortions for sequences

In the Bartlett (1932) experiment, students heard a story called 'The War of the Ghosts' and were asked to recall it after different time intervals. The story, a Native American myth, was chosen to be culturally unfamiliar, making memory distortions more pronounced. Bartlett found that the story was recalled in a way that was consistent with the students' background knowledge of the world, and details were added to explain unusual elements of the story (i.e. confabulation and rationalisation were observed). As replicated by Bergman and Roediger (1999), memory distortion increased over time after encoding. This simulation aims to test the hypothesis that recalled narratives are distorted based on background semantic knowledge (Bartlett, 1932).

To simulate consolidation, I fine-tuned the existing medium-sized GPT-2 model on the Bartlett (1932) story in addition to 'background data'. Recall of the story was explored by giving the network the first few words of the story ('One night two young men from Egulac'), and inspecting the predicted continuation.

To explore the effect of the model's 'priors' on recall of narratives, the background data distribution was varied. The 'Shakespeare model' used 5000 lines from Shakespeare plays as the background data, the 'News model' used 5000 short news stories from the 'AG's News' dataset (Zhang et al., 2015), and the 'Scientific papers' model used 5000 abstracts scraped from PubMed (Cohan et al., 2018).

Between 2 and 10 repetitions of the Bartlett story were then shuffled together with these 5000 items, and the model was trained for 5 epochs on the combined dataset. Note that the number of replays variable in the figures below refers to the *total* number of repetitions of the Bartlett story, so varies from 10 to 50. The temperature for generating continuations was also varied.

Word clouds are used to visualise semantic intrusions. They show terms in the recalled Bartlett stories which do not occur in the original (with common words, i.e. 'stopwords' like 'the', excluded). The 'wordcloud' Python package was used to produce all subsequent word clouds.

When the Bartlett story is 'consolidated' into the generative network memory distortions are observed, as in the human data (see Table 3.3). Distortions in recalled stories reflect the 'priors' of the generative network. The word clouds in Figure 3.8b show that new words added to the story (i.e. 'semantic intrusions') are representative of the background dataset used. Consistent with this, Figure 3.8a shows that the recalled stories move closer towards the background dataset in text embedding space, with the model 'all-MiniLM-L12-v2' from Reimers and Gurevych (2019) used to generate the embeddings. Specifically, the cosine distance between the mean embedding for each category and the *original* story is compared to the cosine distance between the mean embedding for each category and the *recalled* stories. See also Table 3.4 for selected examples of semantic intrusions from the three models.

More 'semantic intrusions' are observed at higher temperatures, as Figure 3.9a and c and Table 3.3 show. However if the original Bartlett story is replayed enough times, the level of distortion reduces (Figure 3.9a and b). This is consistent with the view that memories are replayed to neocortex until they are remembered 'well enough', rather than indefinitely until recall is perfect. (Note that the increase in gist-based distortion over time reflects the increasing dependence of recall on the generative network as consolidation progresses, as opposed to the initial hippocampal trace, so is not inconsistent with this finding.)

3.3.5 Event extension and contraction

The brain segments experience into discrete events (Newtson, 1973; Newtson & Engquist, 1976), with distinctive neural activity at event boundaries (Baldassano et al., 2017; Ben-Yakov & Dudai, 2011). Event extension is the tendency to extend certain events in memory, while event contraction is the tendency to curtail them (Raykov et al., 2023). Event extension is typically observed when an event stops before its natural end, while event contraction is typically observed when an event continues beyond this point, mirroring the findings about which image stimuli produce

Type	**Recalled story**
Original	One night two young men from Egulac went down to the river to hunt seals and while they were there it became foggy and calm. Then they heard war-cries, and they thought: "Maybe this is a war-party". They escaped to the shore, and hid behind a log. Now canoes came up, and they heard the noise of paddles, and saw one canoe coming up to them. There were five men in the canoe, and they said: ...
No sampling	One night two young men from Egulac went to the river to hunt seals. They came upon a party of warriors, and they began to fight. But the young man escaped, and hid behind a log. Now the warriors came up to him, and said, "What do you think? You have escaped. Come along." So the young man went along, and the warriors went on up the river to a town on the other side of Kalama. The people ...
Temperature of 0.25	One night two young men from Egulac went to the river to hunt seals. The sun rose and the fog rose, and they heard war-cries and saw one canoe coming up to them. There were five men in the canoe, and they said: "What do you think? We wish to take you along. We are going up the river to make war on the people." One of the young men said,"I have no arrows." "Arrows are in the canoe," they said. "I ...
Temperature of 0.5	One night two young men from Egulac went to the river to hunt seals. They came upon a party of warriors, and one of the young men said: "Quick, let us go home. We are going up the river to fight. They are coming up the river to cut us down." One of the warriors said, "I will not go along. I might be killed. My master will not let me go." So the other returned home. The young man went back to Egulac ...
Temperature of 0.75	One night two young men from Egulac went to the river to hunt seals. One of them was a boy and the other was a girl. The boy was well armed, and his arrows were well-carved. The boy said he would go with them, and the girl said she would not go with him. Now is the season for the hunt: the water is calm, and the sun is up. But the young man said he would not go, and they heard war-cries, and ...

Table 3.3: Recalled stories for different temperatures, for a model trained on 30 replays of the Bartlett story with Shakespeare plays as the background data distribution.

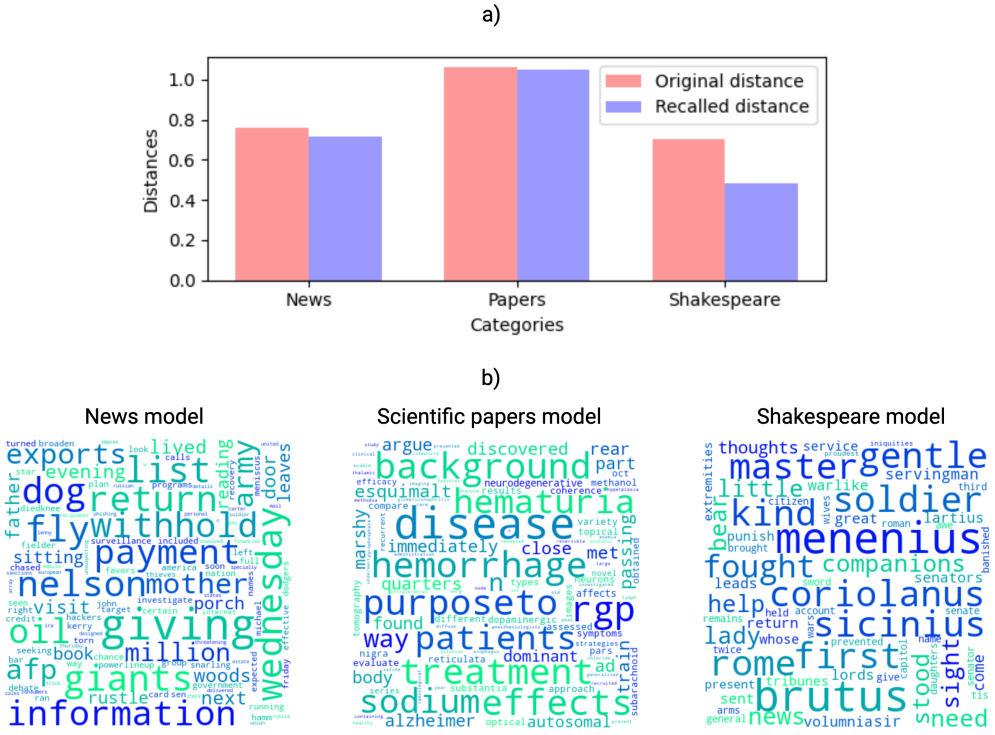

Figure 3.8: The effect of the background data distribution on narrative distortions. The 'Shakespeare' model is trained on lines from Shakespeare plays, the 'News' model is trained on the 'AG's News' dataset (Zhang et al., 2015), and the 'Scientific papers' model is trained on abstracts of papers scraped from PubMed (Cohan et al., 2018). a) The cosine distance between the mean embedding for each category and either the original story (red) or the recalled story (blue). The embeddings of the training data plus the recalled story for each model are obtained using 'all-MiniLM-L12-v2' from Reimers and Gurevych (2019). Recalled stories become more similar to the background dataset. b) The word clouds show terms in the recalled Bartlett stories which do not occur in the original (with common words, i.e. 'stopwords', excluded) for the three models. The 'semantic intrusions' at a temperature of 0.75, after ten replays of the Bartlett story, reflect the background data distribution.

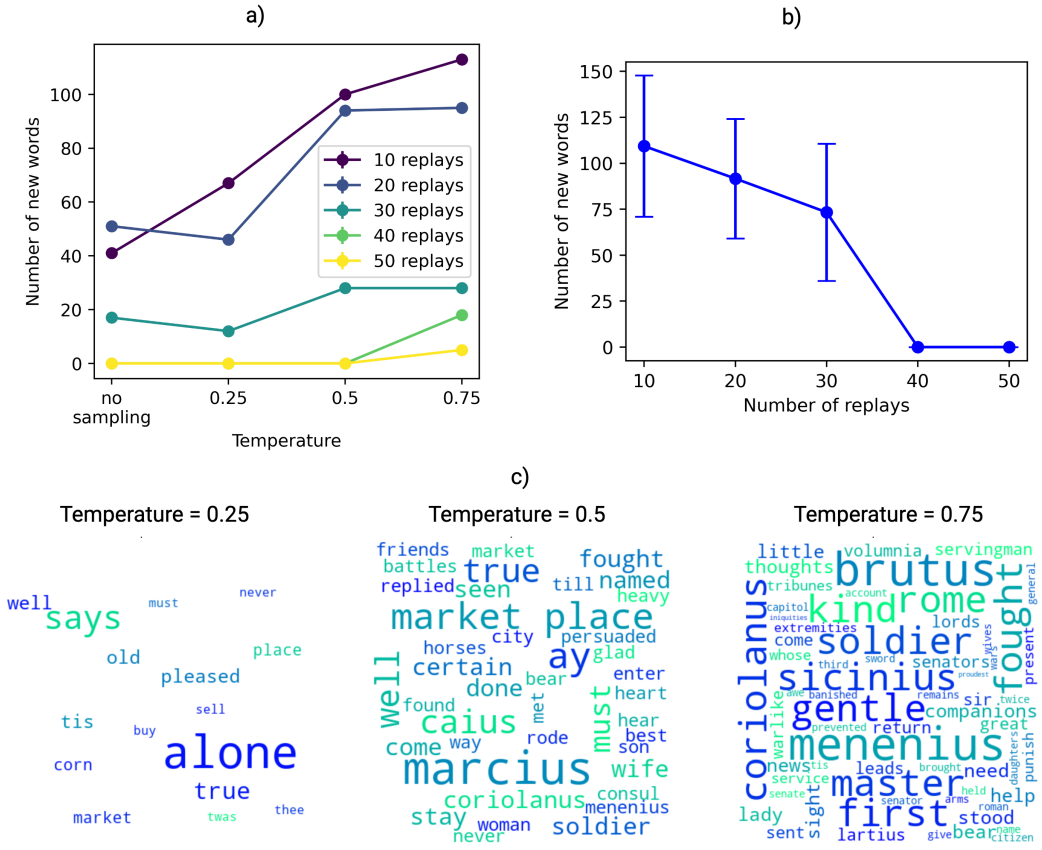

Figure 3.9: Effect of temperature and replay quantity on narrative distortions. a) The effect of temperature on narrative distortions is explored using the model with Shakespeare lines as the background data distribution. The plot shows the number of new words against temperature, for differing amounts of replay. More 'semantic intrusions' are observed at higher temperatures. The more times the story is replayed, the more accurate the recall. b) The number of new words in the recalled story that did not feature in the original story, for different numbers of replays of the Bartlett story (i.e. occurrences in the training data). There are fewer semantic intrusions with more repetitions. Results are averaged across models trained on three different 'background' datasets. Errors bars give the SEM. c) The word clouds show terms in the recalled Bartlett stories which do not occur in the original (with common words, i.e. 'stopwords', excluded) for a range of temperatures, with ten replays of the Bartlett story.

Model	Recalled story
Shakespeare	"Behold I accompanied the canoes to make war on the people. He who does this will pay for it in blood."
	He was dead. The gods have mercy on thee!
	"I have killed many of your fellows, and I have many arrows in my hand. I would fain have killed you."
News	So Egulac became a war-zone.
	But the young man escaped and went to the police station
	Now Egulac is a city on the western coast of Lake Superior
Scientific papers	Nowhere do we find this statement in the traditional accounts of the war-party, but it is clearly there in the oral tradition of the warriors.
	The young man received intravenous medication
	So we have a case of two young men who went down to the river to hunt seals.

Table 3.4: Recalled stories for different models, showing how semantic intrusions reflect the 'priors' of the generative network. Examples are selected from a range of replay counts and temperatures.

boundary extension vs. contraction.

Raykov et al. (2023) explore the recall of three types of video, at the end of the encoding session and after a week's delay: complete videos end at a natural event boundary, incomplete videos are curtailed before this point, and updated videos are extended beyond it. The authors find that in the incomplete condition, 'participants often falsely recalled additional details going beyond the last interrupted action', known as extension errors, whereas in the updated condition 'participants often omitted the entirety of the new scene from their recall', known as omission errors (Raykov et al., 2023, Discussion). The authors compared the incomplete vs. updated conditions at a long vs. short retention intervals, finding a significant interaction between condition and delay for the frequency of both extension and omission errors. The increase in the event extension and contraction effects over time indicates that the errors cannot be attributed solely to processes at encoding time, suggesting that consolidation promotes these distortions. (See Figure 3.10a.)

This simulation aims to test the hypothesis that the sequential model captures the effect of consolidation in Raykov et al. (2023) (see Section 3.3.6 for how the immediate effects could be explained). Event extension and contraction are modelled with simple stories in text form, each only a few lines long (Rashkin et al., 2018). Three types of narrative are used. Firstly, a large majority are typical stories, unmodified from the Rashkin et al. (2018) dataset. Secondly, an atypically shortened (or 'incomplete') story has 100 characters removed. Thirdly, an atypically lengthened (or 'updated') story has 100 characters appended from another, randomly selected story. See Table 3.5 for examples. Note that this is a challenging task as some stories could be interpreted as either atypically shortened or lengthened. I did not control for the story length in characters, so one cannot easily separate the effect of the content from the effect of length alone, however the stories did vary in length. Complete stories had a mean of 264.30 characters (SD = 36.21), incomplete stories had a mean of 132.89 characters (SD = 39.11), and updated stories had a mean of 346.90 characters (SD = 59.12).

Training data consisted of 100 complete stories, 10 incomplete stories, and 10 updated stories, with more complete examples used than incomplete or updated ones to establish a 'prior' for typical stories. The experiment was repeated five times, with a different sample of stories used in each trial.

The medium-sized GPT-2 model (Radford et al., 2019) was fine-tuned on this set of stories for five epochs and then recall was tested by giving the model the first few words and observing the output. The highest probability token was taken at each step (i.e. 'greedy decoding'), rather than sampling from the distribution.

Before training on the stories, neither the typical nor atypical stories are memorised, and only generic continuations are generated. But after some training, the atypical stories are remembered with characteristic extension and omission errors: on average the incomplete (atypically shortened) stories are lengthened, and the updated (atypically lengthened) stories are shortened, while complete stories stay approximately the same length (Figure 3.10). By bringing the story to a more natural conclusion or by removing an incongruous ending, the stories are made more similar to the complete

stories. See Table 3.5 for examples. (Note that even for the complete stories, some distortion is observed.)

One interpretation of these results is that these event boundary distortions are a special case of the Bartlett (1932) findings, in which confabulation, omission, and substitution are observed in memory for narratives; memories of atypically shortened events display confabulation, and memories of atypically lengthened events display omission or substitution. That is, confabulation, omission, and substitution reflect the behaviour of generative models more broadly.

There is also a clear link between the boundary extension and contraction results in Chapter Two (see Figure 2.5) and the event extension and contraction results here, despite the differing types of neural network used. In both cases, memories are distorted towards a prior encoded in the network by its previous 'experience', consistent with Bayesian views of memory (Hemmer & Steyvers, 2009).

3.3.6 Retrieval augmented generation

Memories can be used to support problem solving immediately after encoding, not just after consolidation. This simulation aims to test the hypothesis that the generative network and hippocampal network could work together to achieve this in a way resembling 'retrieval augmented generation' (RAG). Inference from recent memories is modelled as a process whereby relevant sequences from the hippocampus are retrieved and used to condition the generative model.

I create a 'toy example' of retrieval augmented generation using the two models from the structural inference results above (one model trained on spatial graphs, and one trained on family tree graphs). 100 graphs were constructed, each missing one edge, so that inference from memory could be tested. A walk on each of these graphs was stored in the 'hippocampus' (simply a list of strings in this example). For each missing edge, a query of the form 'ab EAST' or 'cd PARENT_OF' was constructed for the spatial and family tree graphs respectively. In other words, if

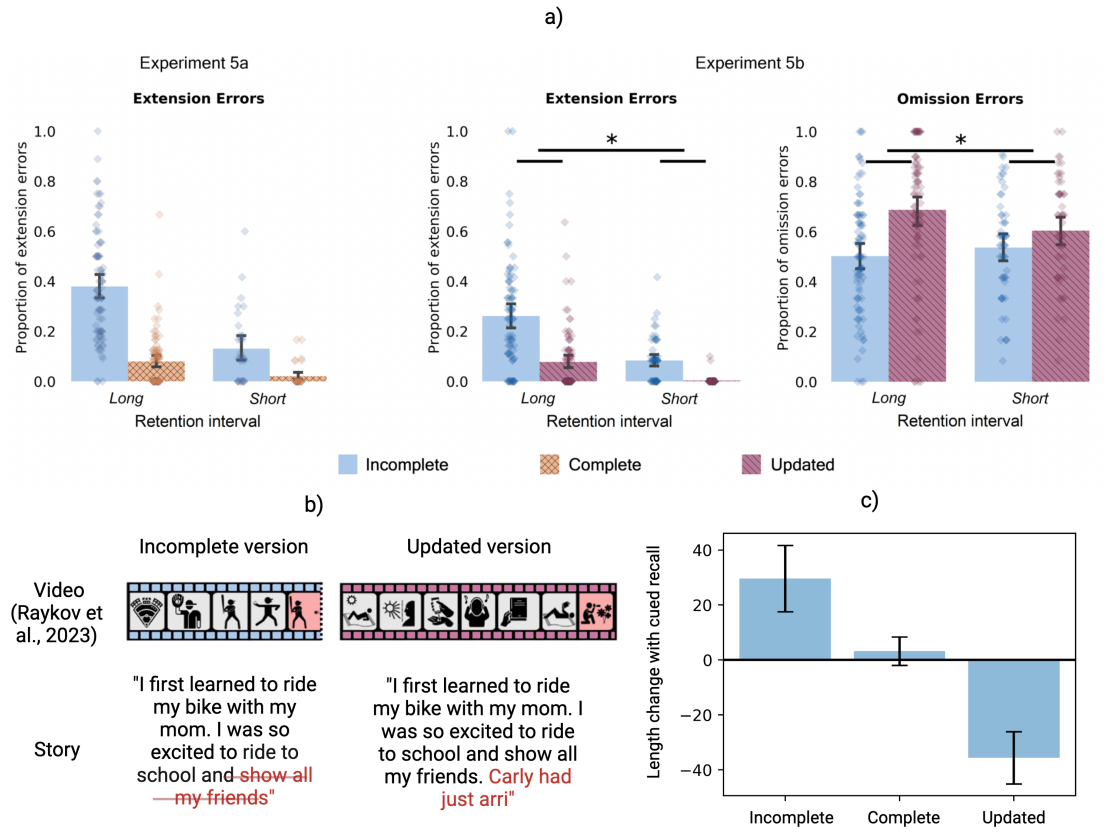

Figure 3.10: a) Figure reproduced from (Raykov et al., 2023). b) Comparison of the video task in (Raykov et al., 2023) and the text narrative analogue. c) The mean difference between the length of the recalled and encoded stories, in the incomplete, complete, and updated conditions. The incomplete stories are recalled as longer than the original text, whereas the updated stories are recalled as shorter.

the 'cd PARENT_OF ef' edge was omitted from the graph, the test would be the model's continuation from 'cd PARENT_OF'. (Beam search with five beams was used to generate predictions.)

Testing involves two stages, retrieval followed by generation: first the hippocampus is queried for relevant traces, simply by finding sequences containing the node in the query. Then the generative network produces an output conditioned on the retrieved sequence concatenated with the sequence for the task (see Figure 3.11 for examples).

The results show that this supports structural inference immediately after encoding sequences in the hippocampus, whereas relying on either the hippocampal network or generative network alone gives worse results (Figure 3.11).

This is not quite the same as the extended model for sequences, as it is still assumed that hippocampal sequences are stored veridically. However see the Discussion for how these ideas could be developed to store 'conceptual gists' (together with unexpected details) in the hippocampus and use retrieval augmented generation to reconstruct memory based on these gists.

3.4 Discussion

In this chapter I have extended the model of memory construction and consolidation in Chapter Two to sequential stimuli by changing the generative network to an autoregressive sequence model (Radford et al., 2019), allowing a wider range of phenomena to be explored. However the idea of consolidation as self-supervised learning (i.e. learning to reconstruct patterns) during replay from a hippocampal 'teacher' is the same.

The resulting generative network exhibits a number of capabilities in addition to the memorisation of 'replayed' sequences. In particular, the generative network supports statistical learning of transition probabilities (Section 3.3.1), inferring new relationships from limited observations by learning an implicit structure (Section

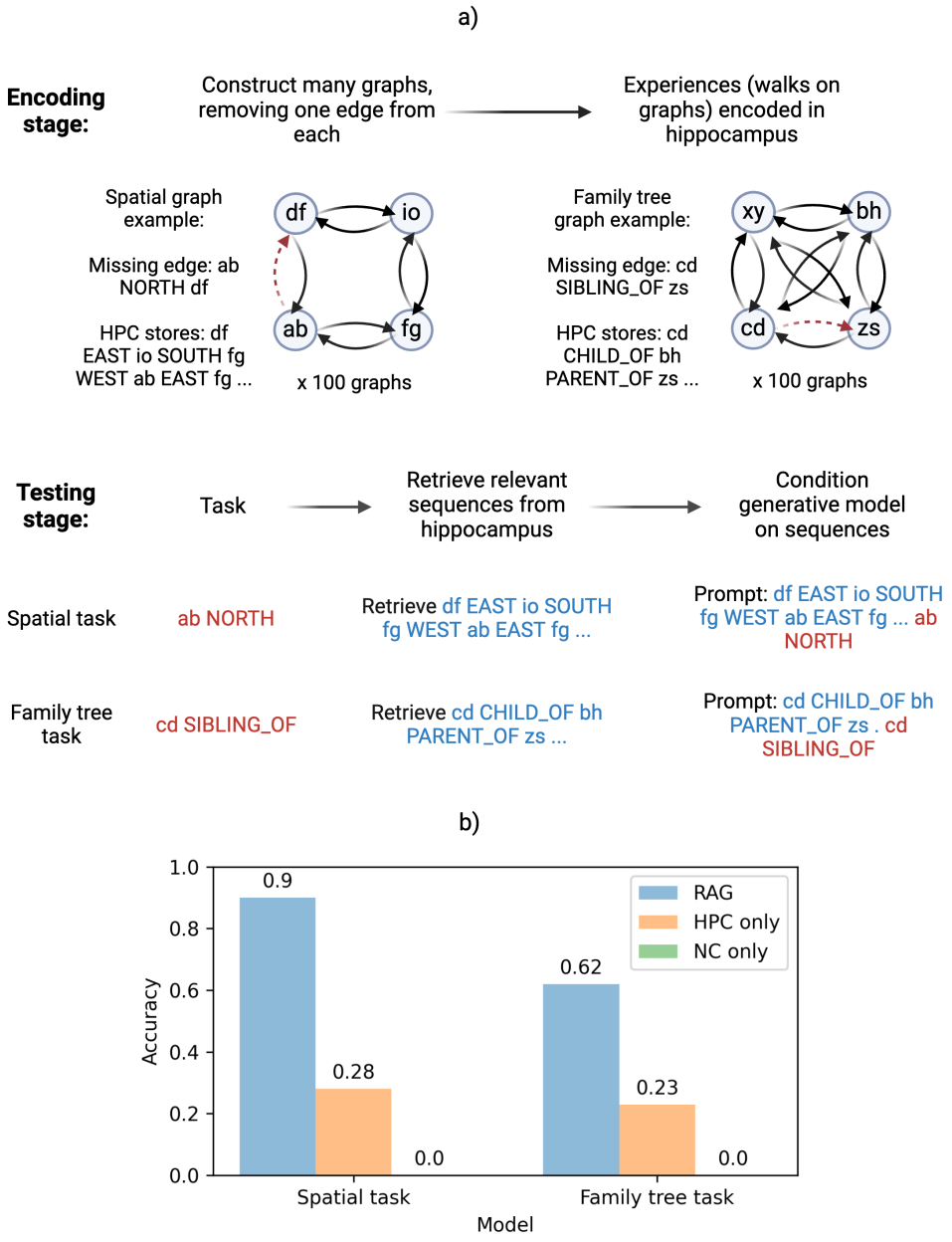

Figure 3.11: Retrieval augmented generation and inference. a) The procedure for the simulations, repeated for the spatial and family tree tasks. In each case, the system encounters 100 new graphs (with a simplified structure), each missing a single edge. A sequence of observations from each graph is then encoded in the hippocampus (so that 100 sequences are stored). The task is then for the system to infer the missing edge, e.g. predict the next location after 'ab NORTH' or 'cd SIBLING_OF' as shown, prior to any consolidation of the new graph. First the system retrieves relevant sequences, and then uses these to condition the generative network. The models trained in Section 3.3.2 are used. b) Results and a 'hippocampus only' and 'neocortex only' baseline for comparison. The 'hippocampus only' baseline randomly selects one of the locations / people in the retrieved sequence. The 'neocortex only' baseline conditions the generative network on the task alone, without retrieving sequences from the hippocampus.

3.3.2), and model-based planning (Section 3.3.3). The computational approach taken is applicable to any sequence of symbols, meaning that linguistic and non-linguistic sequences can be modelled in a consistent way: I also demonstrated how distortions arise in narratives, and how these reflect priors in the generative model derived from previous experience (Sections 3.3.4 and 3.3.5). Gist-based distortions in narratives arguably reflect the claim of Zwaan and Radvansky (1998) that a story is 'a set of processing instructions on how to construct a mental representation of the described situation', and it is this situation model that is remembered rather than the text itself.

The static extended model suggests how memories are encoded as a combination of sensory and conceptual features, and how the generative model contributes to recall from encoding onwards. The results so far are the sequential equivalent of the 'basic' model in Chapter 2, i.e. the simplifying assumption is made that sequences in the hippocampus are encoded 'in full' even if they are well predicted. This raises the question of what constitutes the extended model for sequences. As with the static extended model, this is required i) because storing veridical sequences is inefficient, ii) to explain why we observe distortions immediately after encoding (rather than only after some consolidation has occurred), and iii) to account for the presence of abstract conceptual representations in the hippocampus (Quiroga, 2012).

The simple demonstration of retrieval augmented generation in Section 3.3.6 shed some light on this question, showing how hippocampal and generative networks could jointly draw novel inferences, with the generative network conditioned on retrieved hippocampal sequences. This could also explain gist-based distortions to sequences that are observed even prior to consolidation (Bartlett, 1932; Raykov et al., 2023). But this is only one aspect of the extended sequential model, as it is still assumed hippocampal sequences are stored veridically. Further work could explore ways to store a compact vector representation of each experienced sequence in the hippocampus together with unexpected elements. However intermediate layer representations in GPT-2 are not directly comparable to the latent variable representations in a VAE, since they are less compact, as there is no 'bottleneck' in GPT-2, and do not support

sampling.

The principle of the extended sequential model can also be demonstrated with narratives by using the generative network to generate a 'gist' for each 'experience', and then storing the gist with unexpected elements in the model hippocampus. The 'full' memories could then be reconstructed from the gist through retrieval augmented generation, i.e. by conditioning the generation of the output on the gist retrieved from the hippocampus. (See Figure 3.12 for a summary of this proposal, and Section B.2 of the Appendix for further details and examples.) However what a non-linguistic gist or summary would involve is unclear. In addition, further thought is required regarding the connection between 'gists' and compact vector representations of sequences, and which of these options best reflects the conceptual component of sequential memory in the hippocampus.

There are several other directions for future research. Firstly, I treat the stimuli in the simulations as though they are already segmented, and do not address how continuous experience is discretized into events (Newtson, 1973; Newtson & Engquist, 1976). The relationship between event segmentation theory (Zacks et al., 2007), the Structured Event Memory (Franklin et al., 2020) account, and this model should be explored, as prediction error according to the generative network could be used to segment the sequence. One complexity is that event segmentation occurs at multiple levels of granularity; this could potentially involve rolling means of prediction error over different time periods, or alternatively variable error thresholds (so that more fine-grained segmentation occurs when the error exceeds a lower threshold, and more coarse-grained segmentation a higher threshold). However how these segments of varying lengths would be stored is unclear.

A more fundamental issue is how to learn cross-boundary transition probabilities at all. Whilst cross-boundary transitions are remembered less well (DuBrow & Davachi, 2013), they are not forgotten completely, which is what would happen if no sequence in the hippocampal network captured the transition. The hippocampal network is thought to preferentially encode unfamiliar (i.e. high prediction error) experiences (Hasselmo et al., 1996), such as an unexpected transition to a new context, and

reconciling this with the mechanism for event segmentation is a challenge.

Secondly, much more could be done to explore the connections between language, memory, and imagination. As demonstrated in this chapter, advances in generative models allow certain psychological phenomena involving language to be modelled more fully than was previously possible. A number of mental health symptoms and conditions are thought to relate to narratives, e.g. rumination (Nejad et al., 2013), delusion (Coltheart et al., 2011), and confabulation (Kopelman, 2010), so this may be fruitful from a computational psychiatry perspective. (I revisit rumination in Chapter Four.)

Thirdly, this is primarily a model of psychological, rather than neural, data. Much more work could be done to bridge the gap between the ideas in this chapter and the growing understanding of sequence representations at a neural level. In particular, the sequential model allows the consideration of navigation, which is associated with distinctive cell types such as grid cells, so the connection between these different levels of explanation should be explored.

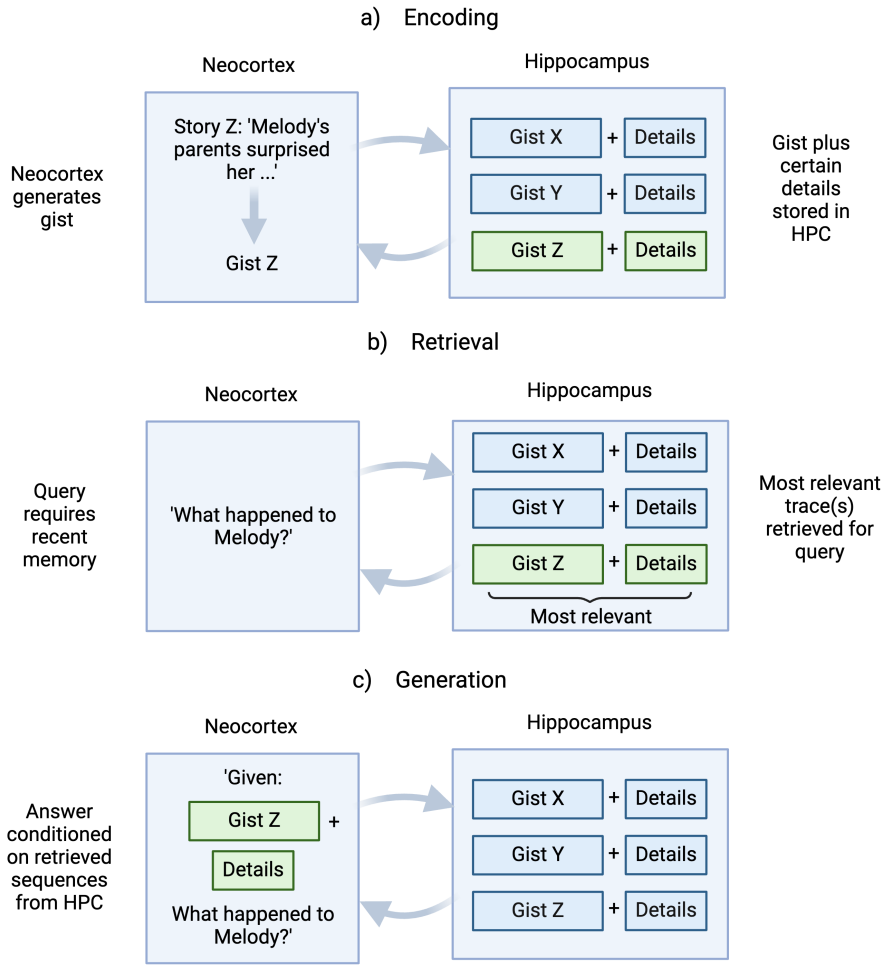

Figure 3.12: A simplified model of hippocampal-neocortical interactions as retrieval augmented generation with stories. a) The neocortex generates a gist for the narrative, which is encoded in the hippocampus together with unpredictable details (not simulated here). b) Given a query as the input to recall, the neocortex searches the hippocampus for relevant traces, e.g. by finding nearby embeddings in this simulation. c) The generative network in neocortex produces an 'answer' conditioned on the retrieved hippocampal trace(s). See Section B.2 of the Appendix for further details and examples.

Type	Original	Recalled
Incomplete	I first learned to ride my bike with my mom. I was so excited to ride to school an	I first learned to ride my bike with my mom. I was so excited to ride to school. My mom was so proud of me.
	Jean had a job interview for a position she really wanted. The interviewer said the company would call the next day. That day, Jean discovered her phone had no dial tone. She frantically searched and f	Jean had a job interview for a position she really wanted. The interviewer said the company would call the next day. That day, Anna was at work. She got the call and was so excited. The interviewer had no idea Anna was so excited!
Complete	Tom had a very short temper. One day a guest made him very angry. He punched a hole in the wall of his house. Tom's guest became afraid and left quickly. Tom sat on his couch filled with regret about his actions.	Tom had a very short temper. One day a guest made him very angry. He punched a hole in the wall of his house. The guest got scared and left quickly. Tom regretted it very much that night.
	Jim got a tattoo when he was younger. It was an impulsive decision. After a while Jim really didn't like it. He decided to get it removed. It was costly but worthwhile.	Jim got a tattoo when he was younger. It was an impulsive decision. After a while, Jeff really didn't like it. He decided to get it removed. Jeff was very embarrassed by the decision.
Updated	Max's car was old. It needed new tires desperately. He bought a set from the discount tire store. They worked very well for a month. Then the rubber came apart and they went flat! Rita was so excited she couldn't sleep. That night her Mom took her camping. They had a wonderful time.	Max's car was old. It needed new tires desperately. He bought a set from the discount tire store. They worked very well for a month. Then the rubber came apart and they went flat!
	Sarah was tired of seeing spiders in her room. So she decided to use special bug spray. But later, she read that birds are a better solution. So she built a birdhouse to attract birds to her backyard. And she didn't see spiders in her room anymore. The performer was ready to go! She walked out on stage and heard the applause and cheers. The stage	Sarah was tired of seeing spiders in her room. So she decided to use special bug spray. But later, she read that birds are a better solution. So she built a birdhouse to attract birds to her backyard. And she didn't see spiders in her room anymore.

Table 3.5: Event extension and contraction examples.

Chapter 4

Consolidation and continual learning

4.1 Introduction

Catastrophic forgetting or interference occurs when newly learned information overwrites previously learned information in a neural network, leading to poor performance on the previously learned tasks (McCloskey & Cohen, 1989). Continual learning is the ability to learn, or memorise, a series of tasks, or items, sequentially without the occurrence of catastrophic forgetting (Hadsell et al., 2020). During consolidation, it is thought that hippocampal memory traces 'teach' predictive models of the world new information (Chapter Two), but how this process avoids destroying previous learning is not fully understood.

Beyond memory consolidation, this is a general problem for connectionist models of the brain (French, 1999). The training data for machine learning problems is carefully curated, e.g. to balance the number of examples of different categories, and shuffled before training. Reality is far messier, so biological learning must be robust to changing distributions over time. For example, it might be important for survival for a single highly salient event to be remembered, even in the absence of

reminders and the presence of conflicting data. The field of lifelong learning tries to make machine learning capable of continual learning over long timescales in more naturalistic contexts.

This problem is closely linked to the stability-plasticity dilemma (Carpenter & Grossberg, 1987). Networks in the brain must be *plastic* enough to learn from a single event, but *stable* enough for existing knowledge to be preserved. Whilst the complementary learning systems account (CLS; McClelland et al., 1995) made progress on this problem, explaining how a single event can be gradually assimilated into neocortex, there is still the issue of how old knowledge can be retained in the absence of hippocampal traces (in CLS all memories are retained to train the neocortical network).

Building on the account of memory consolidation and construction presented so far, this chapter discusses the following points:

1. How the brain avoids catastrophic forgetting - i.e. achieves only gradual forgetting - in memory consolidation is unclear. If the neocortical 'world model' is trained only on new memories, old memories degrade faster in neural network simulations than observed in reality. This is a general issue for all accounts of memory consolidation, including classic views like CLS, which is only solved in machine learning through the careful interleaving of training data (Norman et al., 2005).

2. Learning from self-generated data ('generative replay') during the integration of new knowledge into a generative model alleviates catastrophic forgetting. Whilst generative replay can be used to reduce catastrophic forgetting in a separate classifier (Shin et al., 2017), it can also enable continual learning in the generative model itself (Sun et al., 2019; Van de Ven & Tolias, 2018). However too much generative replay may lead to model degradation (Alemohammad et al., 2023).

3. Many variables affect the results of this approach; I demonstrate a few of these factors with causal language models trained on paths in consecutive

environments. In particular, I explore different ways of sampling from the generative model, and show that a small number of self-generated examples can suffice to stabilise old knowledge.

4. Learning from self-generated experience can aid inference and generalisation, preventing overfitting to real experience as well as catastrophic forgetting (Hoel, 2021; Kurth-Nelson et al., 2023). This gives clues to how generative models contribute to offline learning. However, the model can also degrade as a result of self-generated learning (Alemohammad et al., 2023; Shumailov et al., 2023), with the pathological effects of rumination simulated as one example of this.

5. The stages of sleep may be optimised to interleave new and old knowledge, enabling continual learning (Hoel, 2021; Norman et al., 2005; Singh et al., 2022). Although this is more speculative, I explore the effect of variables such as the number of sleep cycles and ratio of NREM to REM 'sleep'.

I focus on exploring these issues in a simplified spatial task, because much of the literature in this area relates to rodent spatial cognition. This makes use of the sequential model proposed in Chapter Three.

4.1.1 The problem of catastrophic forgetting

Most research on catastrophic forgetting focuses on classification tasks, where in each phase the distribution of the training data changes in some way. Suppose a classifier learns to categorise an input as one of four classes. First it is trained on classes A and B (phase one), distinguishing between them with high accuracy. If it is then trained on classes C and D only (phase two), the performance on classes A and B is likely to deteriorate. This is an example of catastrophic forgetting in a class-incremental learning task (Van de Ven & Tolias, 2018), but this issue can be found in many settings.

Continual learning also has clear relevance to generative models. The previous chapters propose that consolidation involves training a generative model of experience on replayed hippocampal memories. This model needs to retain its ability to represent

a certain type of event even when it has not been experienced for months or years. Without this, the ability to recall or imagine remote events would be lost over time.

Computational models of memory ought to display gradual forgetting, but catastrophic forgetting is a more dramatic decline which is not observed in healthy adult human memory. (However, Darby and Sloutsky (2015) suggest that in young children retroactive interference effects appear similar to catastrophic interference.) Gradual forgetting could reflect degradation of the memory trace due to either 'intrinsic' causes, or 'extrinsic' interference from new memories, whereas catastrophic forgetting is due to interference.

One might expect catastrophic forgetting to be alleviated if unexpected memories are stored preferentially, as in the model presented so far - when a category of memory begins to decay, the generative network's prediction error for that category increases, meaning that examples are stored. With consolidation these memories refresh the category's representation in the generative network. However this does not fully solve the problem, as some memories are of rare or unique categories, for which 'reminders' are experienced rarely or not at all.

Note that we are less concerned with the problem of catastrophic forgetting in classification tasks in this chapter. In Chapter Two I suggest that semantic memory could be supported by projections from latent variable representations, and show that lightweight models can be trained on relatively few examples to decode latents into categories. Learning multi-purpose representations arguably makes continual learning more feasible; if only a few parameters are task-dependent, it is more reasonable to assume that the brain could learn new 'classifiers' for new categories. (For example, logistic regression classifiers on top of CLIP image embeddings are highly competitive with deep neural networks trained on particular tasks (Radford et al., 2021), making it possible to have many task-specific classifiers without adding many parameters.) Catastrophic forgetting in category learning is therefore a more tractable problem than catastrophic forgetting in the 'world model' itself.

4.1.2 Catastrophic forgetting and consolidation

CLS makes progress on the stability-plasticity dilemma with two networks that learn at different rates, the more stable cortex learning gradually from repeated replay, following one-shot encoding in the more plastic hippocampus. If the environment is mostly stationary with occasional surprises, CLS works very well, preserving knowledge of the environment in the cortex and also integrating the rare events. As Norman et al. (2005) put it, the hippocampus serves as a 'training trial multiplier'. However, as the authors point out, CLS and subsequent models run into problems when the statistics of the environment change so that there are no longer hippocampal 'reminders'.

Norman et al. (2005) illustrate this by considering the effect on the concept of birds when a person moves to Antarctica, if CLS were true with respect to this issue. Initially, memory of typical birds, and facts such as 'most birds can fly', would persist thanks to hippocampal replay of typical bird memories. But over time, the likelihood of typical bird memories being replayed relative to penguin memories would become lower and lower, if only penguins were observed in waking life (ignoring other Antarctic bird species for the sake of argument). Eventually the cortex would only be exposed to penguins, whether awake or asleep, and knowledge of typical birds would suffer catastrophic forgetting.

Norman et al. (2005) argue that 'the CLS model can be supplemented by a new kind of off-line learning where cortex and hippocampus separately rehearse stored memories, thereby repairing damage to these memories' (Introduction), and that this occurs during REM sleep. Singh et al. (2022) develop this proposal further; in their model, during NREM sleep the hippocampal network replays its attractor states to the cortex, whereas during REM sleep the cortex autonomously activates its own attractors due to network oscillations.

To support this view Norman et al. (2005) review evidence that 'SWS may support hippocampal replay of new memories to cortex, and REM may support tuning of pre-existing cortical and hippocampal representations' (Data section). This includes

findings that hippocampus and neocortex are highly synchronised during NREM but not REM sleep, suggesting that REM sleep involves hippocampus-independent processes. (But on the other hand, hippocampal damage leads to dreams in REM sleep lacking in vividness and detail (Spanò et al., 2020), in line with the many findings discussed in Chapter One implicating HF in all 'event generation'.)

A few other papers in the neuroscience literature connect consolidation and/or hippocampal replay with continual learning. For example, González et al. (2020) propose that sleep may be a mechanism for reducing catastrophic forgetting of old memories due to interference from new memories, specifically by 'combining consolidation of new memory traces with reconsolidation of old memory traces to minimize interference' (Abstract).

Káli and Dayan (2004) show that experience replay is required to avoid catastrophic forgetting in their model of consolidation. They find that without reactivating memories stored in the hippocampus, consolidated episodic memories are easily forgotten due to cortical plasticity. Replay 'allows access to episodes stored in the hippocampus to be maintained, by keeping them in appropriate register with changing neocortical representations' (Káli & Dayan, 2004, Abstract).

Their model is a restricted Boltzmann machine (RBM; see Section 1.6.2). The input layer of the RBM represents the sensory neocortex (SNC), and the hidden layer represents the medial temporal neocortex (MTNC). The hippocampus (HPC) is connected to MTNC, and stores patterns of MTNC activity. In recall, partial or corrupted activity in SNC propagates to MTNC, activating the most similar stored memory in HPC, with MTNC activity then propagating back to a complete version of the memory in SNC.

The data used in the simulations are binary patterns, with 8,000 possible patterns overall. During semantic pre-training, the RBM is trained on a large number of random patterns to learn their statistics (but none are encoded in HPC at this stage). Next, the network encodes a smaller number of patterns sequentially, this time adjusting the weights of the RBM but also storing the resulting MTNC pattern

in the HPC. Káli and Dayan (2004) alternate these experience-initiated events with replay-initiated events to achieve successful 'consolidation'.

Experience-initiated learning works as follows: Activity in SNC first spreads to MTNC. As the signal bounces back and forth, the weights are adjusted using the contrastive divergence algorithm (see Section 1.6.2). To encode an episode, the pattern in MTNC is written to HPC. Replay-initiated learning works as follows: A stored pattern in HPC is randomly activated, producing activity in MTNC which spreads to SNC. As the signal bounces back and forth, the weights are again adjusted using the contrastive divergence algorithm, but with the layers 'flipped'. In other words, the simulation involves the training of an RBM as usual, but with an extra stage (offline learning from replay) that keeps the 'meaning' of the hidden variables more stable.

Káli and Dayan (2004) consider what happens when replay is inactivated: they find that even if an episode 'remained perfectly stored in the hippocampus throughout ... neocortical learning came to erase the route to recall', because 'continued semantic learning after the storage of the episode caused its MTNC representation to move away from the version with which the stored hippocampal trace was associated' (Index Maintenance section). In other words, the authors show that without memories being replayed, the network's representations drift away from their initial 'meaning' at the time of encoding. The stored memories become out of sync with MTNC representations, leading to rapid forgetting. Crucially, this applies to consolidated memories too.

My approach to this problem differs in several ways. Firstly, I do not assume that events can be stored indefinitely in the hippocampus. Secondly, Káli and Dayan (2004) model hippocampal traces as latent representations, whereas I model them as sensory features (in the basic model) or a combination of latent representations and sensory features elements (in the extended model). This avoids representation drift to the same extent. Thirdly, I explore continual learning in the neocortical model, rather than how the ability to retrieve hippocampal traces is maintained.

4.1.3 Continual learning techniques

An obvious strategy to avoid catastrophic forgetting is to interleave examples from the different classes in the training data, e.g. by shuffling together a balanced dataset of examples. But this requires retaining all the training data, which is not biologically plausible. In the machine learning literature, and especially in reinforcement learning, the hippocampus is often thought of as a 'memory buffer' to reduce catastrophic forgetting (e.g. Roscow et al., 2021). As described above, whilst this may be true over short timescales, it does not solve the lifelong learning problem, since the hippocampus does not store all its memory traces indefinitely (even if it is always required for episodic recall).

However, multiple approaches have been suggested for how to avoid catastrophic forgetting *without* keeping a copy of all training data. These mainly fall into 'regularisation-based', 'expansion-based' and 'rehearsal-based' categories (Van de Ven & Tolias, 2018). Elastic weight consolidation (Kirkpatrick et al., 2017) is the most well-known regularisation-based approach. It aims to prevent the subset of weights that are most important for a task from changing too much in subsequent learning. Another way to get different weights to 'specialise' in different tasks is by adding new weights during learning; this is the expansion-based category. For example, Rao et al. (2019) propose a 'dynamic expansion approach in which capacity is added as needed' for new tasks. However this may not be scalable as the number of tasks grows.

Early approaches that attempt to 'remind' the network of the original data fall into the rehearsal-based category. The 'pseudo-rehearsal' strategy (Robins, 1995) involves giving a classifier random inputs at the end of a phase, recording its outputs, and then interleaving these examples during the next phase. These approaches use a form of teacher-student learning, as they train the 'student' model in phase two on inputs labelled by the 'teacher' model in phase one, in order to maintain the behaviour in phase one. Unfortunately this becomes less successful as tasks get more complex, as the random inputs are not representative enough to capture the model's behaviour in their outputs.

4.1.4 Generative replay in machine learning

Generative replay (Shin et al., 2017; Van de Ven et al., 2020; Van de Ven & Tolias, 2018) is a more recent variant of the rehearsal-based category. When this approach is applied to classification tasks, a generative model learns to generate representative examples of the classifier's training data, including those from previous phases of training. These are then labelled by the classifier at the end of a phase, and mixed into the training data in the next phase. If the generative model suffers from catastrophic forgetting, generative replay just moves the problem elsewhere, so how might continual learning in the generative model itself be achieved? The generative model can be trained on its own self-generated data, as Shin et al. (2017) propose. Note that this is connected to the concept of self-distillation (Furlanello et al., 2018), which I revisit later.

The following example provides some intuition for the effect of generative replay on the generative model itself: consider a generative model trained on images of different animals consecutively. If in phase one the model was trained on cats, and then in phase two on dogs, it would lose the ability to (re)construct cats. But if self-generated cats were mixed in with dogs in phase two, memory of both classes would be preserved.

Van de Ven and Tolias (2018) combine a classifier and generative model in a single network, thus avoiding the training of two separate models, in order to make generative replay more efficient. Specifically they use a VAE with an additional classification layer in the middle, trained on the sum of a classification loss and the usual VAE loss. It learns from both experience and replay, the latter initiated by sampling from the latent variables then decoding them into generated examples. (The authors observe that generative replay with distillation - where the classifier is trained not just on the predicted labels of the previous model for generated inputs, but the predicted scores - performs better than plain generative replay. So for the replayed data, the classification loss is replaced by a distillation loss.) Van de Ven et al. (2020) develop this model further with various refinements, such as the use of a different prior in the VAE to enable class-conditional generation. They contrast their approach

with both the 'standard' view of the hippocampus as a 'memory buffer' replaying veridical memories to neocortex, and the view of the hippocampus as a standalone generator.

I build on the idea of training generative models on self-generated data (Shin et al., 2017; Van de Ven et al., 2020; Van de Ven & Tolias, 2018), but there are several differences from previous work. Firstly, consolidation is not a primary focus in these existing studies, so they do not explore how 'standard' hippocampal replay of new memories could be combined with generative replay. Shin et al. (2017) and Van de Ven et al. (2020) imply all replay is generative and new knowledge is instead acquired with online learning from direct experience, which is arguably implausible for the reasons McClelland et al. (1995) explain, whereas I suggest recent memories are intermixed, or interleaved, with generated events. Secondly, Van de Ven et al. (2020) and Van de Ven and Tolias (2018) use a specific VAE variant architecture and test continual learning performance on an image classification task, whereas I explore continual learning primarily in autoregressive sequence models, and test tasks performed by the generative model itself. These tasks capture the ability to remember an environment and the ability to draw novel inferences about it.

More recent work shows that generative replay can be applied to many types of generative model. LAMOL (Sun et al., 2019) — 'LAnguage MOdeling for Lifelong language learning' - is an approach to continual learning for large language models (LLMs). The authors show that replaying 'pseudo-samples' of old tasks whilst a new task is learned reduces catastrophic forgetting, outperforming rival continual learning techniques. It does even better when samples are conditioned on a task-specific token, ensuring that there are sufficient reminders for each previously learned task; without this, the authors note that samples of more remote tasks become infrequent over time. More recent research explores other ways to control the sample generation of LLMs to optimise continual learning; for example, Maekawa et al. (2023) use a model 'hippocampus' that stores the beginning of a subset of items from the training data. The LLM produces the generated samples conditioned on these hippocampal inputs, fusing the rehearsal-based and generative replay approaches.

4.1.5 Self-generated training data in machine learning

Generative models can be trained on their own generated data as they learn new tasks, reducing catastrophic forgetting in the generative model, but this can come with drawbacks.

Alemohammad et al. (2023) argue that generative models trained on their own data, in what the authors call an 'autophagous ("self-consuming") loop', can degenerate over time. They show that with fully synthetic data 'the quality (precision) or the diversity (recall) of the generative models decreases over generations', however 'with enough fresh real data, the quality and diversity of the generative models do not degrade' (Contributions section). Similarly, Shumailov et al. (2023) find that 'model collapse is universal among generative models that recursively train on data generated by previous generations' (Theoretical Intuition section), where model collapse is a process whereby a generative model produces an increasingly limited set of outputs.

In other words, with insufficient real data, training on self-generated data reinforces the generative model's statistical biases, leading to forgetting of the true data distribution. If the quality is too low, errors compound over time, e.g. Alemohammad et al. (2023) show that image generation artefacts become more pronounced in an 'autophagous' generative adversarial network. But if the diversity is too low, the model 'collapses' to just a few examples. This concept of 'model collapse' is related to the concept of 'runaway consolidation' (Norman et al., 2005), where rehearsal of the strongest memories leads to a vicious circle of forgetting.

Such research might lead to scepticism about generative replay improving generalisation, even if it can support continual learning. On the other hand, offline learning from a model of the environment has been used extensively in reinforcement learning with great success. For example, the DYNA architecture (Sutton, 1991) is a model-based reinforcement learning (RL) algorithm in which an agent learns a model of the environment through experience, which captures the transition probabilities between states. As in conventional RL, the agent learns a value function or policy directly from real interactions with the environment; Q-learning is used in Sutton (1991) but

the same approach can be applied more broadly. However, in addition, the DYNA agent uses its learned model of the environment to simulate experiences. During the planning phase, the agent 'imagines' taking actions in various states and observes the predicted outcomes according to its model. These simulated experiences are used to update the agent's value function or policy. The DYNA architecture is particularly powerful when actual interactions with the environment are limited. See also Ha and Schmidhuber (2018) for more recent work that implements the world model as a VAE trained on video game footage.

Similarly, data augmentation is widely used in training classification models, particularly for images; whilst augmentations were previously rules-based (e.g. resizing existing images) they now often involve the creation of variant images with diffusion models (Trabucco et al., 2023). Examples from reinforcement learning and classification tasks demonstrate that generative models can improve the generalisation abilities of other models. But the success of augmenting a *separate* model's training data with 'imagined' examples from a generative model does not necessarily imply that a generative model can learn from its *own* outputs.

Also relevant are 'Born-Again Networks' (Furlanello et al., 2018), in which self-distillation is used to train a student with an identical architecture to the teacher on the teacher's outputs. Distillation traditionally involves a larger teacher model training a smaller student, but Furlanello et al. (2018) use a succession of identical models. The process starts by training a 'first-generation' model on a given classification task. This model's outputs (the distributions of scores rather than the final labels) are used to train a 'second-generation' model. This process can be repeated for multiple generations, with each student becoming the teacher for the next generation. Furlanello et al. (2018) show that student models can outperform their teachers, displaying better generalisation despite having access to no additional 'real' data.

4.1.6 Generative replay in the brain

In neuroscience, the term generative replay is sometimes used in a broad sense to refer to training on self-generated data (Van de Ven et al., 2020), which is the definition adopted here, and sometimes used in a narrow sense to refer to certain kinds of novel sequence in the hippocampus (Schwartenbeck et al., 2021). It is worth expanding on the connection between these concepts.

As Kurth-Nelson et al. (2023) summarise, there is plenty of evidence that the sequences of place cells firing in hippocampal replay do not always reflect real trajectories, but can join together separate subsequences (Gupta et al., 2010), 'diffuse' throughout an open environment (Stella et al., 2019), or traverse regions that have never been visited in reality (Ólafsdóttir et al., 2015; Pfeiffer & Foster, 2015). For example, Gupta et al. (2010) showed that after rats explored a T-shaped maze, the horizontal corridor of the maze was replayed as a single sequence, despite the left and right sides only ever being experienced separately, whilst Ólafsdóttir et al. (2015) showed that rats replayed paths within regions of a maze that they could view but not visit.

There is also evidence from human neuroimaging that sequences (re)played by the hippocampus do not always correspond to real memories (Liu et al., 2019). Participants experienced stimuli in a scrambled order, after learning a pattern according to which the stimuli could be unscrambled. The replayed sequences were unscrambled rather than in the scrambled order experienced by the participants. This has added further support the idea of generative replay as sampling from a model of the world. The 'world model' implied by these findings could be the type of generative model learned through consolidation, as suggested in this thesis, or a more 'neurosymbolic' model where entities are bound to roles to implement compositional reasoning (Kurth-Nelson et al., 2023).

The description of these imagined sequences as generative replay (e.g. Schwartenbeck et al., 2021) suggests that this phenomenon is a variant of 'standard' hippocampal replay. 'Preplay', in which a trajectory is played out by the hippocampus preceding its experience, is another recent addition to this category (Dragoi & Tonegawa, 2011;

Ólafsdóttir et al., 2018), thought to depend on latent codes for space in entorhinal cortex (Bicanski & Burgess, 2018). It is worth noting that observing a novel sequence being (re)activated in the hippocampus does not necessarily imply the sequence is stored there, or that the activity is initiated there. According to the account presented so far, the hippocampal formation is required for many kinds of 'event generation', so arguably the 'generative replay' of novel sequences is more like neocortex-initiated imagination than the reactivation of hippocampal traces in 'standard' hippocampal replay. (However generative replay is a commonly used term in both the machine learning and neuroscience literature, so I use it throughout this chapter.)

Stoianov et al. (2022) propose a computational model of how the hippocampus gives rise to non-veridical sequences (Dragoi & Tonegawa, 2011; Gupta et al., 2010; Liu et al., 2019; Ólafsdóttir et al., 2018; Stella et al., 2019), and of how these contribute to continual learning of spatial environments. They argue that the hippocampal formation is a hierarchical generative model supporting spatial cognition, with items, sequences, and maps corresponding to different levels in the hierarchy. The ability to infer which environment the agent is in, as represented by the final 'map' layer, is used to assess continual learning. As the model learns new environments, it generates samples of previous environments to avoid catastrophic forgetting. This differs from the work in this chapter as it focuses specifically on a hierarchical model of spatial environments, rather than a more general transition model for sequences.

4.1.7 Dreams and continual learning

Whilst most hippocampal replay events occur in NREM sleep, implicating it in memory consolidation (Ego-Stengel & Wilson, 2010; Girardeau et al., 2009), there is less consensus regarding the role of REM sleep. REM sleep is associated with dreaming, which makes it natural to ask whether it could be a mechanism by which self-generated stimuli stabilise the current model of the world.

As described above, Norman et al. (2005) and Singh et al. (2022) argue that REM sleep supports 'stabilisation' of remote memories. Similarly, Walker and Stickgold (2010)

propose that while NREM sleep consolidates new memories, REM sleep 'supports the integration of these and older memories into rich associative networks' (Assimilation section). Furthermore McDevitt et al. (2015) provide experimental evidence that 'the brain can rescue and consolidate memories damaged by interference, and that this process requires REM sleep' (Abstract).

Other models share the view that spontaneous activity during sleep could reactivate weak connections then strengthen them with Hebbian plasticity, saving previous learning or memory from catastrophic forgetting (González et al., 2020; Tadros et al., 2022). As Tadros et al. (2022) describe, 'information about old tasks is not completely lost even when catastrophic forgetting is observed from the performance-level perspective', since traces of information relevant to the old task remain in the synaptic weights, and 'can be resurrected by offline processing' (SRC Algorithm section).

However the connection between REM sleep and declarative memory is controversial. The 'dual process hypothesis' (Marshall & Born, 2007; Smith, 2001) suggests NREM sleep supports the consolidation of declarative memories, whereas REM sleep supports the consolidation of non-declarative memories (such as procedural memory for skills). Meanwhile some research disputes the idea that REM sleep is related to memory at all, e.g. Vertes and Eastman (2000) note that several common antidepressants dramatically suppress REM sleep but memory deficits are not observed. They argue that REM deprivation techniques in experiments linking REM to memory processing often induce stress, which is a confounding factor. See also Siegel (2001).

But more recently, stronger evidence has emerged of the link between REM sleep and consolidation. Noting the limitations of previous studies based on correlation alone, Boyce et al. (2016) used optogenetics to suppress the theta rhythm during REM sleep in mice, without disrupting other aspects of sleep, and found this impaired consolidation.

Other work has also explored the potential functions of dreams. Hoel (2021) suggests that dreams are simulated 'out-of-distribution' events that help the brain avoid

overfitting to real experiences. Hoel (2021) argues that 'the most effective way to trigger dreams about something is to have subjects perform a novel task like Tetris repetitiously ... because the visual system has become overfitted to the task' (Discussion).

Whilst some dreaming does occur in NREM sleep, dreams during REM are thought to be more frequent, longer, richer in detail, and more bizarre (Hobson et al., 2000). Intriguingly, Cavallero et al. (1992) observe that 'semantic knowledge is more frequently mentioned as a dream source' for REM than NREM dreams (Abstract). In addition, (Blagrove et al., 2011) find there is a lag of 5-7 days between experience and REM dreams, but no such lag for NREM dreams, consistent with the idea that dreams originate from consolidated memory. However, some studies dispute the extent of these differences (Oudiette et al., 2012). (As an aside, dreams are reported to become richer and more complex over the course of development (Foulkes, 2009), in fitting with the fact that a generative model trained on more data would generate more varied outputs.)

Deperrois et al. (2022) present an alternative view of the functions of sleep. In their computational model, the authors refer to NREM sleep as 'perturbed dreaming', and REM sleep as 'adversarial dreaming'. In NREM sleep, the hippocampus replays encoded memories (modelled as images), which are then perturbed by the addition of occlusions. The network's weights are adjusted to make the latent codes for the perturbed images more similar to the latent codes of the originals (a common strategy for learning robust representations, as in a noise-reducing autoencoder). REM sleep involves adversarial training: the generator learns to trick the discriminator, while the discriminator learns to distinguish dreams from reality.

It is clear that the brain can generate novel events, and it seems highly likely that these imagined stimuli contribute to learning. But the evidence that REM sleep specifically contributes to learning is weaker. The relationship between generative replay and sleep in the subsequent results is therefore more speculative than the basic idea that generative replay enables continual learning, but I explore it as an interesting hypothesis.

4.1.8 Maladaptive learning from imagination

Learning from imagination, i.e. updating beliefs based on self-generated data, can also have adverse effects. Arguably rumination, delusion, hallucination, and confabulation all involve maladaptive imagination of some kind. Rumination refers to repetitive thinking, often involving episodic recall with a negative emotional bias (Nejad et al., 2013), delusion refers to false beliefs (Coltheart et al., 2011), and hallucination to false perception (Beck & Rector, 2003). Confabulation refers to distortions in episodic memory beyond the typical gist-based distortions discussed in Chapter Three (Kopelman, 2010). These phenomena are thought to be linked, e.g. rumination increases the likelihood of hallucination (Jones & Fernyhough, 2009) and delusion (Carse & Langdon, 2013).

Rumination is particularly relevant to the questions in this chapter. As mentioned above, this term refers to repetitive negative thinking and is associated with multiple mental health conditions including depression and anxiety (Nejad et al., 2013). Rumination is often focused on the past, distinguishing it from more future-oriented 'worry' (Ehring & Watkins, 2008). Specifically rumination involves recalling autobiographical memories with a negative bias, in such a way that negative biases in the memory are further reinforced. (The term 'maladaptive rumination' is sometimes used to distinguish these behaviours from typical rumination on memories and beliefs, which can be adaptive.)

Neuroimaging evidence links rumination to regions in the proposed generative network; this is to be expected given that rumination often involves remembered and imagined autobiographical events, which are known to activate the hippocampal formation as well as association cortex. In an fMRI study, Hamilton et al. (2011) compared activity in the default mode network – the network associated with passive mind wandering in the absence of a specific task – with task-specific activity to quantify 'default mode dominance'. An 'overactive' default mode network, which overlaps with the proposed network for generating autobiographical events, was associated with higher levels of maladaptive rumination in patients with depression.

Recall of autobiographical memories after rumination is negatively biased compared to recall after a distractor task (Lyubomirsky et al., 1998). This is evident in free recall, with more negative memories recalled after rumination, but furthermore distorts judgements. For example, participants recalled the frequency of negative events in their life as higher, and the frequency of positive events as lower, following rumination. Lyubomirsky et al. (1998) note a feedback loop in which 'negative memories may further exacerbate depressed mood through their effects on negative thinking and poor problem solving ... thus feeding a vicious cycle between rumination, mood, and negative thinking' (Introduction).

Mechanistic accounts of rumination are limited, partly because the tools to model narratives computationally are so recent, but there are a handful of relevant studies. Van Vugt et al. (2018) model the 'habits of thought' account of rumination, in which 'patterns of memory associations that are frequently rehearsed can become something like an attractor ... and therefore will be replayed any moment there is time for mind-wandering' (Introduction). (However, this work does not model memories' contents but just their association structure.) In addition, Berg et al. (2022) discuss rumination from the perspective of the active inference framework, and propose that it involves sampling possible policies in a maladaptive manner.

4.2 Methods

4.2.1 Continual learning with sequences

The sequence simulations described here investigate the impact of generative replay on continual learning, focusing on spatial navigation tasks within simple grid environments. I simulate the consolidation of a succession of environments, and test the effect of generative replay on continual learning. Each environment consists of a 3x3 grid, with 9 locations represented by randomly generated nouns, and the trajectories the model is trained on are the shortest paths between points within the environment.

In the pre-training phase, a base model is trained on many such environments, not including the new environments used in the subsequent tasks. This represents the background knowledge in the generative model of the world. In the task, a model is trained on five consecutive environments, with and without generative or experience replay. The experiments are repeated across multiple trials, testing different combinations of training sizes, sample sizes, and temperature parameters to explore their effects on model performance.

Data preparation

As described above, each environment is represented as a grid in which each location (i.e. square) is labelled by a random noun. Routes in the environment can then be represented as sequences of form 'apple EAST pancake NORTH material EAST chair'. (As before, sequences are represented as strings of characters because this makes it straightforward to train GPT-2.) More specifically, the shortest path between two locations can be represented as 'FROM: apple, TO: chair, PATH: apple EAST pancake NORTH material EAST chair'. This enables us to test the ability to infer the shortest path based on a few examples of a new environment.

The paths for each environment are split into training and testing datasets. The code supports three ways of doing this. Firstly, by default a random subset are used for testing and the rest are used for training. Secondly, paths can be sorted by length (in characters), so that shorter paths are used to train the model. This makes the tasks more challenging, as the paths in the testing dataset are less likely to be subsequences of the paths in the training dataset. Thirdly, the twelve horizontal and vertical paths within the grid can be used for training. This provides a small set of paths that are guaranteed to provide sufficient information to solve any problem, if the model is can learn the structure of the tasks adequately.

Simulation procedure

A single generative model is used, which is updated after exposure to each new environment (i.e. it is fine-tuned - the neural network's weights are adjusted from the

existing weights, not initialised from zero). As in the previous chapter, the generative model uses the GPT-2 architecture, an autoregressive sequence model that learns to predict the next item in the sequence.

The starting point is the small (117 million parameter) version of GPT-2, trained by OpenAI on large amounts of text data from the internet (Radford et al., 2019). Whilst the sequences in the task are not human language, they feature words, so the rationale for fine-tuning rather than starting from scratch is that this might accelerate the learning of the task.

In the following experiments the model is first pre-trained, using the approach described in the previous chapter. The pre-trained generative network represents the neocortex, with its existing knowledge of spatial environments, prior to the task. Specifically, 1000 random three-by-three grids are created, and for each 'environment' all shortest paths between all pairs of points are calculated (including multiple paths of equal length where applicable). This gives 140 sequences per grid, so a total of 140,000 sequences for pre-training. The model was trained for five epochs (iterations through the full dataset). However the pre-trained model was not trained on any of the environments used in the subsequent simulations, i.e. new environments were created for the task itself.

Task procedure

The steps of the simulation when using generative replay are as follows:

1. Create five new three-by-three grids of random nouns to represent environments 1 to 5.
2. Fine-tune the base generative model on environment 1 for 10 epochs, using a learning rate of 5e-05 and a batch size of 1.
3. Then for each environment x (where x is from 2 to 5):
    (a) Generate n sequences from the updated generative model with temperature T, as per Section 3.1.1. (The model continues generating sequences up to

a certain number of tokens, but only the first sequence is taken from the output.)

(b) Mix together the generated sequences with 100 real sequences from environment x.

(c) Fine-tune the generative model on this combined dataset (randomly oversampled to 1000 sequences to keep the total amount of training data constant), using a learning rate of 5e-05 and a batch size of 1 for 10 epochs.

(d) Generate the metrics described below for the test data for each of the five environments.

The number of generated samples (n) and the sampling temperature (T) are varied in Figures 4.2 and 4.4 respectively.

Two baselines are tested for comparison. For the first baseline, which exhibits classic catastrophic forgetting, the generative model is trained on only the most recently encoded environment (representing the recent memories stored in the hippocampus and replayed during sleep and rest). For the second baseline I test 'experience replay', which just replays n random samples from the training data for each previous stage of the task.

Evaluation Metrics

Model performance is evaluated using two metrics.

The first is the next location prediction accuracy: given a sequence 'FROM apple, TO: pear, PATH: apple NORTH' can the generative model predict the next location? (Note that the 'TO:' part of the sequence is irrelevant to the task, but is included because the model has been trained on this format of sequence.) After training on each environment, performance is tested on this task across all five environments. For each environment, the model predicts the next locations given partial sequences from the test data (where partial sequences end in directions so that each problem has a single solution), and these are compared with the true next locations. For testing,

the highest probability token was picked at each stage (i.e. greedy decoding), rather than sampling from the probability distribution.

The second is the shortest path accuracy: given a sequence 'FROM apple, TO: pear, PATH:' can the generative model provide a valid shortest path? Since there are multiple such paths of equal length for some pairs of points, the subset of test problems with no solutions in the training data are used. In other words, a good score on this metric requires inferring a novel solution, not just remembering a training example.

I also analyse the generated sequences at each stage of the task to explore the effect of generative replay on learning. Firstly, I calculate the number of unique locations from each environment that feature in the generated sequences. Secondly, I categorise generated sequences into i) 'real' shortest paths that were present in the training data, ii) 'valid' shortest paths that are consistent with the grid environment but were not present in the training data, and iii) 'invalid' sequences. This third category could include sequences that are not consistent with the grid, e.g. a sequence that predicts 'table NORTH chair' when 'chair' is in fact south of 'table', or sequences which are consistent with the grid but are not shortest paths.

4.2.2 Continual learning with images

I also explore whether generative replay reduces catastrophic forgetting with the variational autoencoder (VAE) used in the first chapter.

The MNIST dataset was used to represent the 'old environment', and the inverted MNIST dataset was used to represent the 'new environment'. The latter simply inverts all pixel values in MNIST images so that they display black digits on a white background rather than white digits on a black background.

Before simulating consolidation of the inverted MNIST dataset, a VAE with 20 latent variables was trained for 25 epochs on the MNIST dataset (with early stopping based on the loss enabled). A learning rate of 0.001 and batch size of 32 was used. The architecture of the VAE was the same as in the MNIST simulations in Chapter

Two.

I then tested the performance of three different models: firstly, I tested the VAE prior to any training on the inverted MNIST dataset. Secondly, I fine-tuned the VAE on 1000 inverted MNIST examples (reflecting consolidation of the new 'environment' without generative replay). Thirdly, I generated images (which were representative of the MNIST dataset) using the VAE, by sampling from the standard normal distribution for each of the 20 latent variables, and converting the result into an image with the VAE's decoder. I then mixed together 500 inverted MNIST images with 500 generated images, and fine-tuned the VAE on this randomly shuffled dataset. 10 epochs of training with a learning rate of 0.001 were used in each case. Histograms of the VAE's reconstruction errors for images from inverted MNIST and MNIST datasets were then plotted. Finally, I tested the effect of different numbers of self-generated samples, holding the number of new images constant at 500. Results were averaged across three trials.

4.2.3 Continual learning and sleep

The sleep simulations use the same method as in the sequence simulations described above, but rather than mixing the 'recent memories' and 'generated data' together, they are used in alternating stages. NREM sleep is modelled by replaying sequences from the most recently encoded environment to the model while REM sleep is modelled by sampling from the current generative model, and training the generative model on these samples.

100 epochs are used per simulation in total (an epoch is a complete iteration through the current training data). Epochs per sleep cycle is calculated as total epochs divided by the number of cycles. Some fraction of the epochs per cycle are REM epochs and some fraction are NREM epochs (giving a total of *rem_epochs* and *nrem_epochs* respectively). The base model is first fine-tuned on *train_size* training sequences from environment one, representing the model before consolidation of environment two, then the sleep simulation begins. During each NREM stage, the model is trained

for *nrem_epochs* epochs on *train_size* items (sampled from all possible sequences in environment two). During each REM stage, the model is trained for *rem_epochs* epochs on *train_size* items sampled from the current generative model. The perplexity after each stage is recorded, and at the end of 'sleep' the 'next location prediction' accuracy across the test data for both environment one and environment two is measured.

Note that because the generated sequences are from the current generative model, not the initial one, the distribution of the samples changes over time. (For example, one would expect all the samples to reflect environment one at first, but to reflect envionments one and two during REM sleep after some consolidation of environment two.)

There are many parameters that can be varied in the code in future experiments, including:

1. The number of 'sleep cycles' for which to train the VAE, each cycle consisting of a non-REM phase and a REM phase
2. The fraction of epochs per cycle allocated to the REM phase at the beginning of the training, and at the end of training, with the fraction increasing linearly between these values
3. The number of items to use for training during each phase of each sleep cycle
4. The learning rate (different learning rates could be used for REM and NREM sleep, although I did not explore this)
5. The sampling parameters for generative replay, including the temperature

As described above, the fraction of NREM vs. REM sleep was set by changing the number of epochs (so that the model saw each sample multiple times), rather than the number of samples. However the latter would also be a sensible way to implement this fraction in future work.

4.2.4 Simulating rumination

Rumination can be modelled as the generative network consolidating a memory as usual, but then rehearsing the memory repeatedly, and learning from these outputs according to the procedure described above.

The generative network underwent initial training on the Bartlett (1932) story plus one of three datasets. To explore the effect of the generative network's 'priors' on the effect of of rumination, the simulation was run with three different background data distributions, as in the memory distortion simulations in Chapter Three: the 'Shakespeare' model uses lines from Shakespeare plays, the 'News' model uses the 'AG's News' dataset (Zhang et al., 2015), and the 'Scientific papers' model uses abstracts of papers scraped from PubMed (Cohan et al., 2018).

Two further stages of training followed the initial consolidation of the story. In each stage, the model recalled the Bartlett story, given the sequence 'One night two young men from Egulac' as a prompt. The model was was then trained on the recalled story, plus 5000 items from the background dataset as before. Recall was tested at the end of each stage with a sampling temperature of 0.25.

Word clouds show all words in the recalled story (up to the length of the original Bartlett story). Words that feature in the original story are in grey, whereas new words ('semantic intrusions') are in red.

4.3 Results

4.3.1 Mixing self-generated with new memories

In these simulations I explore consolidation over longer timescales in a constantly changing environment. Suppose an animal explores five environments consecutively, e.g. one in the first week, another in the second, and so on. I make the simplifying assumption that sequences are only stored in the hippocampus for the most recently experienced environment; even if this would not be true after a week, I assume that

fully consolidated sequences are not all stored in the hippocampus indefinitely. (For convenience I do not explicitly simulate the hippocampal network, and the replayed hippocampal memories are simply sampled from a list of sequences.)

The environments are represented as grids in which each location (i.e. square) is labelled by a random noun, and shortest paths are represented as sequences of the form 'FROM: apple, TO: chair, PATH: apple EAST cat NORTH chair' (Figure 4.1). Using shortest paths enables us to test the ability to *infer* the shortest path based on a few examples of a new environment. As in the previous chapter, I use GPT-2 (Radford et al., 2019), an autoregressive model that learns to predict the next item in a sequence, to represent the generative network trained through consolidation. The pre-trained GPT-2 model is further pre-trained on a large set of random three-by-three grids, in order to reflect the generative network's 'background knowledge' at the start of the task. It is then trained on five environments consecutively, with and without generative replay. See Methods for further details.

In the baseline case, the generative model is trained on only the most recently encoded environment at each stage (representing the recent memories stored in the hippocampus and replayed during sleep and rest). As expected, catastrophic forgetting is observed (Figure 4.2a) - performance is only high on the most recently consolidated environment, and decays at an unrealistic rate on the previously learned environment. (It is worth noting that this is a large model that definitely has capacity for many environments to be memorised simultaneously, if they were interleaved. So catastrophic forgetting does not reflect capacity limitations in the network, but instead interference of new with old knowledge.)

Let us now compare this with the effect of generative replay. At each stage of consolidation, 100 samples are replayed from the new environment as before, but in addition n self-generated sequences are added into the training data (where n is 10, 50, or 100 in Figure 4.2b-d respectively). The results show that forgetting of the old environments is more gradual thanks to the generative replay; that is, generative replay alleviates catastrophic forgetting during consolidation into a generative model.

The more self-generated samples are used, the better performance on the old environments' tasks is preserved. Using as few as 10 self-generated samples, intermixed with 100 sequences from the new environment, makes a substantial difference. How little data is needed for either rehearsal or 'pseudo-rehearsal' has been noted in previous work, e.g. Scialom et al. (2022) preserve performance of a large language model on previous tasks during successive fine-tuning by replaying a 'memory buffer' containing only 1% of every previous dataset. As Van de Ven et al. (2020) observe, even a small amount of generative replay is beneficial, since not forgetting is easier than learning from scratch.

I also corroborate the claim that generative replay reduces catastrophic forgetting with the VAE used in the first chapter, showing that this is not specific to one particular type of generative model. To demonstrate this, a VAE was trained on MNIST, and the histogram of reconstruction errors for new images from MNIST and inverted MNIST datasets was plotted. Unsurprisingly errors were low for MNIST and higher for inverted MNIST (Figure 4.3, part a). Then the model was trained on 1000 inverted MNIST images, leading to catastrophic forgetting of MNIST (part b). However, interleaving 500 samples generated by the VAE at the end of MNIST training with 500 new inverted MNIST images alleviates this problem, and error for the MNIST dataset remains low (part c). As above, surprisingly few self-generated samples are required, with a sudden change to the MNIST dataset's reconstruction error once 30-40 samples are used. See Methods for further details.

Further research is needed to explore if the latent representations move more or less in the latent space as a result of generative replay. If they move less between phases, i.e. are more stable, this may be advantageous for systems that store latent codes, as Káli and Dayan (2004) describe. This is because if they are unstable, decoding a latent code that was stored some time ago may produce a meaningless output. In addition, recent research proposes ways to keep representations more stable in generative replay (Caccia et al., 2021). However, whether this would occur the expense of other abilities - e.g. integrating knowledge across old and new memories - is unclear.

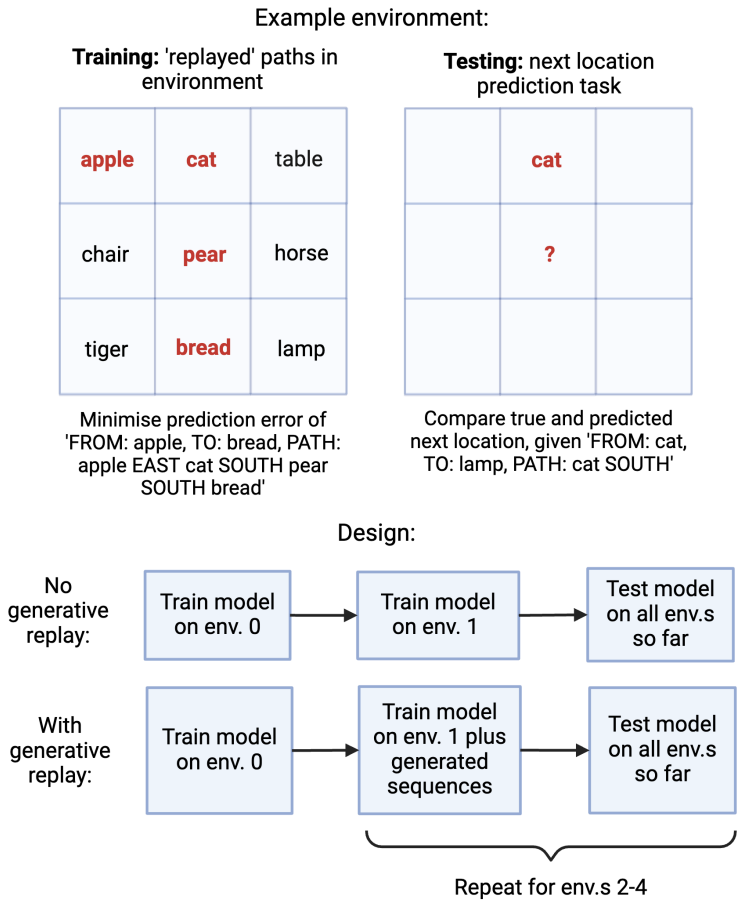

Figure 4.1: Design for the simulations. Each 'environment' is a 3x3 grid with different randomly selected nouns as locations. Sequences from the environment, specifically shortest paths between two points, are used as training data. This is intended to mimic 'standard' hippocampal replay, but I do not simulate the hippocampus explicitly. The main task used for testing is next location prediction given the sequence so far (including the final direction, so that there is a single solution to each problem). In trials without generative replay, five environments are 'consolidated' consecutively. In trials with generative replay, self-generated samples are added to the training data. Note that the samples are drawn from the current generative model, not the initial one.

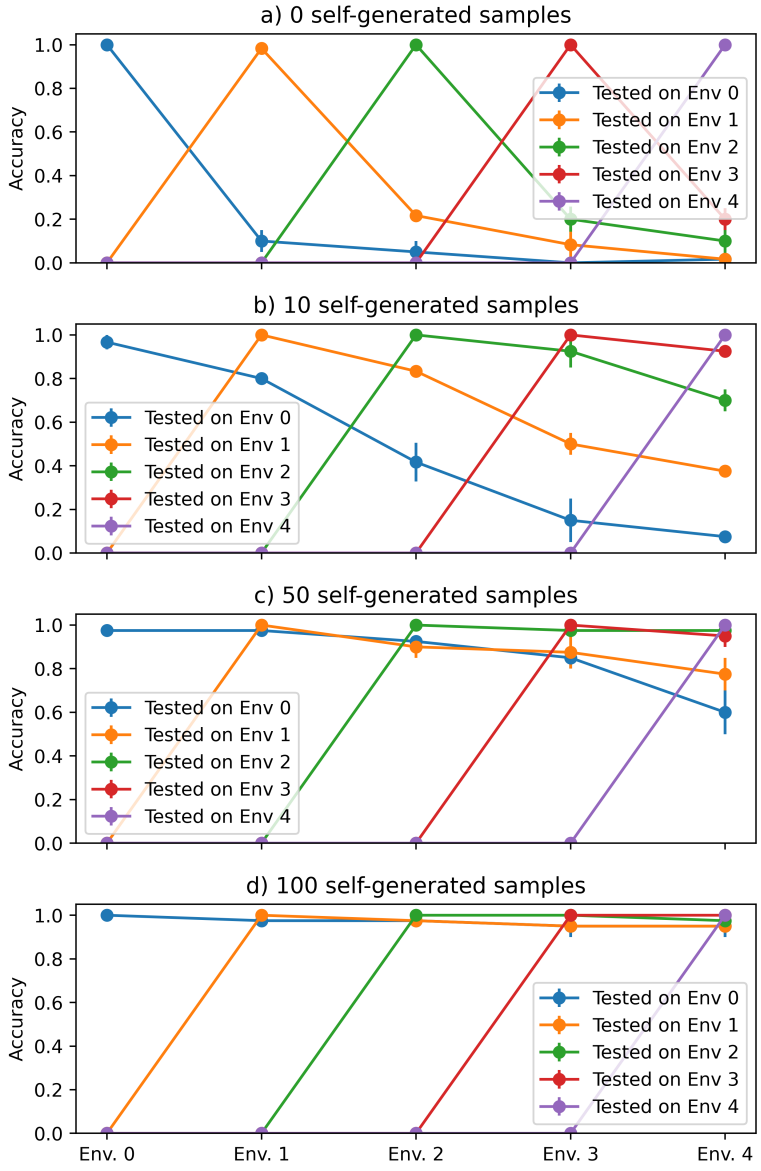

Figure 4.2: The effect of generative replay. a) The rate of forgetting as a sequence model consolidates five environments consecutively. The legend gives the environment the model is tested on. The x-axis gives the most recently consolidated environment. The mean across three trials is shown, and error bars give the standard error of the mean. b-d) At each stage of consolidation, 100 samples are replayed from the new environment as before, but in addition n self-generated sequences are added into the training data.

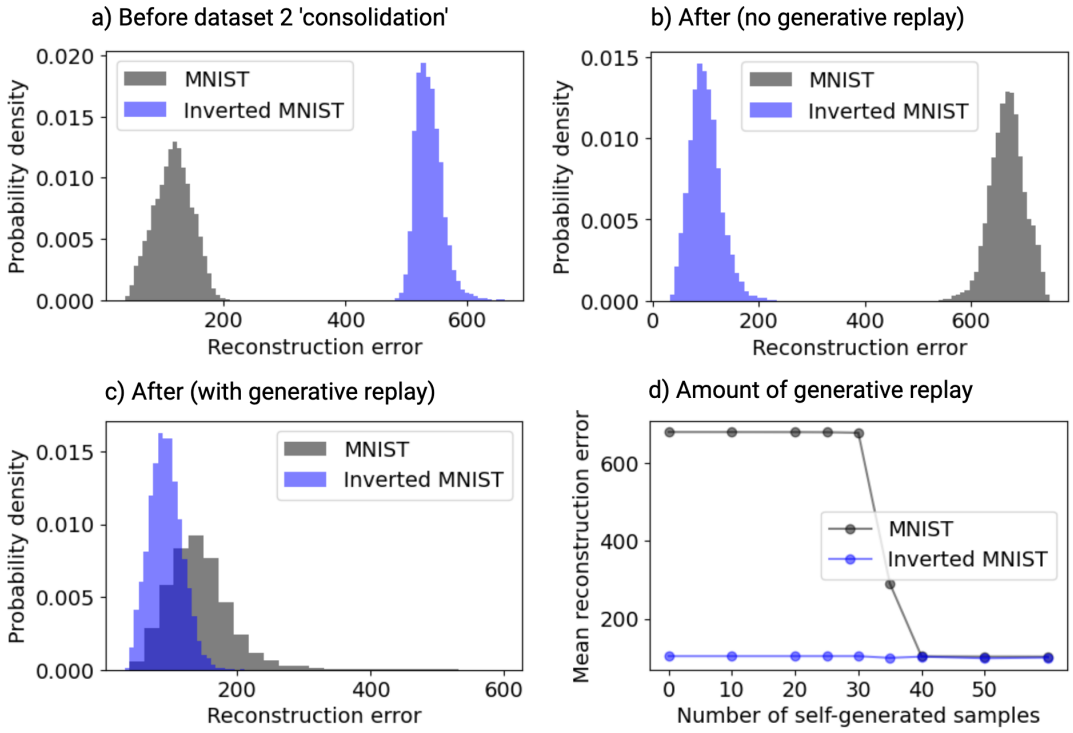

Figure 4.3: Corroborating the effect of generative replay with a variational autoencoder as the generative model. a) The histogram of reconstruction errors for new images for a model trained on the MNIST dataset. As expected, the error is low for MNIST (dataset one) and higher for inverted MNIST (dataset two). b) Results for a model trained on MNIST then inverted MNIST. Catastrophic forgetting of MNIST (indicated by the high reconstruction errors) occurs. c) Results for a model trained on MNIST (phase one) then inverted MNIST (phase two), but with samples generated by the VAE at the end of phase one added to the training data in phase two. d) Mean reconstruction error on the MNIST and inverted MNIST datasets against the number of self-generated samples.

4.3.2 Varying the sampling parameters

Generative replay can produce samples along a spectrum from 'remembered' to 'imagined'. Furthermore 'memories' in the generative model vary in strength; replaying only the strongest consolidated memories produces different results to replaying a a wider variety of weaker memories too. The diversity of the generated samples turns out to be a key factor in the results of generative replay.

For models like GPT-2, this can be manipulated by varying the temperature parameter, which controls the 'sharpness' of the probability distribution from which the next token is drawn. (At a low temperature, already high probability tokens are boosted further, whereas at a high temperature, lower probability tokens are more likely to be picked, producing more 'imaginative' outputs.) Accordingly, I now explore the effect of the 'imaginativeness' of generated samples on continual learning. (There are similar ways to control the variation of data generated using other types of generative model; for example, in a variational autoencoder, sampling from a larger region of latent space would also produce more 'imaginative outputs'.)

Figure 4.4 shows the effect of temperature on generative replay. As the temperature increases (i.e. as the outputs become more varied) knowledge of old environments is initially better preserved, until performance deteriorates again at the highest temperature tested. Figure 4.5 provides an explanation for this. Each plot shows the number of unique locations featuring in the generated data at each stage of training, across a range of temperatures. At lower temperatures, fewer environments are represented in the generated data. For example, at the end of the simulation, no sequences from environment one are generated at a temperature of 0.3, but sequences from all five environments are generated at a temperature of 2.1, together with sequences that do not exist in any environment (which may be either sequences from imagined environments or invalid sequences from learned environments).

The results in Figures 4.4 (summarised in Figure 4.7a) and 4.5 reflect a trade-off between the variety and the quality of the samples. At low temperatures a subset of real memories are generated - the quality of the data is higher but variety is lower.

At high temperatures a wider range of real memories are generated, but also some 'imagined' ones, which may include sequences that are inconsistent with the real environments - the variety of the data is higher but quality is lower. When the quality degrades beyond a certain point, performance deteriorates.

Norman et al. (2005) note that 'an important problem that autonomous rehearsal mechanisms need to solve is runaway consolidation', where 'strong memories are rehearsed more often than weak memories ... [leading] to a positive feedback loop' (Implementation of REM section). In the simulations, a low temperature might lead to only a subset of sequences with the highest probabilities being generated, and whilst this subset of strong memories would be preserved very well weaker memories would be forgotten completely. Figure 4.5b appears to show the 'runaway consolidation' of the first environment at the cost of retaining subsequent environments (since it was first to be consolidated its sequences were generated the most, compounding the issue with each stage of generative replay).

Figure 4.6 breaks down the generated sequences at each temperature into three categories: those that are 'remembered' (i.e. appeared in the training data for the environment), those that are 'imagined' and consistent with the environment, and those that are 'imagined' but inconsistent. For example, if 'apple NORTH table' is part of a sequence but in fact 'chair' is north of 'apple' in the grid, this is inconsistent. Note that this is with only 20 sequences used for training - the number of valid sequences increases with more training data, but the purpose of the figure is to illustrate the trade-off.

In summary, a higher temperature gave better results up to a point (Figure 4.7a), but this could be because more real sequences were generated from remote environments (Figure 4.5). In other words, the results so far do not necessarily imply that 'imagined' rather than 'remembered' sequences are helpful for learning. (Perhaps the best sequences for the model to learn from are veridical memories, but in the trade-off between variation and accuracy, higher variation is worth reduced accuracy.)

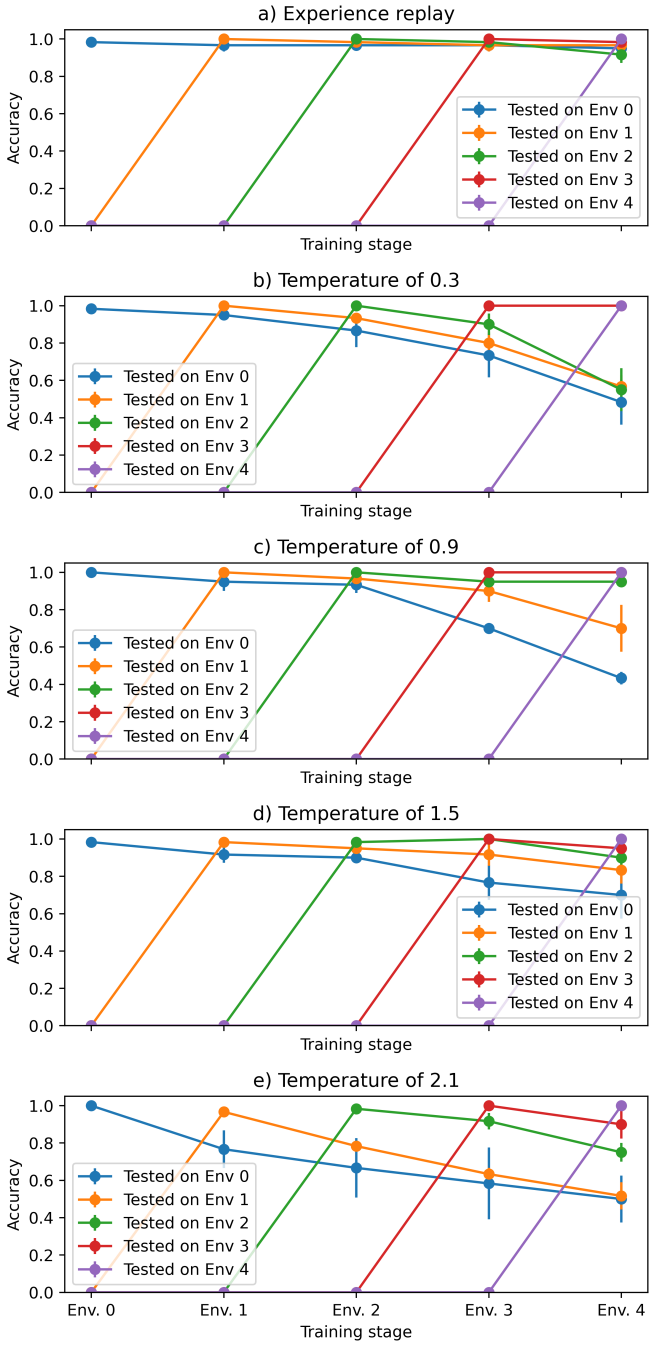

Figure 4.4: The effect of temperature on generative replay. Each plot shows the rate of forgetting as a sequence model consolidates five environments consecutively. At each stage of consolidation, 100 samples are replayed from the new environment as before, but in addition 50 self-generated sequences, sampled using the given temperature, are added into the training data. The legend gives the environment the model is tested on. The x-axis gives the most recently consolidated environment.

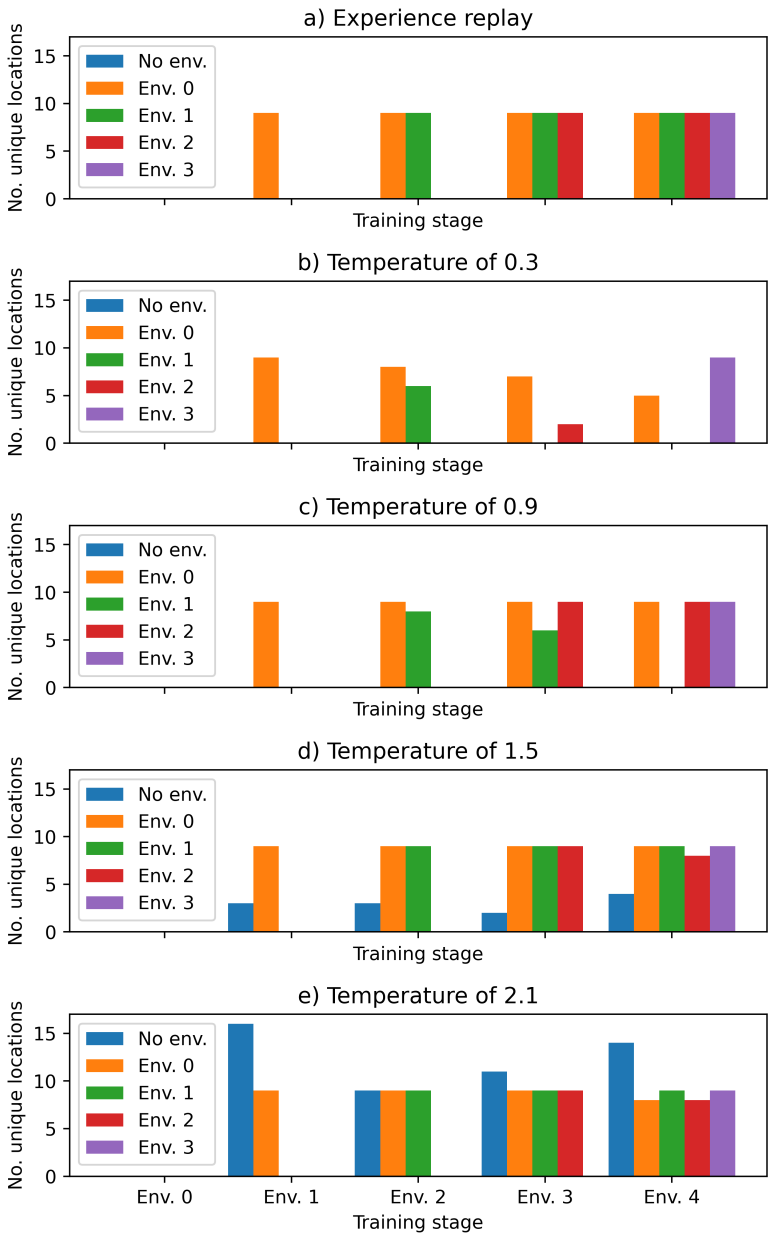

Figure 4.5: The distribution of locations generated at different temperatures. Each plot shows the number of unique locations featuring in the generated data at each stage of training, where the generated data is sampled at the given temperature. At lower temperatures, fewer environments are represented in the generated data. For example, in part a, only locations from the most recent and second most recent environments feature in generative replay, whereas in part c, all environments that have been seen so far feature at each stage.

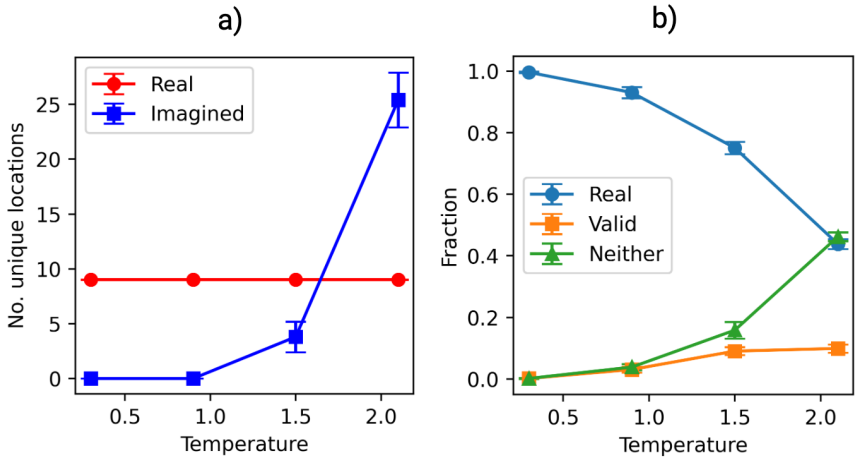

Figure 4.6: Analysing the generated sequences. a) The total number of unique real locations (red) and imagined locations (blue) at each temperature tested, for a model trained on 20 real sequences. Real locations are those that occur in the environment the model was trained on, while imagined locations are those that do not. Note that generated sequences at the end of the first stage of training were used in this analysis, such that there are only nine real locations. b) The number of real sequences (blue), novel but valid sequences (orange), and other sequences (green) in generative replay, using the same model as in part a. Valid sequences are correct shortest paths in the training environment that were *not* included in the training data. Errors bars give the SEM.

4.3.3 Generative replay and generalisation

Whilst the primary focus of this chapter is on generative replay for continual learning, it may also help with another objective: supporting generalisation from a few examples. One might hypothesise that novel sequences which are consistent with the structure of the environment aid generalisation (much like other forms of data augmentation), especially when real memories of the environment are limited. This is a claim that generative replay provides a benefit above and beyond approximating experience replay without the need to store hippocampal memory traces forever.

To investigate this, I compare the case when there is experience replay of real memories of old environments to the case where there is generative replay (keeping the total number of additional 'replays' constant). The rationale for this is that if generative replay provides no benefit beyond experience replay, that suggests generative replay is useful insofar as it reactivates real memories in a far more memory-efficient way.

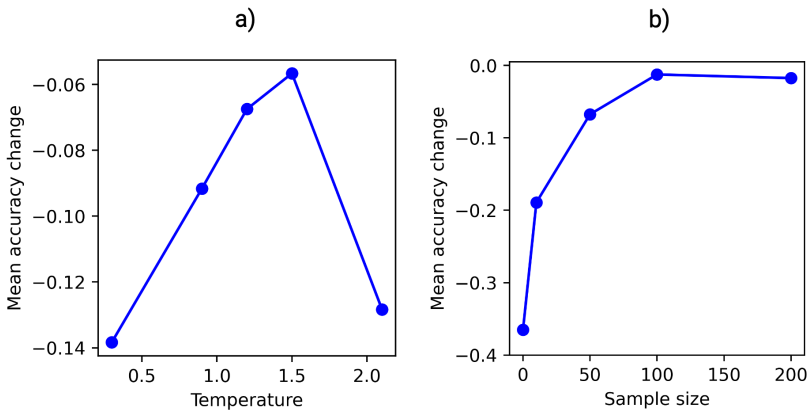

Figure 4.7: Mean accuracy change against temperature and number of generative replay samples (with 100 real sequences from the new environment at each stage). a) The mean accuracy change on the test set for an environment at each stage after its consolidation, for a range of temperatures. The change in 'next location prediction' accuracy between stages n and $n+1$ was calculated for each environment at each stage of training, excluding stages before the environment was consolidated. The average was taken across all of these accuracy differences, across three trials. 50 generative replay samples were used. b) The mean accuracy change on the test set (as described above) for a range of numbers of generative replay samples. A temperature of 1.2 was used.

But if generative replay gives better results, this suggests the 'imagined' rather than 'remembered' sequences are aiding learning.

So far the accuracy has been measured with a simple next location prediction task, however now also I test whether the model can correctly complete the shortest path, which reflects the ability to draw novel inferences from memories. Furthermore, to avoid a 'ceiling effect', rather than training the model on a random subset of n paths, I select the n shortest strings (measured by number of characters). This places further demands on the model to extrapolate beyond the training data.

To avoid the trade-off between quality and variety mentioned previously, I use a different sampling approach to in the previous experiments. Instead of unconditional sampling, which led to repetition of sequences with the same start and end location, generative replay is conditioned on sequences of the form 'FROM: x, TO: y' for all pairs x and y in the previously learned grid. This makes the generated sequences more diverse, without needing to raise the temperature. In this experiment beam search

with five beams (as opposed to sampling) is used to continue the sequence.

Figure 4.8b compares the shortest path and next location accuracy with 100 items of experience replay to 100 items of generative replay, with the generative replay conditioned on pairs of locations as described above. Generative replay does considerably better than experience replay on the shortest path task. This provides tentative support for the view that under certain circumstances, on a task requiring inference, generative replay may benefit generalisation more than an experience replay baseline.

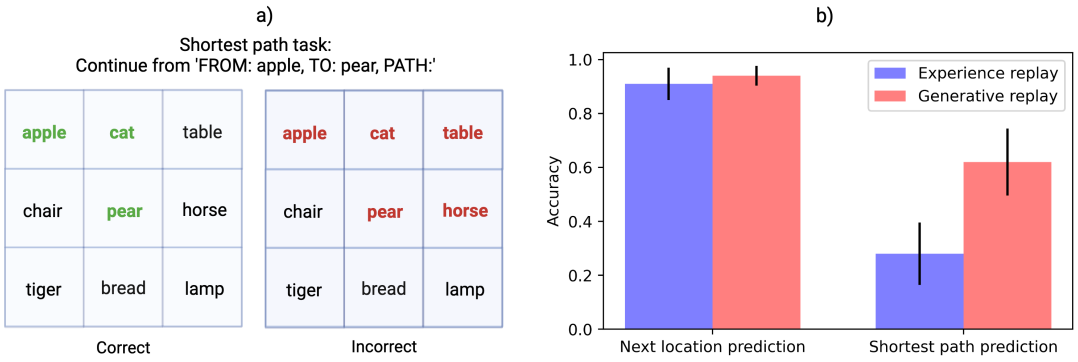

Figure 4.8: Experience vs. conditioned generative replay. a) The shortest path accuracy tests the ability to draw novel inferences from memories. I prompt the model with sequences of the form 'FROM apple, TO: pear, PATH:', and calculate the fraction for which a valid shortest path is generated. I exclude pairs for which one or more routes feature in the training data. b) I compare the next location and shortest path prediction accuracy with 100 items of experience replay to 100 items of generative replay. The simulation involves two successive environments, with models tested on the first environment after training on the second, with either generative or experience replay of the first environment. Generative replay is conditioned on sequences of the form 'FROM: x, TO: y' for all pairs in the previously learned grid, and then greedy decoding (as opposed to sampling) with beam search is used to continue the sequence. The mean across three trials is shown on both the next location and shortest path tasks. Errors bars give the SEM.

4.3.4 Exploring the link to sleep

Papers such as Singh et al. (2022) suggest remote memories are replayed by neocortex in REM and recent memories are replayed by hippocampus in NREM, with this combination preventing catastrophic interference of recent memories with remote

memories. This follows on from papers like Norman et al. (2005), which suggests REM involves 'autonomous memory rehearsal'. Let us now consider how sleep might relate to generative replay, extending the ideas in the papers above to generative models.

The sleep simulations follow the same procedure as in the sequence simulations described above, but rather than mixing the 'recent memories' and 'generated data' together, they are used in alternating stages. NREM sleep involves replaying sequences from the most recently encoded environment to the model (note that whilst this represents hippocampal replay, the hippocampus is not modelled explicitly). REM sleep involves sampling from the current generative model, and training the generative model on these samples. Because the generated sequences are from the current generative model, not the initial one, the distribution of the samples changes over time.

There are many variables that can be manipulated, such as the number of sleep cycles, the temperature for sampling, the learning rates in NREM and REM sleep, the ratio of NREM to REM 'sleep', and the change in this ratio over the course of 'sleep'. Figure 4.9 shows the effect of the number of sleep cycles. More sleep cycles are advantageous in the range tested, but further experiments are required to see if this trend continues. (This is unsurprising - if a block of NREM sleep is too long, remote memories may be forgotten to the extent that they cannot be recovered by REM sleep.) Figure 4.10 shows the effect of the ratio of NREM to REM sleep. This demonstrates that the ratio determines the trade-off between learning the new environment and retaining the old one; the optimal value would be a happy medium between these objectives.

4.3.5 The effect of rumination

Offline rehearsal of memories after consolidation – i.e. ruminating on particular memories – leads to further distortion. Word clouds of the three stages of training for the three models are shown. The word clouds show that the accuracy of recall

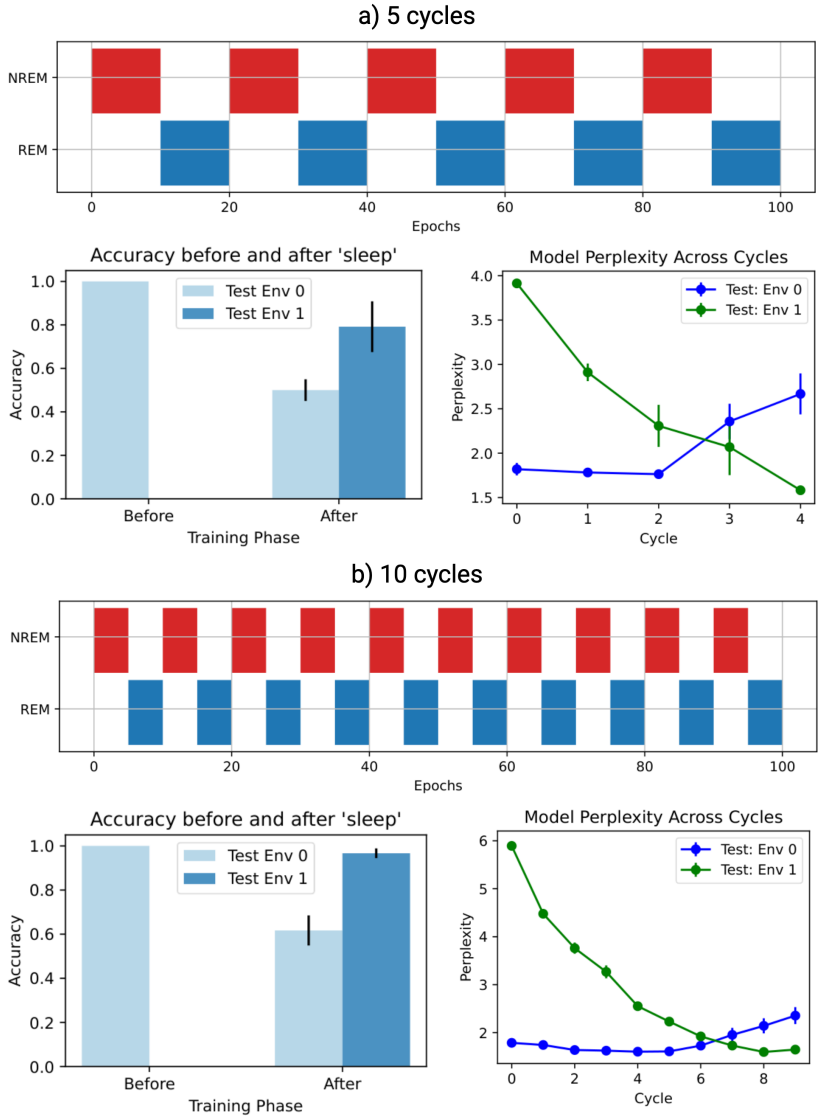

Figure 4.9: The effect of the number of sleep cycles. a) The schedule consisting of five 'sleep cycles' for the simulation, the next location prediction accuracy before and after sleep, and the perplexity per test dataset over time. b) As above but for ten cycles.

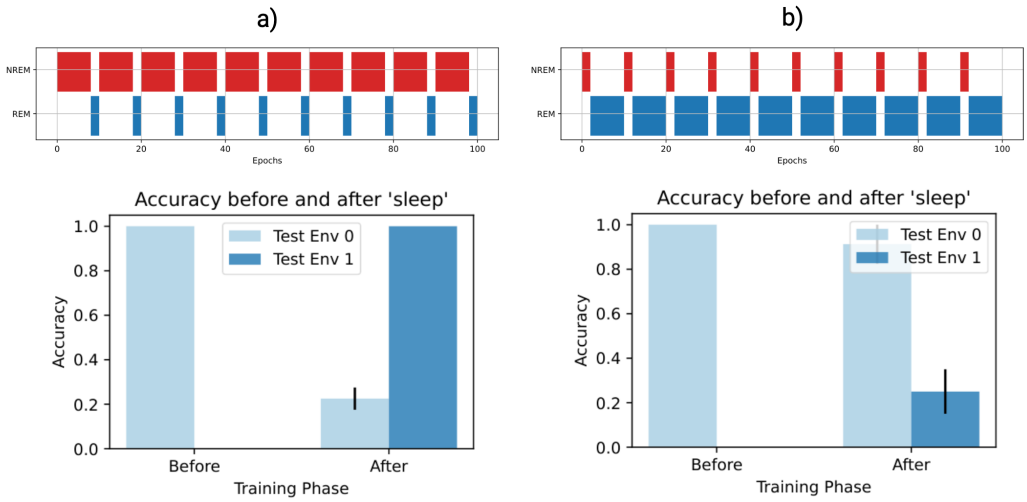

Figure 4.10: The effect of the ratio of REM to NREM sleep. I model REM sleep as training on self-generated sequences, and NREM sleep as training on real sequences from the new environment, representing (veridical) hippocampal replay. The same set of 'memories' are sampled from in each cycle of NREM sleep, whereas new sequences are generated from the current generative model at the start of each phase of REM sleep. a) The schedule consisting of ten 'sleep cycles' for the simulation, each with 20% REM sleep, the next location prediction accuracy before and after sleep, and the perplexity per test dataset over time. b) As above but with 80% REM sleep.

decreases with successive stages of training on self-generated data, and that the semantic intrusions reflect the 'priors' of the model. This is unsurprising given the results in Chapter Two.

This is corroborated by some more qualitative analysis of the stories. Consider the model trained on Shakespeare lines as the background distribution. In stage one, the gist of the memory across the samples is fairly accurate, but there are a few lines like 'My relatives do not know where I have gone ... but I may be killed' suggesting a threat of violence. In stage two, the belief that the people in the canoe are pirates emerges, with several samples containing text such as: 'They thought: "Oh, they must be pirates"'. Then in stage three, this belief strengthens, with other aspects of the story changing to match the belief that the people in the canoe are pirates. For example, one story starts as follows:

> 'One night two young men from Egulac went to the river to hunt seals. The sun rose and they heard war-cries. They thought:"Oh, they must be pirates". They crept up to the shore and hid behind a log. Suddenly a canoe came up to them and out of the canoe came five men armed with bows and arrows. One of the young men said: "What do you think? We wish to take you along. We are going up the river to make war on the people." "I do not wish to go along," the other replied. "Then you must do as I say," the captain said.'

Terms like 'the captain' show how aspects of the story have been distorted in line with the false belief that the canoe is a pirate ship. Similarly, the following lines occur later in the story: 'So they went back to Egulac and the young man told everybody and said: "Behold I accompanied the pirates and we went up the river to make war on the people."'

The model trained on scientific papers displays similar effects. At first small distortions are observed, but by the end of the third stage, lines such as 'Now you may well be thinking:"Oh, that's a canoe full of people." But you would be mistaken' are added to the Bartlett story. Meanwhile, the model trained on news stories initially exhibits

only a few semantic intrusions related to crime and punishment, but after the third stage includes lines like: 'Now this is the law: A man who kills a seal ... shall be punished by death' or 'The court ... in the province of Egulac has pronounced the following verdict:'.

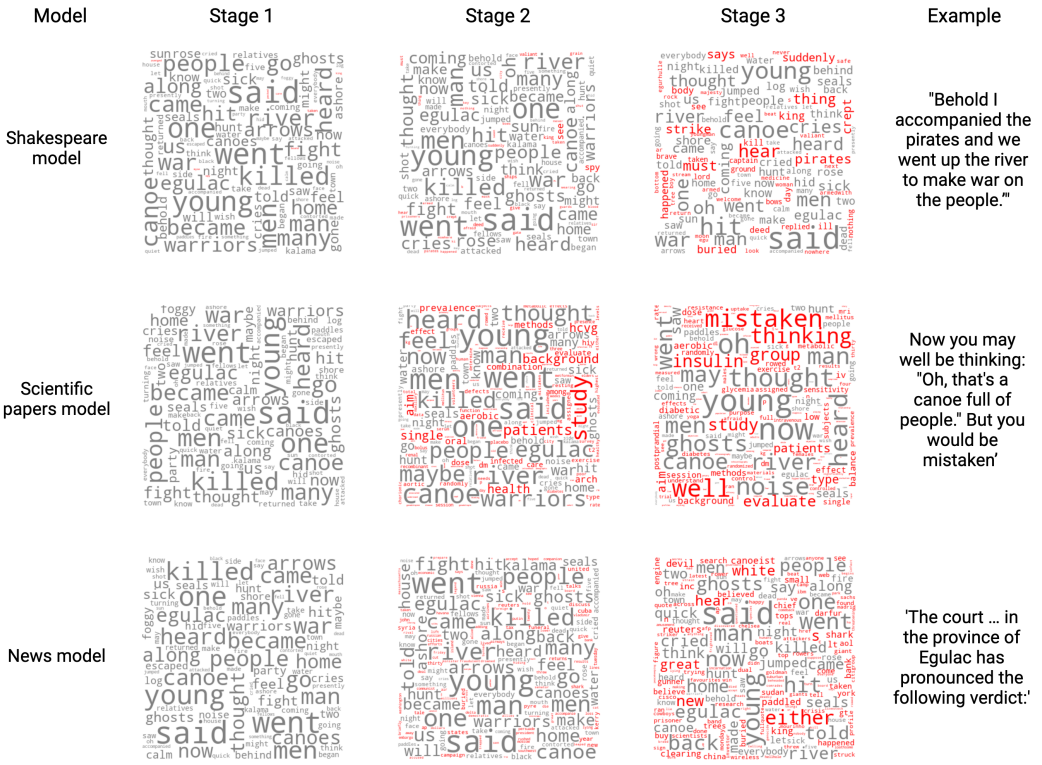

Figure 4.11: Modelling the effect of rumination on recall of narratives. Word clouds showing the effect of rumination on recall of narratives, for each of three 'background datasets'. From left to right, the three stages of training are shown. Stage one is the initial consolidation, and stages two and three involve training on the recalled story.

4.4 Discussion

In a memory system in which new memories in the hippocampus are consolidated into a generative model, how can old knowledge in the generative model be preserved, without retaining memories indefinitely in the hippocampus? This chapter tries

to connect several strands of research relevant to this question, and proposes that training the generative model on its own 'imagined' sequences during consolidation could help to integrate new and old memory without catastrophic forgetting.

In the neuroscience literature, there are a few studies of how the brain might address catastrophic interference during consolidation by rehearsing memories in neocortex (e.g. Norman et al., 2005; Singh et al., 2022). In the machine learning literature, there are many studies of generative replay as an approach to continual learning, including some where generative models are trained on their own samples (Shin et al., 2017; Van de Ven & Tolias, 2018). A subset of these studies have explored generative replay from a neuroscience perspective, most notably Van de Ven et al. (2020), but with limited attention to consolidation. As this thesis models the neocortex as a generative network that can produce events as well as learning from them, this framework is well suited to simulate the role of generative replay in consolidation.

I considered the consolidation of consecutive spatial environments, using the model of sequential memory construction and consolidation described in Chapter Three. The results show that generative replay reduces catastrophic forgetting in the model. This can be effective even with relatively little self-generated data; it is easier to not forget an environment than to learn it in the first place. But the effect is sensitive to how the generative model is sampled from. Varying the temperature allowed us to see how the distribution of generated examples affects learning in the simulations. A higher temperature appeared to be beneficial because a wider range of locations were visited in the generated sequences, i.e. as the temperature increased (up to a certain point), generative replay better approximated experience replay. More speculatively, I showed that alternating blocks of veridical hippocampal replay and generative replay, representing NREM and REM sleep respectively, also prevented catastrophic forgetting, with results depending on the number of 'sleep cycles' and fraction of NREM vs. REM sleep.

In the experiments, generative replay did not just reactivate experienced sequences, but also novel schema-congruent ones, e.g. a valid shortest path between a new pair of points in a learned environment, or even a path in an entirely imagined environment. I

tried to distinguish between two views: Firstly, one could hypothesise that generative replay is more effective than experience replay, as it introduces more variation to the generative model's training data and prevents overfitting to real experiences (Hoel, 2021; Kurth-Nelson et al., 2023). Secondly, one could hypothesise that generative replay is only helpful insofar as it approximates experience replay without the need to store the original data. In a 'shortest path prediction' task requiring novel inferences to be drawn across memories, generative replay gave better results than experience replay (but only when sequences were conditioned in a certain way to maximise variation whilst ensuring quality, and when the number and complexity of real memories was limited).

Other results complicate the picture of generative replay as beneficial for learning. I modelled rumination as maladaptive offline learning from self-generated data, which could be thought of as a human analogue of the 'model autophagy' described in Section 4.1.5, in which negative beliefs 'self-reinforce' over time. Under certain conditions learning from self-generated data is highly adaptive, stabilising old knowledge during new learning, and perhaps even helping the brain to generalise from limited experience. But the results show that under different conditions, learning from self-generated data reinforces negative beliefs, leading to a vicious cycle in which statistical biases in the generative model increase over time.

These results raise several interesting issues. They show that generative replay can beat experience replay in certain scenarios, but this only scratches the surface of a broader question about the feasibility of offline learning from imagination. Generative models can improve other models' ability to generalise, as seen in classification and reinforcement learning tasks, but that involves two separate models, with the larger generative model teaching the smaller classifier about its 'world model'. In other words, the generative model is providing an external source of information, so it is unsurprising the smaller model improves. We cannot conclude that a generative model can augment its *own* training data in the same way, so under what conditions these models can 'self-improve' through offline learning is an open question.

Some research suggests that generative networks inevitably deteriorate when trained

on their own outputs, with the quality or variety of their samples decreasing as their statistical biases are reinforced (Alemohammad et al., 2023; Shumailov et al., 2023). On the other hand, if the system encompasses a generative model and a memory store, it is clear that offline learning can be effective (as demonstrated by the brain). So if a generative model is capable of memorising examples from the training data, as LLMs are, one might argue that offline learning on self-generated outputs should be possible in this case. However, as we have seen, even when generative models memorise their examples 'semantic distortions' are introduced, which may undermine this view.

One might hypothesise that there is a threshold for model performance below which subsequent generative replay is unhelpful - in other words, if the environment or task has not been learned well generative replay will just compound the errors. Conversely, if the model already understands the environment or task well, or at least has memorised relevant experiences, learning from its own 'imagination' can be more helpful. One might expect that the better the 'world model', the more 'imaginative' sequences can be whilst still benefitting learning; further research could therefore explore whether there is an interaction between model performance and the optimal temperature. (Alternatively, one might hypothesise that generative replay only helps learning in networks that can memorise their training data with sufficient reliability, as otherwise error compounding is too serious a problem.)

The fact that training on too much self-generated data can lead to model degeneration is a challenge for certain models in which the hippocampus only encodes latent representations (Benna & Fusi, 2021; Káli & Dayan, 2004). In such cases the full memory must be decoded from stored latent representations, and training on these generated outputs might inevitably reinforce statistical biases within the model. I suggest that memory traces combining different levels of abstraction would be sufficient to alleviate this problem.

One more speculative suggestion I explore in this chapter is that dreams implement generative replay. This is obviously a stronger claim than suggesting that some kind of offline learning from self-generated data takes place, since there are several other

phenomena this could correspond to. The 'dreams aid generalisation' (Hoel, 2021) and 'dreams promote continual learning' (Norman et al., 2005; Singh et al., 2022) hypotheses are not mutually exclusive – dreams as (one variety of) self-generated training data for a predictive 'world model' could explain both of these benefits. However there are limitations to the modelling of sleep which make it hard to draw clear conclusions about the optimal 'schedule' of sleep stages for integrating old and new knowledge. In particular, neural networks learn better from interleaved than from blocked training, whereas humans can learn well from both (Flesch et al., 2018; Flesch et al., 2023); some of the sleep results, e.g. better performance with more cycles, might just display a limitation of connectionist models rather than useful predictions about sleep.

Despite this, much more work could be undertaken on sleep and dreams using the framework. For example, there is intriguing evidence that the nature of dreams varies throughout sleep, e.g. dreams in REM sleep are more vivid and bizarre than those in NREM sleep. As mentioned above, there may be a correlation between the ideal level of 'creativity' and model performance; one might expect less 'imaginative' sampling (a lower temperature) to be optimal early on in learning, but more 'imaginative' sampling (a higher temperature) to be optimal later on. Further experiments could cycle between temperature values rather than using a fixed value. In addition, there are developmental changes in dreaming that could be explored. It is striking that dreams become richer and more complex in early childhood (Foulkes, 2009) as the neocortical 'world model' matures, and that this coincides with the end of infantile amnesia, which could potentially reflect catastrophic forgetting (Darby & Sloutsky, 2015).

There are many other directions for future research. Firstly, hippocampal forgetting in the simulation could be refined to be more realistic. Here I tested a very simplified scenario in which all traces of environment n were gone from the 'hippocampus' when environment $n+1$ was encountered, but really consolidation is gradual. I suggest events in an environment would be consolidated at different rates, with each sequence 'marked for deletion' in the hippocampus once prediction error is low enough. A more

sophisticated simulation could 'delete' hippocampal traces individually rather than the entire environment in one go. One would then observe a smoother transition from experience to generative replay for each environment.

Secondly, solving the tasks in this chapter required only one environment at a time, but the experiment could be adjusted to explore more complex tasks, in which inferences must be drawn based on memories from multiple environments. For example, if the last 'column' of one grid environment was the first 'column' of another, finding the shortest route between environments would require the integration of memories from different stages. One might expect that this kind of task would particularly benefit from generative replay (perhaps more so than experience replay).

Thirdly, generative replay is likely to be one of several mechanisms for avoiding catastrophic forgetting during consolidation. Future work could explore how generative replay could be combined with other mechanisms in the context of consolidation. For example, could particularly salient remote memories in the neocortical network be protected by elastic weight consolidation (Kirkpatrick et al., 2017)? Or might there be some targeted expansion of neocortical networks if the statistics of experience change drastically (Rao et al., 2019)? In addition, these alternative mechanisms could be compared to generative replay. One might speculate that the relative performance would depend on how similar successive environments are, and whether the task requires integrating knowledge across environments; expansion-based or regularisation-based approaches might do less well when information must be synthesised, as they tend to 'partition' tasks in different areas of the network.

Fourthly, the experiments could be repeated with larger models. It may be that the generative models used here, while clearly demonstrating the utility of generative replay for continual learning, are too basic to fully demonstrate the role of imagination in learning. There are plenty of other phenomena that 'emerge' as generative models get larger, so perhaps the ability to do more complex offline learning based on self-generated thoughts is one of them.

Finally, more complex ways of sampling from the generative model could be explored,

which might improve the quality and diversity of imagined sequences. For example, beam search constructs several sequences in parallel before selecting the one with the highest overall probability (see Methods), and its effect could be tested more systematically. Techniques that guide generative models from the GPT family to generate a certain kind of sequence may also be relevant to modelling generative replay (Dathathri et al., 2019; Ouyang et al., 2022). In addition, given the positive impact of conditioning generative replay on certain subsequences in the generalisation results, this could be investigated further.

In summary, I have shown that generative replay supports continual learning in the model, stabilising old knowledge as new memories are consolidated into the neocortical generative network. I also found evidence that generative replay is helpful for generalization under certain conditions. However, in simulations of rumination generative replay was found to compound errors, introducing more 'semantic intrusions' over time. Given this complex picture, more research should explore the conditions under which self-generated data is beneficial or harmful for learning in brains and machines.

Chapter 5

Discussion

Brains need to make predictions to survive, and to achieve this must extract statistical structure from experience. Generative neural networks provide a mechanism for learning to do this by 'prediction error' minimisation. In this thesis I explored how memories may be replayed over the course of systems consolidation to train a predictive model of the world, which supports multiple cognitive functions. These include episodic memory, semantic memory, imagination, and inference. This provides a more mechanistic account of the theory that episodic memories are reconstructions that are influenced by our beliefs, i.e. recall involves 'predicting' the past.

In Chapter Two I presented a computational model in which episodic memories are initially encoded in the hippocampus, then replayed to train a neocortical generative network to (re)construct sensory experiences via latent variable representations. Using images, I simulated how this network can *reconstruct* scenes from partial inputs according to learned schemas (which produces gist-based distortions) and *construct* novel scenes consistent with those schemas. That is, through consolidation we 'learn to imagine'. The generative model could correspond to a network including the hippocampal formation and association cortex which is implicated in many kinds of event generation, including imagination, dreaming, and day-dreaming. (Furthermore, multiple generative networks could be trained concurrently from a single autoassociative

network, with different networks optimised for different tasks.)

Even right after an experience, remembering involves imagining the past based on concepts, combining some stored details with our expectations about what happened. I also showed how unique and predictable elements of memories could be stored and reconstructed by efficiently combining both hippocampal and neocortical systems, optimising the use of limited hippocampal storage.

Experiences and our memories thereof are sequences, not snapshots; in Chapter Three I extended the ideas above to the construction and consolidation of sequential memory, using networks trained to predict the next item in a sequence during replay. I applied this model to statistical learning, relational inference, and planning tasks, and considered distortions in our memories of narratives and events. Recent work on large language models suggests how parametric and non-parametric memory can be combined using 'retrieval augmented generation', in which sequence generation is conditioned on relevant 'memories'. I also explored this as a potential model for hippocampal-neocortical interaction during recall.

Having considered learning *to* imagine, Chapter Four explored some effects of learning *from* imagination. I explored how predictive models of the world trained through consolidation avoid catastrophic forgetting, and suggested that learning from self-generated events (i.e. generative replay) could stabilize old knowledge as new knowledge is assimilated into neocortical networks. One might think that the value of generative replay is just to approximate experience replay without the need to store memories in the hippocampus forever. However, I showed that under certain conditions generative replay may help generalisation as well as continual learning, with imagined sequences extrapolating beyond limited experience to enable novel inferences. On the other hand, under certain conditions generative replay reinforces errors in a 'vicious cycle' of increasing distortion.

I now discuss limitations of this research, themes emerging from it, and ideas for future work.

5.1 Limitations

Marr (1982) proposed that cognitive processes can be studied at three levels: the computational level (the high-level computation being performed), the algorithmic level (the algorithm used to perform the computation), and the implementation level (the neural implementation of the algorithm). Understanding a cognitive process requires an integrated explanation across all three levels. This thesis addresses the computational and algorithmic levels of memory construction and consolidation, but has less to say about the implementation level, as the results relate to behavioural rather than neural data. This is the main limitation of the model.

In particular, the types of generative model used are not realistic at a neural level. Whilst biological intelligence is thought to involve learning via prediction error minimisation (e.g. Friston, 2010), the implementation of such algorithms is unclear. The deep neural networks used in this thesis are trained by computing error gradients via backpropagation, and then performing gradient descent to optimise the weights. But deep learning by backpropagation is not thought to be biologically plausible (with issues including the use of non-local information, symmetric synaptic weights, and neurons with continuous outputs), let alone the elaborate architectures of variational autoencoders and generative pre-trained transformers (Whittington & Bogacz, 2019). Further work is therefore required to bridge the gap between realistic models of learning at the synaptic level and the generative models used here. Approximations to error backpropagation do exist (Whittington & Bogacz, 2019), as well as other families of network that mimic the brain more closely (Dayan et al., 1995; Friston, 2010; Rao & Ballard, 1999). Whilst these alternatives cannot yet accomplish the feats of GPT-2 (Radford et al., 2019) etc., there is no reason to think this is impossible in principle, so these issues do not necessarily undermine the main arguments of the thesis.

Further work is also required to fully specify the extended sequential model, and to incorporate certain aspects of the static model. As described in Chapter Three, the extended sequential model would store a compressed conceptual representation of

a sequence, but whether this would be an intermediate vector representation from the generative network or some other kind of sequential 'gist' is unclear. Also, 'pixel-level' reconstruction errors were used to disentangle predictable and unpredictable elements of memories. But the analogue for sequences is not obvious - perplexity could perhaps be used to identify unexpected subsequences within a sequence, and retrieval augmented generation could recombine a stored 'gist' with such elements, but the mechanism for deconstructing and reconstructing memories was not simulated explicitly.

The sequential model captures the sequential nature of experience, but in reality each moment is 'high-dimensional' (more like the images in the static model than the 'tokens' in the sequential one). For example, video data, whilst still far less rich than reality, better reflect this: a video is made up of a succession of frames, each of which is rich in data, and recalling it requires 'filling in the gaps' of a particular frame as well as predicting successive ones. This would require changes to the associative and generative networks. Firstly, since pattern completion of both the current stimulus and the next stimulus would be required in the associative network, a combination of autoassociative and heteroassociative connectivity in the hippocampal network might be required. Secondly, the generative model could be replaced with one for video data (e.g. Yan et al., 2021), although the time and cost required to train such a model might make this impractical.

Another simplification in the current model is the division between sensory and conceptual representations (where conceptual representations are latent variables extracted from the most compressed layer of a VAE). The extended static model captures how memories bind together both sensory and conceptual representations from the outset. But really such representations must span a hierarchy of abstraction; for example, visual stimuli could be represented at one level as small patches of colour, and at another level as shapes and patterns. Going further up the hierarchy, the stimuli could be represented as objects in a particular configuration, and at the top as a schema for the scene (without specifying particular objects). These could align with representations in successive layers of a neural network, or the hierarchy

of a predictive coding network. How might more than two levels of abstraction be integrated into the model? One option is to store multiple representations and the poorly predicted elements relative to each. This could potentially correspond to the gradient of representations in the hippocampus, from fine-grained, perceptual features in posterior hippocampus to coarse-grained, conceptual ones in anterior hippocampus. (Note that sensory representations bound together in an episodic memory are already highly processed, unlike 'pixels' making up an image, so this is another aspect of the model that could be refined.)

5.2 Neural foundation models

Recent years have seen a move in machine learning from task-specific models to larger task-general ones, sometimes referred to as 'foundation models' (Bommasani et al., 2021). Similarly, neuroscience has seen a move from a modular view of many semi-independent networks learning particular tasks to a focus on the learning of multipurpose representations. We should perhaps think of the brain as learning neural 'foundation models' too, and the work in this thesis suggests how memory consolidation could contribute to their development.

A foundation model is 'any model that is trained on broad data (generally using self-supervision at scale)' (Bommasani et al., 2021, Introduction), and that can be used for, or adapted to, many different tasks. The use of a self-supervised task to train such models is key to their scaling, as human-annotated data is not required. Other advantages of this approach include enabling superior generalisation, and multimodal representation learning. One drawback of using a few task-general networks as opposed to many task-specific ones is greater susceptibility to catastrophic forgetting, as the same weights support a new task as the previous ones.

Crucially, these models can generalise to tasks they were not explicitly trained on, as large language models (LLMs) demonstrate (Brown et al., 2020). In 'zero-shot' inference, an LLM performs a task without any specific examples. For instance, when given a prompt such as 'Write a Shakespearean sonnet about the hippocampal

formation' an LLM like GPT-3 can infer how to solve the task from the prompt alone, applying its pre-existing general knowledge to generate a response (despite not having observed this combination of style and topic before). In 'few-shot' inference, an LLM can perform a task based on just a small number of examples in the prompt. An example could be asking an LLM to write a poem about the hippocampal formation in the style of a given poet, providing it with one or two examples of the poet's work. These examples highlight how foundation models enable generalisation with little or no task-specific training.

Arguments that the brain learns something akin to foundation models include the growing consensus that prediction error minimisation is key to biological intelligence (e.g. Friston, 2010), neuroimaging evidence of large task-general, or even task-negative, networks (e.g. Raichle et al., 2001), and the relative lack of external supervision in learning. This last point is more debatable, but only a small minority of the 'training data' for a human is 'labelled' by other humans (e.g. by a parent pointing out objects to a child), and for all animals rewards seem too scarce for reinforcement learning to be the main driver of knowledge acquisition. This is consistent with the view that self-supervised learning is the predominant kind of learning in the brain.

5.3 Generative models and generalisation

Chapter Four scratched the surface of a key question: once we've learned *to* imagine, how can we learn *from* imagination? What additional advantages does the brain's 'simulation machine' provide? The machine learning literature explores the advantages and disadvantages of synthetic training data, and further research could explore the implications for offline learning in the brain.

Note that the feasibility of training a model on its own outputs is controversial. There is strong evidence that a separate generative model can successfully augment the data for a classifier (Trabucco et al., 2023) or reinforcement learning agent (Sutton, 1991); it is essentially acting as a teacher for a simpler student. But this does not imply that a generative model can improve when trained on its own outputs, rather than

simply reinforcing its existing biases. Even if it can, the conditions for improvement vs. degeneration are unclear.

It has been suggested that generative replay is compositional in nature, enabling the recombination of components to imagine new events. For example, in their MEG study Schwartenbeck et al. (2023) design a task in which participants must infer which 'Tetris-style' shapes make up a novel silhouette. Representations in hippocampus and vmPFC were found to support 'vector arithmetic', such that the sum of activities for constituent shapes was closer to the true combined shape than a control shape. In other words, representations in these regions appear to be conjunctive.

Further research could explore how generative models of the kind used in this thesis relate to compositional reasoning. There are experimental and theoretical reasons to think that representations which factorise / disentangle / decompose events are useful for generalisation, and one way to think of this is as the separation of roles and entities filling those roles, as Kurth-Nelson et al. (2023) describe. When stimuli can be factorised neatly based on prior knowledge, as in the Tolman-Eichenbaum Machine (Whittington et al., 2020), this works very well, but how compositional representations are learned automatically seems more mysterious. Whether symbolic reasoning of this kind is a 'side effect' of prediction error minimisation in a deep neural network or requires some level of 'hard-coding' is unclear. In addition, the extent to which representations should be disentangled may depend on the task, and how a system would determine this is an open question.

There are many other questions relating to how generative models might be used to simulate events optimally. Firstly, how can generative models self-correct or filter out low-quality outputs? Secondly, can learning be regulated by the confidence in the imagined data? For example, one could weight imagined events so that less confident outputs have less effect on learning and vice versa (perhaps by using a low perplexity as a measure of confidence). Thirdly, how should the 'imaginativeness' of simulated events be controlled? As touched on in Chapter Four, early in learning generated data is likely to be misleading, so perhaps the model should become more 'imaginative' as training progresses.

5.4 Modelling language and cognition

Many aspects of human cognition are inextricably linked to language. Some of the computational modelling approaches in this thesis could be applied more broadly to studies involving narratives.

Firstly, Loftus and Palmer (1974) showed that answers to questions about a remembered story could be biased by language at recall time (with different estimates of the speed of a car in an accident depending on the choice of verb, e.g. with 'smashed' leading to a higher speed estimate than 'bumped'). Semantic influences at recall time could be easily modelled with GPTs encoding the stimulus story in the same way as the Bartlett story was encoded in my simulations. Secondly, Bransford et al. (1972) showed that changes congruent with the gist of a sentence were more likely to trigger false recognition memory than incongruent ones. This would be straightforward to model in the same way as the Bartlett experiment, with perplexity (how 'surprising' a candidate sentence is to the model) used as a measure of recognition.

The findings in this thesis align with an extensive literature on memory for narratives. The gist of a narrative is perhaps best described as a situation model (Zwaan & Radvansky, 1998). Bransford et al. (1972) drew the conclusion that narratives 'are information which [people] can use to construct semantic descriptions of situations' (p. 194), and it is these semantic constructions that are remembered. Whether the situation model itself is non-verbal or verbal, as in the retrieval augmented generation demonstration in Section B.2 of the Appendix, is an interesting topic for further research. (In other words, does understanding and remembering language involve constructing a non-verbal representation, or is the linguistic gist sufficient?)

More speculatively, one might wonder whether language comprehension involves the construction of situation models, and whether these topics can be explored with generative pre-trained transformers (but obviously there is a risk of anthropomorphising LLMs when discussing their 'comprehension' of a story). This has interesting implications for comprehension as a constructive process that involves generating a mental model consistent with the stimuli. This would align with the idea that a

neocortical generative network generates a gist representing the situation model, and it is this gist which is encoded, together with certain details.

There is also a connection between the literature on situation models and event boundaries and the model in this thesis. Event boundaries are thought to occur when situation models change (Zacks et al., 2007), and situation models might change when the situation model is no longer predictive of the current sequence. Suppose the system generates a situation model ('gist') and holds this in working memory. Then when the situation model is no longer predictive of new information, the current situation model would be written into memory, and a new situation model would be created. (Note that this differs from the proposal that prediction error relative to the narrative so far determines the event boundaries.) This suggestion could be explored further.

5.5 Psychiatric symptoms and conditions

At several stages of memory processing, a balance must be struck between experience and priors. One might speculate that a variety of psychiatric symptoms could be related to this, with one cluster of symptoms involving hypo-priors, and the other involving hyper-priors. In the hypo-priors case, too much weight is given to noisy input data, with bottom-up processing dominant, whereas in the hyper-priors case, too much weight is given to priors, with top-down processing dominant. (Note that the terms 'hypo-priors' and 'hyper-priors' obviously have a specific Bayesian meaning, but I am using them more loosely.)

How might this manifest at different stages of memory processing? In Chapter Two, it was suggested that when memories are encoded in the hippocampus, the prediction error threshold determines how much detail is stored (as opposed to relying on a conceptual 'gist'). When the threshold is at a 'happy medium' it strikes a balance between detail and efficiency, with some gist-based distortion from the outset that further increases with consolidation. When the threshold is low, many sensory details are encoded, but this is very inefficient in terms of storage. When the threshold is

high, few sensory details are encoded, which would produce more gist-based distortion than typical controls, with consolidation reinforcing these errors.

In Chapter Three, it was suggested that learning a 'world model' involves a combination of replay of recent memories and generative replay to stabilise older knowledge. With the right balance of these data, learning from imagination can not only help continual learning in a constantly changing world, but also aids inference and generalisation. With too little generative replay, the 'world model' changes too much based on new information, and inference and generalisation may be impaired. With too much generative replay, the model may degenerate, reinforcing its existing errors in a vicious cycle. See Table 5.1 for a summary.

Consistent with these ideas, one prominent hypothesis for autism is the hypo-priors account (Pellicano & Burr, 2012), which suggests that cognition in autism is less affected by prior beliefs. Symptoms like hypersensitivity to sensory data and difficulties in resolving ambiguity are consistent with this view (Pellicano & Burr, 2012). Relative strengths like superior attention to detail and reduced perceptual biases also align with this account. For example, people with autism are less susceptible to optical illusions (Happé, 1996), and better able to distinguish between similar stimuli (Plaisted et al., 1998). More broadly, autistic patients may struggle to apply complex priors to new experiences, e.g. social cognition involves a great deal of 'filling in the blanks' using priors about others' mental states (Frith, 2003). A number of studies look at autism from a predictive coding perspective, for example Van de Cruys et al. (2014) argue that 'errors resulting from violations to ... predictions are given a uniform, inflexibly high weight' in autism (Abstract).

The hypo-priors account (Pellicano & Burr, 2012) is potentially consistent with the 'weak coherence' account of autism (Frith, 2003; Happé & Frith, 2006), which suggests a bias towards detail-oriented rather than gist-oriented processing, with a focus on local rather than global features. Future work could disentangle the implications of the weak coherence account, predictive coding account, and account outlined here, as many of their predictions overlap.

Conversely, hyper-priors could be connected to certain symptoms and conditions involving false beliefs. This could be linked to rumination, as simulated in Chapter Four, but also a broader range of phenomena like delusion and confabulation. Let us consider a patient who holds some false beliefs (as all humans obviously do), and then undergoes a shift to the hyper-priors state of the model. If the false belief is that pirates are attacking, to use a 'belief' in the rumination simulation as an example, a memory of a boat trip may rely heavily on conceptual representations in which the sea is linked to pirates. This could mean that stronger gist-based distortions are observed than in typical controls, with pirate attacks imagined in memories in which they did not occur. Consolidation might then strengthen this belief further by assimilating the episode into a neocortical model that believes pirate attacks are common. If episodes that contradict the belief are encoded, these may be drowned out by generative replay reinforcing the current model, such that the patient is unable to correct their false belief with experience. In other words, maladaptive learning from imagined pirate attacks might further entrench the false belief.

One might predict more neocortical involvement in memory from the outset and more generative replay in patients with the hyper-priors cluster of symptoms, and the reverse in patients with the hypo-priors cluster of symptoms. Further work could test these predictions, and explore how the hyperparameters of the model (e.g. prediction error threshold and fraction of generative replay) might be set. In addition, more work could explore whether the predictive coding account and approach outlined here make different predictions relating to computational psychiatry, and if so how experimental tests could compare them.

5.6 Episodic and semantic memory

A theme throughout this thesis is that the classic distinction between episodic and semantic memory (Tulving, 1985) is much blurrier than it seems. For example, generative networks such as large language models (LLMs) can memorise 'event-

	Hypo-priors	**Typical cognition**	**Hyper-priors**
Definition	Too much weight given to noisy input data. Bottom-up processing dominates.	Delicate balance between input data and priors.	Too much weight given to priors. Top-down processing dominates.
Memory encoding	Many sensory details of an event are stored. Very inefficient in terms of hippocampal storage.	The prediction error threshold for encoding strikes a balance between efficiency and detail. Gist-based distortion from the outset, increasing with consolidation.	Few sensory details of an event are stored. More gist-based distortion / confabulation than typical controls. Consolidation reinforces current errors.
Learning a 'world model'	Online and offline learning from reality. Veridical replay dominates. Catastrophic forgetting once hippocampal memory traces fade. Poorer generalisation and inference.	A combination of veridical and generative replay consolidates new memories while stabilising 'world model'. Learning from imagination helps inference and generalisation, but with sufficient external inputs to prevent model degenerating.	Offline learning from imagination. Generative replay dominates. Model degenerates if external inputs are insufficient, producing rumination, delusion, etc.

Table 5.1: Comparison of hypo-priors, typical cognition, and hyper-priors.

unique' specifics as well as generalities in a single network (Carlini et al., 2022). If a series of narratives representing 'episodes' are 'consolidated' into an LLM, the resulting LLM could support both memory for specific episodes and for semantic 'facts', with the latter learned as a side-effect of reconstructing the former. Not only are 'beliefs' influenced by 'episodes' the network was trained on, but the 'episodes' are reshaped by the 'beliefs', as we saw in a range of memory distortion simulations. My interpretation of these findings is that generative networks can support truly episodic memory even once the hippocampal trace is erased, and that consolidated memories can consist entirely of generative predictions; one could go as far as to

say that remote episodic memory involves imagination constrained by beliefs about the past. (This raises many definitional, and even philosophical, questions about the boundary between memory and imagination, including how we could distinguish between the two.)

A different view is that episodic memory must include some stored details in a hippocampal trace to be experienced as memory rather than imagination. Even though consolidation lessens dependence of episodic memory on the hippocampus, memories arguably retain some limited trace in the hippocampus proper for a long time. Thus the retrieval of event-unique details which have not yet been consolidated could be how we distinguish memories from imagination. This could align with hippocampal deficits in recognition memory when tested with 'close foils' that differ only in small details from the previously seen item (Migo et al., 2009).

A final possibility is that truly episodic memory *can* eventually be supported by the generative network, but that the hippocampus proper plays a greater role in generation than in the models investigated so far. For example, memory fragments stored in the hippocampus proper could augment event construction with further detail drawn from real memories, or the hippocampal conceptual representations could provide more than links to generative latent variables, e.g. ensuring that retrieved scenes correspond to single viewpoints via place and head direction cells (Becker & Burgess, 2000; Bicanski & Burgess, 2018). This would align with the finding that the construction of complex and coherent scenes benefits from an intact hippocampus proper, even though the entorhinal cortex can generate simple scenes (Hassabis et al., 2007).

5.7 From behavioural to neural data

There are many other ways in which the neural substrates of the model could be better understood (in addition to points in Sections 5.1 and 5.6).

Firstly, the differing roles of anterior hippocampus (aHPC) and posterior hippocampus

(pHPC) could be explored; as discussed in Chapter One, aHPC is associated with coarse-grained conceptual representations of memory, while pHPC is associated with fine-grained perceptual representations (Robin & Moscovitch, 2017; Zeidman & Maguire, 2016). The extended model proposes that memories are encoded as a combination of conceptual and sensory features (the latent variables and the poorly-predicted elements respectively). One might predict that the former corresponds more to aHPC and the latter more to pHPC. A test of this could be whether 'expectations' are more easily decoded from aHPC while elements deviating from expectation are more easily decoded from pHPC. There is some evidence that aHPC activity increases relative to pHPC activity as a memory is consolidated (Robin & Moscovitch, 2017), consistent with the predictions of Zeidman and Maguire (2016) and Moscovitch et al. (2016); one might also speculate that the generative network is more dependent on aHPC, so this could be explored further.

Secondly, how this relates to the lateralisation of hippocampal function could also be investigated. Maguire and Frith (2003) found that right hippocampus was less active for remote than recent memories, with no such temporal gradient for left hippocampus. This is intriguing as the right hippocampus is more associated with memory for visuospatial detail, while the left hippocampus is more associated with memory for narratives (Burgess et al., 2002). As above, whether 'expectations' are more easily decoded from left hippocampus while elements deviating from expectation are more easily decoded from right hippocampus could be tested.

Thirdly, the link between the latent spaces of generative models and cognitive maps in the hippocampal formation (Behrens et al., 2018) could be explored further. Grid cells, with their characteristic hexagonal firing pattern as an animal moves through an environment (Moser et al., 2008), are thought to be a mechanism behind path integration and vector navigation (Bush et al., 2015). This model is consistent with earlier proposals that EC encodes latent representations, capturing shared structure to enable inference in spatial and non-spatial tasks (Whittington et al., 2020). But this thesis does not explore how grid cells emerge, and whether this happens automatically as a result of prediction error minimisation, or depends on additional constraints.

Finally, the correspondence between components of the hippocampal autoassociative network and cell types in the hippocampal formation requires further thought. One view is that episode-specific neurons (Kolibius et al., 2021) bind memories, acting as an abstract 'index' for a pattern across the feature units. These episode-specific neurons could correspond to memory units in a modern Hopfield network (Krotov & Hopfield, 2020; Ramsauer et al., 2020). Another view is that episode-specific neurons simply reflect event-unique features of memory. These could also be sufficient to bind a trace together, without requiring an explicit 'index' for the pattern, but this would align less closely to the modern Hopfield network.

5.8 Neurodevelopmental implications

Memory and learning change dramatically over the course of development, so the neurodevelopmental implications of the model in this thesis could be explored further. For example, adults are generally unable to recall episodic memories from before the age of three to four, with very limited recall up to age seven or so (Bauer & Larkina, 2014), a phenomenon known as infantile amnesia (Miles, 1895).

Infantile amnesia seems puzzling given the effects of early experiences on shaping brain development (Alberini & Travaglia, 2017; Josselyn & Frankland, 2012). One might assume that memories are not encoded properly in early childhood, but there is strong behavioural evidence that infants are capable of memory (Mullally & Maguire, 2014), e.g. babies rapidly learn to recognize their mother's face and voice (DeCasper & Fifer, 1980). One suggestion is that systems consolidation cannot happen effectively for infants and young children because the neocortical networks into which memories are consolidated are immature (Alberini & Travaglia, 2017). Infancy can be seen as a period of great plasticity during which the brain 'learns to remember' (Alberini & Travaglia, 2017; Mullally & Maguire, 2014). This is consistent with the proposal in this thesis; if consolidation involves assimilation into a generative model, an immature and unstable generative model would mean there is not a mature conceptual framework into which to integrate experiences.

Note that there are alternative explanations including retrieval failure and instability due to hippocampal neurogenesis. The former suggests that memories from early childhood are encoded adequately, but cannot be retrieved, since neocortical representations have changed so much that the original memory is no longer accessible (Munakata, 2004); this is reminiscent of the Káli and Dayan (2004) description of catastrophic forgetting. For example, Hayne and Rovee-Collier (1995) argue that our earliest memories are encoded in non-verbal terms, which our adult brains cannot interpret. This invokes the encoding specificity principle (Tulving, 1983); the context of recall in adulthood may be very different to the context of encoding in infancy, decreasing the recall accuracy. Another view is that neurogenesis in the hippocampus causes accelerated forgetting (Akers et al., 2014; Josselyn & Frankland, 2012), as the rate of neurogenesis is much greater in early childhood than in adulthood (Knoth et al., 2010).

In Chapter Four I explored how catastrophic forgetting could be avoided in the generative model during consolidation. Another hypothesis might be that some initial consolidation occurs, but the mechanisms to avoid catastrophic forgetting are immature. Darby and Sloutsky (2015) note that in young children retroactive interference effects are 'reminiscent of the catastrophic interference effects observed in simple connectionist models' (Implications section). Intriguingly the period of infantile amnesia appears to coincide with reduced dreaming (Foulkes, 2009), which could also be consistent with an immature generative model that cannot 'stabilise' old memories.

5.9 Conclusion

Consolidation can be thought of as a process of learning to construct events, or in other words, a process of 'learning to imagine'. In this thesis I suggested that memories are replayed over the course of consolidation to train a predictive model of the world, which supports functions including episodic memory, semantic memory, and imagination, and forms of statistical learning, inference, and planning. Generative

neural networks provide a mechanism for extracting conceptual structure through prediction error minimisation, eventually encoding 'priors' for the reconstruction of stimuli such that recall involves 'predicting' the past. I suggest this corresponds to a network including the hippocampal formation and association cortex which is involved in many kinds of event generation.

I presented a computational model in which episodic memories are initially encoded in the hippocampus, then replayed to train a neocortical generative network to (re)construct sensory experiences. I also showed how this could be extended to sequential stimuli to capture the sequential nature of experience. Even right after encoding an experience, remembering involves imagining the past, combining some stored hippocampal details with neocortical expectations. Finally, generative networks trained through consolidation are continually updated over a lifespan, maintaining a relatively stable 'world model' without catastrophic forgetting. One strategy to achieve this may be learning *from* imagination, with generative replay stabilising old knowledge as new knowledge is assimilated into neocortical networks.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, *9*(1), 147–169.

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377.

Akers, K. G., Martinez-Canabal, A., Restivo, L., Yiu, A. P., De Cristofaro, A., Hsiang, H.-L., Wheeler, A. L., Guskjolen, A., Niibori, Y., Shoji, H., et al. (2014). Hippocampal neurogenesis regulates forgetting during adulthood and infancy. *Science*, *344*(6184), 598–602.

Alammar, J. (2018). The illustrated transformer [Accessed on March 19, 2024]. https://jalammar.github.io/illustrated-transformer/

Alammar, J. (2019). The illustrated GPT-2 [Accessed on March 19, 2024]. https://jalammar.github.io/illustrated-gpt2/

Alberini, C. M., & Travaglia, A. (2017). Infantile amnesia: A critical period of learning to learn and remember. *Journal of Neuroscience*, *37*(24), 5783–5795.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2023). Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the national academy of sciences*, *91*(15), 7041–7045.

Andersen, P. (1975). Organization of hippocampal neurons and their interconnections. In *The hippocampus: Volume 1: Structure and development* (pp. 155–175). Springer.

Arbib, M. A. (2020). From spatial navigation via visual construction to episodic memory and imagination. *Biological cybernetics*, *114*(2), 139–167.

Bainbridge, W. A., & Baker, C. I. (2020a). Boundaries extend and contract in scene memory depending on image properties. *Current Biology*, *30*(3), 537–543.

Bainbridge, W. A., & Baker, C. I. (2020b). Reply to intraub. *Current Biology*, *30*(24), R1465–R1466.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01), 217–233.

Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, *1*(3), 295–311.

Barry, D. N., & Love, B. C. (2021). A neural network account of memory replay and knowledge consolidation. *bioRxiv*.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Bauer, P. J., & Larkina, M. (2014). Childhood amnesia in the making: Different distributions of autobiographical memories in children and adults. *Journal of Experimental Psychology: General*, *143*(2), 597.

Beck, A. T., & Rector, N. A. (2003). A cognitive model of hallucinations. *Cognitive therapy and research*, *27*, 19–52.

Becker, S., & Burgess, N. (2000). Modelling spatial recall, mental imagery and neglect. *Advances in neural information processing systems*, *13*.

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

Bein, O., Plotkin, N. A., & Davachi, L. (2021). Mnemonic prediction errors promote detailed memories. *Learning & Memory*, *28*(11), 422–434.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P., & Wiener, S. I. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. *Neuron*, *66*(6), 921–936.

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature neuroscience*, *15*(10), 1439–1444.

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

Ben-Yakov, A., & Dudai, Y. (2011). Constructing realistic engrams: Poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *Journal of Neuroscience*, *31*(24), 9032–9042.

Berg, M., Feldmann, M., Kirchner, L., & Kube, T. (2022). Oversampled and undersolved: Depressive rumination from an active inference perspective. *Neuroscience & Biobehavioral Reviews*, 104873.

Bergman, E. T., & Roediger, H. L. (1999). Can Bartlett's repeated reproduction experiments be replicated? *Memory & cognition*, *27*(6), 937–947.

Bicanski, A., & Burgess, N. (2018). A neural-level model of spatial memory and imagery. *Elife*, *7*, e33752.

Biderman, N., Bakkour, A., & Shohamy, D. (2020). What are memories for? the hippocampus bridges past experience with future decisions. *Trends in Cognitive Sciences*, *24*(7), 542–556.

Bisby, J. A., Burgess, N., & Brewin, C. R. (2020). Reduced memory coherence for negative events and its relationship to posttraumatic stress disorder. *Current Directions in Psychological Science*, *29*(3), 267–272.

Blagrove, M., Fouquet, N. C., Henley-Einion, J. A., Pace-Schott, E. F., Davies, A. C., Neuschaffer, J. L., & Turnbull, O. H. (2011). Assessing the dream-lag effect for REM and NREM stage 2 dreams. *PLoS One*, *6*(10), e26708.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, *76*, 198–211.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bonnici, H. M., Chadwick, M. J., & Maguire, E. A. (2013). Representations of recent and remote autobiographical memories in hippocampal subfields. *Hippocampus*, *23*(10), 849–854.

Boyce, R., Glasgow, S. D., Williams, S., & Adamantidis, A. (2016). Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science*, *352*(6287), 812–816.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive psychology*, *3*(2), 193–209.

Bright, I. M., Meister, M. L., Cruzado, N. A., Tiganj, Z., Buffalo, E. A., & Howard, M. W. (2020). A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proceedings of the National Academy of Sciences*, *117*(33), 20274–20283.

Bright, P., Buckman, J., Fradera, A., Yoshimasu, H., Colchester, A. C., & Kopelman, M. D. (2006). Retrograde amnesia in patients with hippocampal, medial temporal, temporal lobe, or frontal pathology. *Learning & Memory*, *13*(5), 545–557.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burgess, C., & Kim, H. (2018). 3D shapes dataset.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological review*, *106*(3), 551.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–641.

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

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological review*, *114*(2), 340.

Caccia, L., Aljundi, R., Asadi, N., Tuytelaars, T., Pineau, J., & Belilovsky, E. (2021). Reducing representation drift in online continual learning. *arXiv preprint arXiv:2104.05025*.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Carmichael, L., Hogan, H. P., & Walter, A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of experimental Psychology*, *15*(1), 73.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, *37*(1), 54–115.

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nature neuroscience*, *14*(2), 147–153.

Carse, T., & Langdon, R. (2013). Delusion proneness in nonclinical individuals and cognitive insight: The contributions of rumination and reflection. *The Journal of nervous and mental disease*, *201*(8), 659–664.

Cavallero, C., Cicogna, P., Natale, V., Occhionero, M., & Zito, A. (1992). Slow wave sleep dreaming. *Sleep*, *15*(6), 562–566.

Chambers, D., & Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: Human perception and performance*, *11*(3), 317.

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., & Rossor, M. N. (2001). Patterns

of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of neurology*, *49*(4), 433–442.

Chaudhry, H. T., Zavatone-Veth, J. A., Krotov, D., & Pehlevan, C. (2023). Long sequence Hopfield memory. *arXiv preprint arXiv:2306.04532*.

Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H., & Wagner, A. D. (2011). Associative retrieval processes in the human medial temporal lobe: Hippocampal retrieval success and ca1 mismatch detection. *Learning & Memory*, *18*(8), 523–528.

Chiarello, C., & Beeman, M. (1997). Toward a veridical interpretation of right-hemisphere processing and storage. *Brain and Language*, *22*, 253–265.

Chollet, F., et al. (2015). Keras documentation. *keras. io*, *33*.

Cipolotti, L., Shallice, T., Chan, D., Fox, N., Scahill, R., Harrison, G., Stevens, J., & Rudge, P. (2001). Long-term retrograde amnesia... the crucial role of the hippocampus. *Neuropsychologia*, *39*(2), 151–172.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual review of psychology*, *62*, 271–298.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468.

Dandolo, L. C., & Schwabe, L. (2018). Time-dependent memory transformation along the hippocampal anterior–posterior axis. *Nature communications*, *9*(1), 1–11.

Darby, K. P., & Sloutsky, V. M. (2015). The cost of learning: Interference effects in memory development. *Journal of Experimental Psychology: General*, *144*(2), 410.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

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

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural computation*, *7*(5), 889–904.

De Luca, F., McCormick, C., Mullally, S. L., Intraub, H., Maguire, E. A., & Ciaramelli, E. (2018). Boundary extension is attenuated in patients with ventromedial prefrontal cortex damage. *cortex*, *108*, 1–12.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology*, *58*(1), 17.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., & Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, *168*(2), 288–299.

Deperrois, N., Petrovici, M. A., Senn, W., & Jordan, J. (2022). Learning cortical representations through perturbed and adversarial dreaming. *Elife*, *11*, e76384.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241–1242.

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.

Dragoi, G., & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, *469*(7330), 397–401.

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, *142*(4), 1277.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia*, *49*(5), 1322–1331.

Ego-Stengel, V., & Wilson, M. A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, *20*(1), 1–10.

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International journal of cognitive therapy*, *1*(3), 192–205.

Eichenbaum, H. (2014). Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), 732–744.

Ekstrom, A. D., Caplan, J. B., Ho, E., Shattuck, K., Fried, I., & Kahana, M. J. (2005). Human hippocampal theta activity during virtual navigation. *Hippocampus*, *15*(7), 881–889.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, *104*(18), 7723–7728.

Evans, T., & Burgess, N. (2019). Coordinated hippocampal-entorhinal replay as structural inference. *Advances in Neural Information Processing Systems*, *32*.

Fayyaz, Z., Altamimi, A., Zoellner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022). A model of semantic completion in generative episodic memory. *Neural Computation*, *34*(9), 1841–1870.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in cognitive sciences*, *14*(3), 119–130.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322.

Flesch, T., Saxe, A., & Summerfield, C. (2023). Continual task learning in natural and artificial agents. *Trends in Neurosciences*, *46*(3), 199–210.

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

Foulkes, D. (2009). *Children's dreaming and the development of consciousness*. Harvard University Press.

Fox, K. C., Nijeboer, S., Solomonova, E., Domhoff, G. W., & Christoff, K. (2013). Dreaming as mind wandering: Evidence from functional neuroimaging and first-person content reports. *Frontiers in human neuroscience*, *7*, 412.

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature reviews neuroscience*, *6*(2), 119–130.

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, *127*(3), 327.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

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

Frith, U. (2003). *Autism: Explaining the enigma*. Blackwell publishing.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018). Born again neural networks. *International conference on machine learning*, 1607–1616.

Gais, S., Albouy, G., Boly, M., Dang-Vu, T. T., Darsaud, A., Desseilles, M., Rauchs, G., Schabus, M., Sterpenich, V., Vandewalle, G., et al. (2007). Sleep transforms the cerebral trace of declarative memories. *Proceedings of the National Academy of Sciences*, *104*(47), 18778–18783.

George, T. M., Barry, C., Stachenfeld, K. L., Clopath, C., & Fukai, T. (2023). A generative model of the hippocampal formation trained with theta driven local learning rules. *bioRxiv*, 2023–12.

Gershman, S. J. (2019). The generative adversarial brain. *Frontiers in Artificial Intelligence*, 18.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114.

Ghosh, V. E., Moscovitch, M., Colella, B. M., & Gilboa, A. (2014). Schema representation in patients with ventromedial PFC lesions. *Journal of Neuroscience*, *34*(36), 12057–12070.

Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, *21*(8), 618–631.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience*, *12*(10), 1222–1223.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*(4), 491–516.

González, O. C., Sokolov, Y., Krishnan, G. P., Delanois, J. E., & Bazhenov, M. (2020). Can sleep protect memories from catastrophic forgetting? *Elife*, *9*, e51005.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

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

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

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, *24*(12), 1028–1040.

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.

Hamilton, J. P., Furman, D. J., Chang, C., Thomason, M. E., Dennis, E., & Gotlib, I. H. (2011). Default-mode and task-positive network activity in major depressive disorder: Implications for adaptive and maladaptive rumination. *Biological psychiatry*, *70*(4), 327–333.

Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, *36*, 5–25.

Happé, F. G. (1996). Studying weak central coherence at low levels: Children with autism do not succumb to visual illusions. a research note. *Journal of child psychology and psychiatry*, *37*(7), 873–877.

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

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, *11*(7), 299–306.

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

Hasselmo, M. E., Wyble, B. P., & Wallenstein, G. V. (1996). Encoding and retrieval of episodic memories: Role of cholinergic and gabaergic modulation in the hippocampus. *Hippocampus*, *6*(6), 693–708.

Hayne, H., & Rovee-Collier, C. (1995). The organization of reactivated memory in infancy. *Child development*, *66*(3), 893–906.

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, *6*(5), 709–719.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2016). beta-VAE: Learning basic visual concepts with a constrained variational framework.

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4*(5), 5947.

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade: Second edition* (pp. 599–619). Springer.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*(5214), 1158–1161.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Hobson, J. A., Pace-Schott, E. F., & Stickgold, R. (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and brain sciences*, *23*(6), 793–842.

Hodges, J. R., & Graham, K. S. (2001). Episodic memory: Insights from semantic dementia. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1423–1434.

Hoel, E. (2021). The overfitted brain: Dreams evolved to assist generalization. *Patterns*, *2*(5).

Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469.

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

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.

Hou, X., Shen, L., Sun, K., & Qiu, G. (2017). Deep feature consistent variational autoencoder. *2017 IEEE winter conference on applications of computer vision (WACV)*, 1133–1141.

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

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological review*, *98*(3), 352.

Igata, H., Ikegaya, Y., & Sasaki, T. (2021). Prioritized experience replays on a hippocampal predictive map for learning. *Proceedings of the National Academy of Sciences*, *118*(1), e2011266118.

Intraub, H. (2020). Searching for boundary extension. *Current Biology*, *30*(24), R1463–R1464.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 179.

Jones, S. R., & Fernyhough, C. (2009). Rumination, reflection, intrusive thoughts, and hallucination-proneness: Towards a new model. *Behaviour Research and Therapy*, *47*(1), 54–59.

Josselyn, S. A., & Frankland, P. W. (2012). Infantile amnesia: A neurogenic hypothesis. *Learning & Memory*, *19*(9), 423–433.

Káli, S., & Dayan, P. (2000). Hippocampally-dependent consolidation in a hierarchical model of neocortex. *Advances in Neural Information Processing Systems*, *13*.

Káli, S., & Dayan, P. (2002). Replay, repair and consolidation. *Advances in Neural Information Processing Systems*, *15*.

Káli, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature neuroscience*, *7*(3), 286–294.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. *International Conference on Machine Learning*, 15696–15707.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Keinath, A. T., Wang, M. E., Wann, E. G., Yuan, R. K., Dudman, J. T., & Muzzio, I. A. (2014). Precise spatial coding is preserved along the longitudinal hippocampal axis. *Hippocampus*, *24*(12), 1533–1548.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. *The world wide web conference*, 2915–2921.

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

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Knoth, R., Singec, I., Ditter, M., Pantazis, G., Capetian, P., Meyer, R. P., Horvat, V., Volk, B., & Kempermann, G. (2010). Murine features of neurogenesis in the human hippocampus across the lifespan from 0 to 100 years. *PloS one*, *5*(1), e8809.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & memory*, *1*(2), 106–120.

Kolibius, L., Roux, F., Parish, G., Ter Wal, M., Van Der Plas, M., Chelvarajah, R., Sawlani, V., Rollings, D., Lang, J., Gollwitzer, S., et al. (2021). Hippocampal neurons code individual episodic memories in humans. *bioRxiv*, 2021–06.

Kopelman, M. D. (2010). Varieties of confabulation and delusion. *Cognitive neuropsychiatry*, *15*(1-3), 14–37.

Kornmeier, J., & Bach, M. (2005). The necker cube—an ambiguous figure disambiguated in early visual processing. *Vision research*, *45*(8), 955–960.

Koscik, T. R., & Tranel, D. (2012). The human ventromedial prefrontal cortex is critical for transitive inference. *Journal of cognitive neuroscience*, *24*(5), 1191–1204.

Krotov, D., & Hopfield, J. (2020). Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*.

Krotov, D., & Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in neural information processing systems*, *29*.

Kumaran, D. (2012). What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience*, *6*, 157.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in cognitive sciences*, *20*(7), 512–534.

Kumaran, D., & Maguire, E. A. (2006). An unexpected sequence of events: Mismatch detection in the human hippocampus. *PLoS biology*, *4*(12), e424.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory: Insights from semantic dementia. *Annals of the New York Academy of Sciences*, *1124*(1), 61–76.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database. *AT&T Labs [Online].*, *2*.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature communications*, *7*(1), 13619.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, A., Lei, X., Herdman, K., Waidergoren, S., Gilboa, A., & Rosenbaum, R. S. (2024). Impoverished details with preserved gist in remote and recent spatial memory following hippocampal and fornix lesions. *Neuropsychologia*, 108787.

Lifanov, J., Linde-Domingo, J., & Wimber, M. (2021). Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nature communications*, *12*(1), 3177.

Lin, F., Hafri, A., & Bonner, M. F. (2022). Scene memories are biased toward high-probability views. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(10), 1116.

Lin, W.-J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific reports*, *6*(1), 1–10.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex*, *53*, 60–77.

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585–589.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., et al. (2024). Large language models surpass human experts in predicting neuroscience results. *arXiv preprint arXiv:2403.03230*.

Lyubomirsky, S., Caldwell, N. D., & Nolen-Hoeksema, S. (1998). Effects of ruminative and distracting responses to depressed mood on retrieval of autobiographical memories. *Journal of personality and social psychology*, *75*(1), 166.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature communications*, *11*(1), 1–11.

Maekawa, A., Kamigaito, H., Funakoshi, K., & Okumura, M. (2023). Generative replay inspired by hippocampal memory indexing for continual language learning. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 930–942.

Maguire, E. A., & Frith, C. D. (2003). Lateral asymmetry in the hippocampal response to the remoteness of autobiographical memories. *Journal of Neuroscience*, *23*(12), 5302–5307.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, *42*(4), 824–836.

Manns, J. R., Hopkins, R. O., & Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron*, *38*(1), 127–133.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal society of London. Series B. Biological sciences*, *176*(1043), 161–234.

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

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, *11*(10), 442–450.

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.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (pp. 109–165). Elsevier.

McCormick, C., Barry, D. N., Jafarian, A., Barnes, G. R., & Maguire, E. A. (2020). vmPFC drives hippocampal processing during autobiographical memory recall regardless of remoteness. *Cerebral Cortex*, *30*(11), 5972–5987.

McCormick, C., Dalton, M. A., Zeidman, P., & Maguire, E. A. (2021). Characterising the hippocampal response to perception, construction and complexity. *cortex*, *137*, 1–17.

McCormick, C., Rosenthal, C. R., Miller, T. D., & Maguire, E. A. (2018). Mind-wandering in people with hippocampal damage. *Journal of Neuroscience*, *38*(11), 2745–2754.

McDevitt, E. A., Duggan, K. A., & Mednick, S. C. (2015). REM sleep rescues learning from interference. *Neurobiology of learning and memory*, *122*, 51–62.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McKenzie, S., & Eichenbaum, H. (2011). Consolidation and reconsolidation: Two lives of memories? *Neuron*, *71*(2), 224–233.

McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in neurosciences*, *10*(10), 408–415.

Michon, F., Sun, J.-J., Kim, C. Y., Ciliberti, D., & Kloosterman, F. (2019). Post-learning hippocampal replay selectively reinforces spatial memory for highly rewarded locations. *Current Biology*, *29*(9), 1436–1444.

Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, *62*(6), 1198–1215.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Miles, C. (1895). A study of individual psychology. *The American Journal of Psychology*, *6*(4), 534–558.

Millidge, B., Salvatori, T., Song, Y., Lukasiewicz, T., & Bogacz, R. (2022). Universal Hopfield networks: A general framework for single-shot associative memory models. *International Conference on Machine Learning*, 15561–15583.

Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: A theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mokady, R., Hertz, A., & Bermano, A. H. (2021). CLIPCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: The hippocampus and neocortex in transformation. *Annual review of psychology*, *67*, 105–134.

Moscovitch, M., & Melo, B. (1997). Strategic retrieval and the frontal lobes: Evidence from confabulation and amnesia. *Neuropsychologia*, *35*(7), 1017–1034.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*, 69–89.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849.

Mullally, S. L., Intraub, H., & Maguire, E. A. (2012). Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. *Current Biology*, *22*(4), 261–268.

Mullally, S. L., & Maguire, E. A. (2014). Learning to remember: The early ontogeny of episodic memory. *Developmental cognitive neuroscience*, *9*, 12–29.

Mumford, D. (1992). On the computational architecture of the neocortex: Ii the role of cortico-cortical loops. *Biological cybernetics*, *66*(3), 241–251.

Munakata, Y. (2004). Computational cognitive neuroscience of early memory development. *Developmental Review*, *24*(1), 133–153.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, *7*(2), 217–227.

Nader, K., & Hardt, O. (2009). A single standard for memory: The case for reconsolidation. *Nature Reviews Neuroscience*, *10*(3), 224–234.

Nagy, D. G., Török, B., & Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, *16*(10), e1008367.

Nejad, A. B., Fossati, P., & Lemogne, C. (2013). Self-referential processing, rumination, and cortical midline structures in major depression. *Frontiers in human neuroscience*, *7*, 666.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology*, *28*(1), 28.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436–450.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2021). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature*, *595*(7865), 80–84.

Norman, K. A., Newman, E. L., & Perotte, A. J. (2005). Methods for reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks*, *18*(9), 1212–1228.

Norman, Y., Raccah, O., Liu, S., Parvizi, J., & Malach, R. (2021). Hippocampal ripples and their coordinated dialogue with the default mode network during recent and remote recollection. *Neuron*, *109*(17), 2767–2780.

Ochsner, K. N., Schacter, D. L., & Edwards, K. (1997). Illusory recall of vocal affect. *Memory*, *5*(4), 433–455.

Odaibo, S. (2019). Tutorial: Deriving the standard variational autoencoder (VAE) loss function. *arXiv preprint arXiv:1907.08956*.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

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

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), R37–R50.

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., & Csicsvari, J. (2010). Play it again: Reactivation of waking experience and memory. *Trends in neurosciences*, *33*(5), 220–229.

Oudiette, D., Dealberto, M.-J., Uguccioni, G., Golmard, J.-L., Merino-Andreu, M., Tafti, M., Garma, L., Schwartz, S., & Arnulf, I. (2012). Dreaming without REM sleep. *Consciousness and cognition*, *21*(3), 1129–1140.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Park, J., Josephs, E. L., & Konkle, T. (2021). Systematic transition from boundary extension to contraction along an object-to-scene continuum.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, *8*(12), 976–987.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L.-W., Wamsley, E. J., Tucker, M. A., Walker, M. P., & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of learning and memory*, *92*(3), 327–334.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Phillips, C., Degueldre, C., Del Fiore, G., Aerts, J., et al. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron*, *44*(3), 535–545.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences*, *16*(10), 504–510.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in cognitive sciences*, *19*(5), 285–293.

Pezzulo, G., Zorzi, M., & Corbetta, M. (2021). The secret life of predictive brains: What's spontaneous activity for? *Trends in Cognitive Sciences*, *25*(9), 730–743.

Pfeiffer, B. E., & Foster, D. J. (2015). Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science*, *349*(6244), 180–183.

Plaisted, K., O'Riordan, M., & Baron-Cohen, S. (1998). Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual

learning task. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(5), 765–775.

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

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & cognition*, *34*, 1150–1156.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the national academy of sciences*, *98*(2), 676–682.

Ramachandran, V. S., & Hirstein, W. (1998). The perception of phantom limbs. The DO Hebb lecture. *Brain: a journal of neurology*, *121*(9), 1603–1630.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *International Conference on Machine Learning*, 8821–8831.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., et al. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature reviews neuroscience*, *13*(10), 713–726.

Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y. W., & Hadsell, R. (2019). Continual unsupervised representation learning. *Advances in Neural Information Processing Systems*, *32*.

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.

Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018). Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Raykov, P. P., Varga, D., & Bird, C. M. (2023). False memories for ending of events. *Journal of Experimental Psychology: General*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature neuroscience*, *17*(7), 981–986.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. *International conference on machine learning*, 4364–4373.

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current opinion in behavioral sciences*, *17*, 114–123.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, *7*(2), 123–146.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*(3), 231–237.

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.

Roediger, H., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. *Theories of memory II*, 187–245.

Roscow, E. L., Chua, R., Costa, R. P., Jones, M. W., & Lepora, N. (2021). Learning offline: Memory replay in biological and artificial reinforcement learning. *Trends in neurosciences*, *44*(10), 808–821.

Rothschild, G., Eban, E., & Frank, L. M. (2017). A cortical–hippocampal–cortical loop of information processing during memory consolidation. *Nature neuroscience*, *20*(2), 251–259.

Salvatori, T., Song, Y., Hong, Y., Sha, L., Frieder, S., Xu, Z., Bogacz, R., & Lukasiewicz, T. (2021). Associative memories via predictive coding. *Advances in Neural Information Processing Systems*, *34*, 3874–3886.

Salvatori, T., Song, Y., Millidge, B., Xu, Z., Sha, L., Emde, C., Bogacz, R., & Lukasiewicz, T. (2022). Incremental predictive coding: A parallel and fully automatic learning algorithm. *arXiv preprint arXiv:2212.00720*.

Schacter, D. L. (2012). Constructive memory: Past and future. *Dialogues in clinical neuroscience*, *14*(1), 7.

Schacter, D. L., & Addis, D. R. (2007). On the constructive episodic simulation of past and future events. *Behavioral and Brain Sciences*, *30*(3), 331–332.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature reviews neuroscience*, *8*(9), 657–661.

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current opinion in behavioral sciences*, *17*, 41–50.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned

information and predicts memory performance. *Nature communications*, *9*(1), 3920.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2021). Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *bioRxiv*, 2021–06.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885–4897.

Scialom, T., Chakrabarty, T., & Muresan, S. (2022). Fine-tuned language models are continual learners. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6107–6122.

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

Sherman, B. E., Graves, K. N., Huberdeau, D. M., Quraishi, I. H., Damisah, E. C., & Turk-Browne, N. B. (2022). Temporal dynamics of competition between statistical learning and episodic memory in intracranial recordings of human visual cortex. *bioRxiv*.

Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. *Advances in neural information processing systems*, *30*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Siegel, J. M. (2001). The REM sleep-memory consolidation hypothesis. *Science*, *294*(5544), 1058–1063.

Singh, D., Norman, K. A., & Schapiro, A. C. (2022). A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *bioRxiv*.

Smith, C. (2001). Sleep states and memory processes in humans: Procedural versus declarative memory systems. *Sleep medicine reviews*, *5*(6), 491–506.

Sompolinsky, H., & Kanter, I. (1986). Temporal association in asymmetric neural networks. *Physical review letters*, *57*(22), 2861.

Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., & Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *Journal of Neuroscience*, *38*(15), 3767–3775.

Spanò, G., Pizzamiglio, G., McCormick, C., Clark, I. A., De Felice, S., Miller, T. D., Edgin, J. O., Rosenthal, C. R., & Maguire, E. A. (2020). Dreaming with hippocampal damage. *Elife*, *9*, e56211.

Spiers, H. J., Maguire, E. A., & Burgess, N. (2001). Hippocampal amnesia. *Neurocase*, *7*(5), 357–382.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current opinion in neurobiology*, *5*(2), 169–177.

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor perspectives in biology*, *7*(8), a021766.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *216*(1205), 427–459.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643–1653.

Stella, F., Baracskay, P., O'Neill, J., & Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, *102*(2), 450–461.

Stoianov, I., Maisto, D., & Pezzulo, G. (2022). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *Progress in Neurobiology*, *217*, 102329.

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience*, *15*(10), 655–669.

Sun, F.-K., Ho, C.-H., & Lee, H.-Y. (2019). LAMOL: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2021). Organizing memories for generalization in complementary learning systems. *BioRxiv*.

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

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Tadros, T., Krishnan, G. P., Ramyaa, R., & Bazhenov, M. (2022). Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature Communications*, *13*(1), 7742.

Takashima, A., Nieuwenhuis, I. L., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernández, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *Journal of Neuroscience*, *29*(32), 10087–10093.

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M., McNaughton, B., & Fernández, G. (2006). Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences*, *103*(3), 756–761.

Tang, M., Barron, H., & Bogacz, R. (2023). Sequential memory with temporal predictive coding. *arXiv preprint arXiv:2305.11982*.

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

Teyler, T. J., & Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, *17*(12), 1158–1169.

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

Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, *2*(2), 189–199.

Tsao, A., Moser, M.-B., & Moser, E. I. (2013). Traces of experience in the lateral entorhinal cortex. *Current biology*, *23*(5), 399–405.

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

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82.

Tulving, E. (1983). Elements of episodic memory.

Tulving, E. (1985). How many memory systems are there? *American psychologist*, *40*(4), 385.

Umbach, G., Kantak, P., Jacobs, J., Kahana, M., Pfeiffer, B. E., Sperling, M., & Lega, B. (2020). Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proceedings of the National Academy of Sciences*, *117*(45), 28463–28474.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., De-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological review*, *121*(4), 649.

Van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, *11*(1), 1–14.

Van de Ven, G. M., & Tolias, A. S. (2018). Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635*.

Van Der Kolk, B. A., Burbridge, J. A., & Suzuki, J. (1997). The psychobiology of traumatic memory. clinical implications of neuroimaging studies.

Van Kesteren, M. T., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences*, *107*(16), 7550–7555.

Van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in neurosciences*, *35*(4), 211–219.

Van Vugt, M. K., van der Velde, M., & Investigators, E.-M. (2018). How does rumination impact cognition? a first mechanistic model. *Topics in Cognitive Science*, *10*(1), 175–191.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, *277*(5324), 376–380.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vertes, R. P., & Eastman, K. E. (2000). The case against memory consolidation in REM sleep. *Behavioral and brain sciences*, *23*(6), 867–876.

Vikbladh, O., Burgess, N., & Russek, E. (2024). Computational and Systems Neuroscience (COSYNE).

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., & Daw, N. D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron*, *102*(3), 683–693.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: Sleep-dependent memory evolution. *Nature Reviews Neuroscience*, *11*(3), 218–218.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279–292.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences*, *97*(20), 11125–11129.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in cognitive sciences*, *23*(3), 235–250.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.

Whittington, J. C., Warren, J., & Behrens, T. E. (2021). Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*.

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

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society*, *17*(5), 766–780.

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, *48*(8), 2339–2356.

Witter, M. P., Wouterlood, F. G., Naber, P. A., & Van Haeften, T. (2000). Anatomical organization of the parahippocampal-hippocampal network. *Annals of the New York Academy of Sciences*, *911*(1), 1–24.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*.

Yan, X., Yang, J., Sohn, K., & Lee, H. (2016). Attribute2Image: Conditional image generation from visual attributes. *Computer Vision–ECCV 2016: 14th*

European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, 776–791.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yu, J. Y., Liu, D. F., Loback, A., Grossrubatscher, I., & Frank, L. M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory. *Nature communications*, *9*(1), 1–11.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience*, *4*(6), 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, *130*(1), 29.

Zeidman, P., & Maguire, E. A. (2016). Anterior hippocampus: The anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience*, *17*(3), 173–182.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.

Zola-Morgan, S., Squire, L. R., & Amaral, D. G. (1986). Human amnesia and the medial temporal region: Enduring memory impairment following a bilateral lesion limited to field ca1 of the hippocampus. *Journal of Neuroscience*, *6*(10), 2950–2967.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.

Appendix A

Data and code availability

A.1 Data availability

The following datasets (all covered by the Creative Commons Attribution 4.0 License) were used in the simulations:

Dataset	**Origin**
MNIST (LeCun et al., 2010)	https://www.tensorflow.org/datasets/catalog/mnist
Shapes3D (Burgess & Kim, 2018)	https://www.tensorflow.org/datasets/catalog/Shapes3D
ROCStories (Mostafazadeh et al., 2016)	https://cs.rochester.edu/nlp/rocstories
Tiny Shakespeare	https://huggingface.co/datasets/tiny_shakespeare
AG's News (Zhang et al., 2015)	https://huggingface.co/datasets/ag_news
Scientific Papers (Cohan et al., 2018)	https://huggingface.co/datasets/scientific_papers

Table A.1: Overview of data availability.

A.2 Code availability

Code for all simulations can be found at https://github.com/ellie-as. Specifically the following repositories were used:

Code	**Section**
https://github.com/ellie-as/generative-memory	Chapter 2
https://github.com/ellie-as/sequence-memory	Chapter 3
https://github.com/ellie-as/sleep-continual-learning, https://github.com/ellie-as/sequence-continual-learning	Chapter 4

Table A.2: Overview of code availability.

Some diagrams were created using BioRender.com.

Appendix B

Supplementary results

B.1 Chapter Two

Figure B.1 shows results for the 18 remaining Deese-Roediger-McDermott task word lists not shown in Figure 2.8. As in the human data, lure words are often but not always recalled when the model is presented with 'id_n'. The model also forgets some words, and produces additional semantic intrusions. See Methods for further details.

Figure B.2 shows that latent representations support few-shot learning better than intermediate representations extracted from the encoder or the 'sensory' image features. Decoding accuracy is measured by training a support vector machine to classify the central object's shape, varying the input features and the amount of data, and evaluating the resulting model on a held-out test set. The intermediate features tested are the outputs of four convolutional layers in the encoder, flattened to one-dimensional vectors.

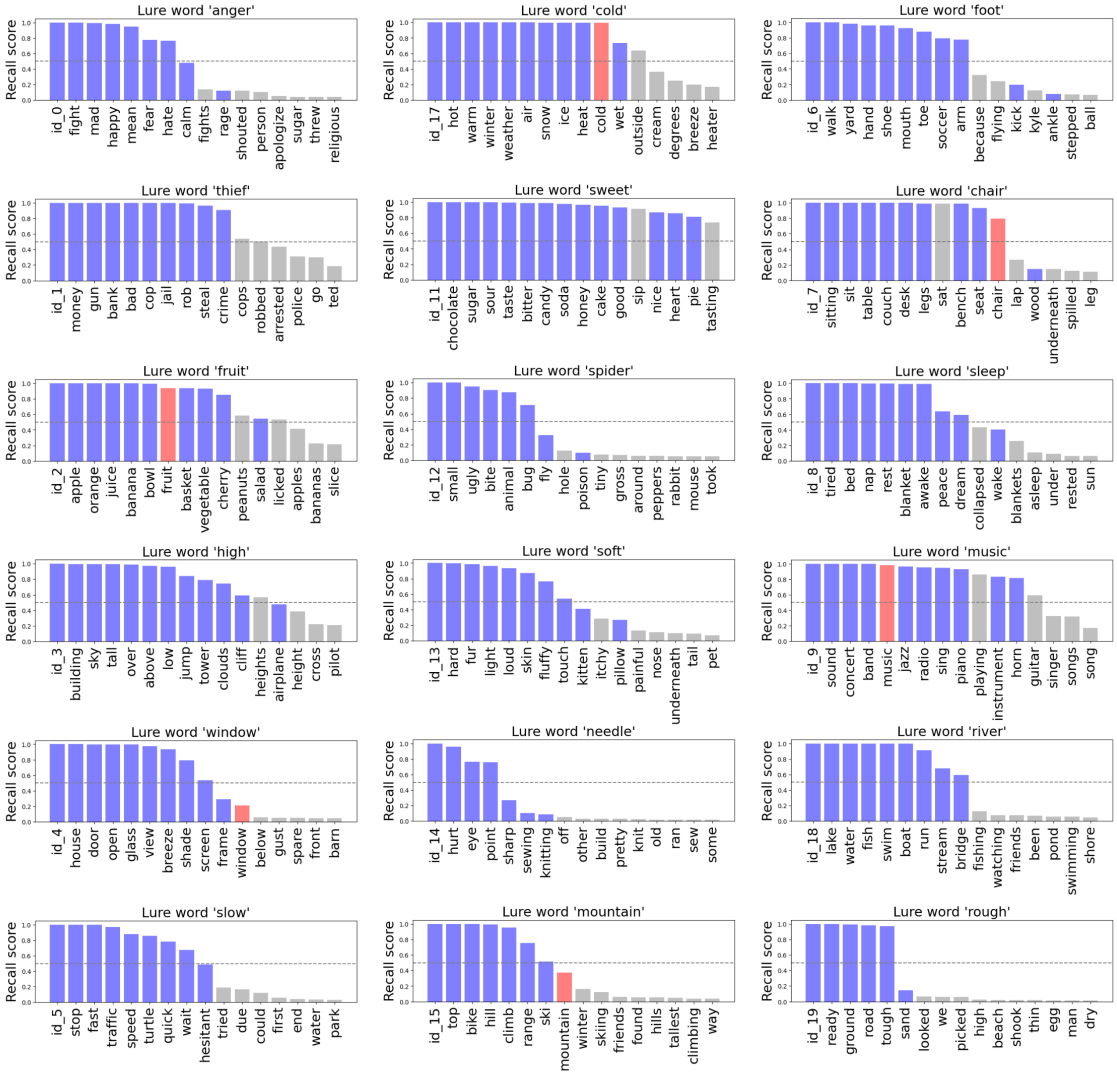

Figure B.1: Additional results for the Deese-Roediger-McDermott task. In the extended model, gist-based semantic intrusions arise as a consequence of learning the co-occurrence statistics of words. First the VAE is trained to reconstruct simple stories (Mostafazadeh et al., 2016) converted to vectors of word counts, representing background knowledge. The system then encodes the lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent representation of the word list. In each plot, recalled stimuli when the system is presented with 'id_n' are shown, with output scores treated as probabilities so that words with a score of above 0.5 are recalled. Words from the stimulus list are shown in blue, and lures in red.

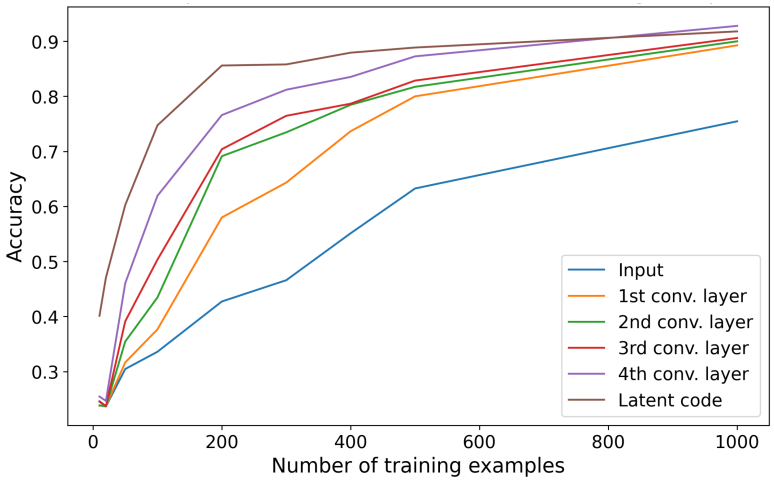

Figure B.2: Latent representations support few-shot category learning. The accuracy of an object shape classifier on a held-out test set is shown for different amounts of training data, with different layers of the VAE as input features. The classifier is a simple support vector machine as in Figure 2.2a.

B.2 Chapter Three

In Section 3.3.6, I suggest retrieval augmented generation (Lewis et al., 2020) may be a potential model of hippocampal-neocortical interactions consistent with the broader proposal in this thesis. In particular, it could help explain how planning and inference based on recent memories are possible, before those memories have been assimilated into the generative network. This can be illustrated more vividly with stories (noting that this is a very simplified version of the extended sequential model, where a neocortex-generated gist represents the 'predictable conceptual features' of a memory, and the 'unpredictable details' are not simulated). This section discusses a toy example of the encoding and recall of narratives. See Figure 3.12 for a diagram showing the steps involved, and Table B.1 for examples of encoded and recalled stories.

Ten stories from the ROCStories dataset (Mostafazadeh et al., 2016) are encoded in the system. First the generative model (GPT-3.5; Brown et al., 2020) generates a gist for each story. This is simply done by prompting the model with the story

followed by 'One line summary:'. Then each gist is stored in the hippocampus; in reality this would be accompanied by unpredictable details but this is not simulated explicitly. To make the hippocampus 'searchable' each gist is encoded together with an embedding capturing its meaning (for simplicity the vectors are also obtained from the OpenAI API, but any sentence embedding model could be used).

Then given a query as the input to recall, the neocortex 'searches' the hippocampus for relevant traces. This is simulated by generating an embedding for the query, e.g. 'What happened to Melody?' in the first example in Table B.1, and then by retrieving the traces with the nearest embedding to the query embedding. However this could be done in many ways, e.g. by finding traces which feature the words in the query.

Finally, the generative network representing the neocortex (GPT-3.5 as before) produces an 'answer' conditioned on the retrieved hippocampal trace(s). This involves constructing a prompt that includes the retrieved traces together with the query. Note that in this simulation, only one trace is retrieved per query. But this mechanism is scalable to cases where many traces are relevant to the query, and is thus a potential model of inference and planning based on multiple recent memories as well as a single memory.

Original event	Gist from HPC	Neocortex-generated answer
Melody's parents surprised her with a trip to the big aquarium. Melody took a nap during the two hour car ride to the aquarium. When they arrived, Melody was energetic and excited. At the aquarium Melody saw sharks, tropical fish and many others. After five hours at the aquarium, Melody and her family drove home.	Melody's parents surprised her with a trip to the big aquarium, where she saw various marine creatures and had an energetic and exciting day.	When Melody's parents surprised her, they took her to the big aquarium. At the aquarium, Melody had the opportunity to see a wide variety of marine creatures. She was likely amazed by the colorful fish swimming gracefully in the tanks, the playful dolphins performing tricks, and the majestic sea turtles gliding through the water. The aquarium would have provided an educational and interactive experience for Melody, allowing her to learn about different species of marine life and their habitats. It is also possible that Melody had the chance to touch some of the marine creatures, such as starfish or stingrays, in a touch tank exhibit. Throughout the day, Melody would have been filled with excitement and energy as she explored the different exhibits and witnessed the beauty and wonder of the underwater world.
John was a pastor with a very bad memory. He tried to memorize his sermons many days in advance but to no avail. He decided to learn to sing to overcome his handicap. He then made all his sermons into music and sang them on Sundays. His congregation was delighted and so was he.	A pastor with a bad memory learns to sing his sermons and delights his congregation.	John, the pastor with a bad memory, faced a unique challenge in delivering his sermons to his congregation. Due to his forgetfulness, he often struggled to remember the content of his sermons, leading to moments of confusion and frustration. However, John was determined to find a solution that would not only help him remember his sermons but also engage and delight his congregation. After much contemplation and experimentation, John discovered a creative approach to overcome his memory issues. He decided to start singing his sermons instead of delivering them in a traditional spoken format. This innovative idea not only allowed him to remember the content of his sermons but also added a new level of excitement and entertainment for his congregation. John began incorporating music into his sermons, composing catchy tunes that conveyed the messages he wanted to share.

Table B.1: Narrative examples of retrieval augmented generation as a model of hippocampal-neocortical interaction. The 'Original event' column gives the story 'experienced' by the system. The 'Gist from HPC' column gives the gist generated by the neocortex and stored in the hippocampus. The 'Neocortex-generated answer' column gives the result when the most relevant gist is retrieved from the hippocampus and used to condition the generative network (GPT-3.5; Brown et al., 2020).

Appendix C

Further model details

C.1 Variational autoencoders

This section provides a more detailed description of the variational autoencoders (VAEs) used in Chapter Two. See Section 1.6.3 for a more general discussion of VAEs.

The VAEs in these simulations use convolutional layers to better encode and decode image features. Convolutional layers learn sliding windows that scan the image for a relevant feature, outputting a stack of feature maps (LeCun et al., 1989). Applying such a layer to the output of a preceding convolutional layer has the effect of finding higher-level features in the stacked feature maps, i.e. if the first convolutional layer learns to identify simple features such as lines at different orientations, the second convolutional layer might learn features consisting of combinations of lines.

A large VAE was used for the Shapes3D dataset (containing RGB images of size 64x64 pixels), and a small VAE was used for the MNIST dataset (containing greyscale images of size 28x28 pixels). In the large model's encoder, four convolutional layers gradually decrease the width and height of the representation and increase the depth (as is standard when using convolutional neural networks to encode images), followed

by a pooling layer and dense layers to represent the mean and log variance of the latent representation. In addition, a dropout layer immediately after the input layer is added to improve the denoising abilities of the model (Srivastava et al., 2014). In the decoder, four convolutional layers alternate with up-sampling layers to increase the width and height of the representation and decrease the depth. The smaller VAE used for the MNIST simulations has a latent dimension of 20, and a reduced architecture with fewer convolutional layers for efficiency (specifically, there are two convolutional layers in the encoder and two transposed convolutional layers in the decoder).

The following list describes the sequence of operations within the large VAE's encoder network, using the layer names from the TensorFlow Keras API (Abadi et al., 2016) (see also Figure C.1):

1. Input layer for arrays of shape (n, 64, 64, 3), representing n 64x64 RGB images
2. Dropout layer with a dropout rate of 0.2 (during training, dropout randomly sets a fraction of the input units to 0 at each step, reducing overfitting and encouraging robustness)
3. Conv2D layer with 32 filters (i.e. convolutional windows, or feature detectors) and kernel size of 4 (i.e. windows of 4x4 pixels)
4. Batch normalisation layer (batch normalisation is a common technique which computes the mean and variance of each feature in a mini-batch and uses them to normalise the activations)
5. LeakyReLU activation layer (LeakyReLU is an activation function that is a variant of the Rectified Linear Unit, ReLU)
6. Conv2D layer with 64 filters and kernel size of 4
7. Batch normalisation layer
8. LeakyReLU activation layer
9. Conv2D layer with 128 filters and kernel size of 4

10. Batch normalisation layer
11. LeakyReLU activation layer
12. Conv2D layer with 256 filters and kernel size of 4
13. Batch normalisation layer
14. LeakyReLU activation layer
15. Global average pooling 2D layer
16. Dense layer to produce the mean of the latent vector
17. Dense layer to produce the log variance of the latent vector (in parallel with the layer above)
18. Custom sampling layer that samples from the latent space, with the mean and log variance layers as inputs

The same information for the decoder network is as follows:

1. Input layer for arrays of shape (n, latent_dimension), where latent_dimension is 20 in these results, representing n latent vectors
2. Dense layer that expands the latent space to a size of 4096
3. Reshape layer to reshape the input to a 4x4x256 tensor
4. Upsampling2D layer with a 2x2 upsampling factor
5. Conv2D layer with 128 filters and kernel size of 3
6. Batch normalisation layer
7. LeakyReLU activation layer
8. Upsampling2D layer with a 2x2 upsampling factor
9. Conv2D layer with 64 filters and kernel size of 3
10. Batch normalisation layer

11. LeakyReLU activation layer
12. Upsampling2D layer with a 2x2 upsampling factor
13. Conv2D layer with 32 filters and kernel size of 3
14. Batch normalisation layer
15. LeakyReLU activation layer
16. Upsampling2D layer with a 2x2 upsampling factor
17. Conv2D layer with 3 filters and kernel size of 3

C.2 Autoregressive sequence models

This section provides a few more details about the training or fine-tuning of models like GPT-2 (Radford et al., 2019).

The main text gives the intuition for attention, but to be more precise the equation is given below:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In this equation, Q, K and V represent the query, key, and value matrices, respectively. These are derived from the input data. d_k represents the dimensionality of the keys (and queries), and is used to scale the dot products in a way that leads to more stable gradients. The softmax function is applied to ensure the weights sum up to 1. When the resulting term is matrix multiplied by V, the result is the relevance-weighted sum of the vector representations in V, for each token in V. In other words, the attention mechanism enriches a token's representations with a relevance-weighted sum of the other tokens' representations. See Alammar (2018) for a detailed description of this process.

Self-attention is a special case of attention in which elements of a sequence 'attend to'

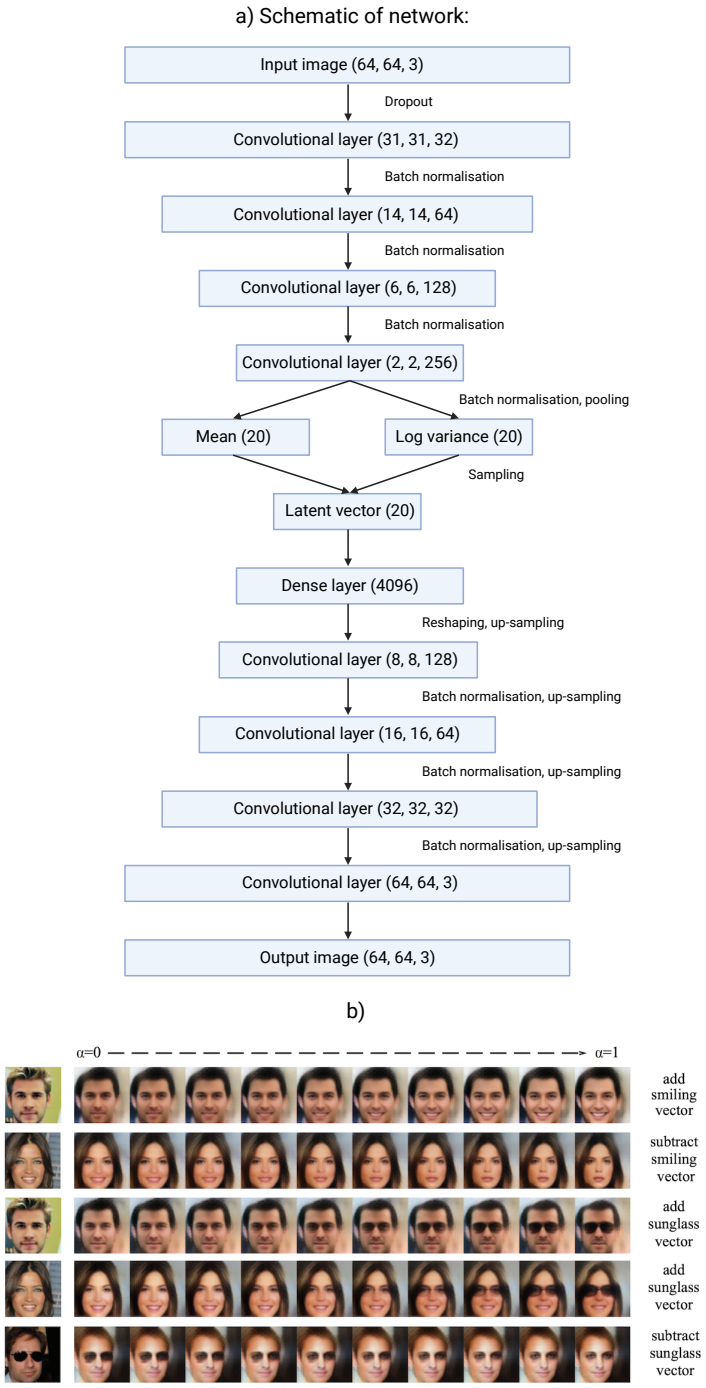

Figure C.1: Additional model details. a) Variational autoencoder architecture. Trainable layers (plus the input, output, and sampled latent vector) are shown in boxes, along with the dimensions of their outputs, and non-trainable operations such as activation functions, batch normalisation, and upsampling are shown as annotations. See the text for more details. b) Figure adapted from Hou et al. (2017) with permission, showing the effect of adding and subtracting a proportion α of various different vectors in the latent space of their VAE. (Diagrams were created using BioRender.com.)

the sequence they are a part of. Specifically, 'masked' self-attention is used, meaning that the representation of a token in an attention block only 'attends to' preceding tokens. Previous models like BERT (Devlin et al., 2018) perform much less well at text generation partly because their attention is not masked in this way.

Transformer blocks come in different varieties. GPT-2 (Radford et al., 2019) is a decoder-only model consisting of a stack of transformer decoder blocks. Each such block consists of a masked self-attention layer followed by a feedforward layer. In a nutshell, inference with GPT-2 works as follows: First an embedding for each token is obtained from a learned embedding matrix. A key point is that position embeddings representing the tokens' positions in the sequence are added to the token embeddings (these are required because self-attention is permutation-invariant, so otherwise information about the order of items in the sequence would be lost). Then the sequence of embeddings passes through a series of transformer decoder blocks, each further enriching the token representations so that they come to capture the meaning of the text. See Alammar (2019) for helpful illustrations of the stages of processing in GPT-2.

GPT-2 (Radford et al., 2019) comes in several sizes, with the number of transformer blocks depending on the size of the model. The small variant has 12 transformer blocks, the medium variant has 24 transformer blocks, and the large variant has 36 transformer blocks.

C.3 Asymmetric modern Hopfield networks

In Section 3.1.2, I discuss several options for how sequential traces might be stored in the initial hippocampal network. One straightforward option is the asymmetric variant of a modern Hopfield network outlined below.

As mentioned earlier, one can combine the concept of the modern asymmetric Hopfield network (Chaudhry et al., 2023; Millidge et al., 2022), with return projections from memory to feature units being the next states as illustrated in Figure 3.1, with a

state representation that captures the history of previous states.

In other words, we want to store patterns X and 'next patterns' X_{next} instead of just patterns, and change the update rule accordingly:

$$\xi_{next} = X_{next} softmax(\beta X^T \xi)$$

To show how the preceding 'context' is captured, consider the example of encoding sentences, i.e. sequences of characters. Each state can be represented by a vector of length equal to the number of symbols, with one at the index of the current symbol, plus the previous state multiplied by some decay factor. In other words, we want a vector representation $\mathbf{x_i}$ of the i$^{\text{th}}$ state which is a sum of the current symbol's vector $\mathbf{v}$ and activity from the previous state $\mathbf{x}_{i-1}$ multiplied by a decay rate λ (noting that the states are then normalised to unit length):

$$\mathbf{x}_i = \frac{\mathbf{v}_i + \lambda \cdot \mathbf{x}_{i-1}}{\|\mathbf{x}_i\|}$$

This has the benefit of still being compatible with one-shot learning, and works in initial testing for a range of sequences represented as characters, e.g. the model of HPC can encode and retrieve sentences (although retrieval performance goes down with sentence length). As noted above, our sequences consist of a single symbol at any moment in time, but other sequences like frames in a video consist of complex representations at each time step. In the latter, the decaying activity from previous states would be hard to distinguish from the current state, so a more complex solution like the temporal predictive coding network of Tang et al. (2023) may be required. In addition the memory capacity may not scale as well as the more complex approaches.

As with all modern Hopfield network (Ramsauer et al., 2020) variants, the inverse temperature, β, is a key parameter determining the network's behaviour. In this case a high value of β would be desirable to avoid composite states. The softmax function

is given below; in the limit of high β, the output of this function becomes a vector with one at the index of the maximum and value and zero elsewhere:

$$softmax(z_i) = \frac{e^{\beta z_i}}{\sum_{j=1}^{K} e^{\beta z_j}}$$

One additional detail is that a special token can be used to represent the start and end of a sequence (e.g. '[' and ']' in my code). This means that 'replay' can be initiated by giving the network the start of sequence token. Since all sequences begin with this character, the dot product of the input and all possible next items is the same, so if a 'random tiebreak' is implemented at the softmax stage, this corresponds to random sampling from the bank of memories.

C.4 Predictive coding networks

This section provides some more details on predictive coding networks, a more biologically plausible alternative to the generative networks used in the simulations.

Bogacz (2017) introduces predictive coding with a simple toy problem involving perceptual inference: an animal receiving noisy sensory data, a single value of light intensity u, must infer the size of an object, v. The variable u is noisy, so is represented as normally distributed with a mean of $g(v)$, which gives the expected light intensity for an object of size v. The variable v is also normally distributed, with a mean of v_p, which represents the prior for v. Note that the animal has two sources of uncertainty in this problem: it doesn't know the true value of v but must infer it from u, and in addition the estimate of u is noisy given v. The problem then is to estimate the most likely size of the object, ϕ, from the noisy observation of light intensity. That is, we want to find the value of v, ϕ, that maximises $p(v|u)$:

$$p(v \mid u) = \frac{p(u \mid v)p(v)}{p(u)}$$

To find the most likely size ϕ, we want to maximise the numerator, or equivalently the logarithm of the numerator, a quantity F:

$$F = \ln p(\phi) + \ln p(u|\phi)$$

What is the relationship between the quantity F we want to maximise and the Helmholtz free energy? In mathematical terms, Bogacz (2017) shows that 'for certain assumptions the negative free-energy is equal (modulo a constant) to the function F' (Free-energy section), so that maximising F minimizes the free energy, which can be thought of very loosely as 'surprise'.

Bogacz (2017) defines two quantities representing prediction error. ϵ_p represents the difference between the inferred object size and the prior, while ϵ_u represents the difference between the observed light intensity and the expected value given ϕ (while Σ_u gives the variance of the light intensity and Σ_p the variance of the size):

$$\epsilon_p = \frac{\phi - v_p}{\Sigma_p}, \epsilon_u = \frac{u - g(\phi)}{\Sigma_u}$$

We want to take the derivative of F with respect to ϕ to find its maximum, at which the derivative is zero. Bogacz (2017) proceeds to show that the following four equations converge to find the maximum of ϕ:

$$\dot{\phi} = \epsilon_u g'(\phi) - \dot{\epsilon}_p, \dot{\epsilon}_p = \phi - v_p - \Sigma_p \epsilon_p, \dot{\epsilon}_u = u - g(\phi) - \Sigma_u \epsilon_u$$

Bogacz (2017) also finds the derivatives of F with respect to Σ_u, Σ_p, and v_p, and these gradients can be used to update Σ_u, Σ_p, and v_p to maximise F.

So in summary, for some sensory input u, and given network weights described by Σ_u, Σ_p, v_p, and $g(\phi)$, we have a set of equations to find the values of phi, ϵ_p, and ϵ_u, based on finding the maximum of F with respect to ϕ (where maximising F is related to minimising the free energy). These values are activities at the nodes of the network.

They converge over time, with activity bouncing back and forth between the layers of the network (unlike in the standard feedforward networks used in deep learning). We also have equations for updating the parameters of the network. As Millidge et al. (2021) describe, the network minimizes prediction error, 'first through the optimization of neuronal firing rates on a fast timescale, and then the optimization of synaptic weights on a slow timescale' (Introduction).

These equations can be implemented by the network shown in Figure C.2, reproduced from Figure 3 of Bogacz (2017). Note that only the prediction error $\epsilon_u g'(\phi)$ is transmitted upwards through the hierarchy (in contrast to autoencoders, as discussed above), while the prediction $g(\phi)$ is transmitted downwards.

This toy example illustrates a local learning algorithm that involves only propagating errors from lower to higher levels, and accounts for uncertainty in both the latent representation given the sensory data, and the sensory data given the latent representation in a mathematically principled way. The authors then develop this simplified version into the full predictive coding network, by generalising to multiple input features, hidden features, and layers in a hierarchy, and by making the distribution $g(v)$, which describes the value of the sensory data u for some inferred cause v, learned rather than fixed.

Predictive coding networks' use outside of neuroscience is limited, partly because they are slower to train than more traditional learning algorithms. However, various modifications have been suggested to address this, such as incremental predictive coding networks (Salvatori et al., 2022).

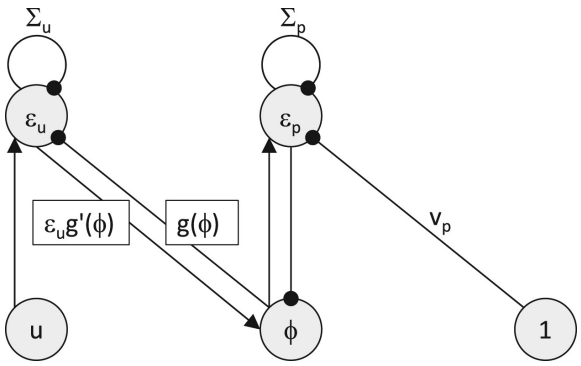

Figure C.2: A toy example of predictive coding reproduced from Figure 3 of Bogacz (2017), in which circles are nodes, arrows are excitatory connections, and lines ending in circles are inhibitory connections.