A Generative Model of Memory Construction and Consolidation

Eleanor Spens and Neil Burgess

Institute of Cognitive Neuroscience, University College London, UK

Author Note

Code for all simulations can be found at https://github.com/ellie-as/generative-consolidation. This study was not preregistered. Correspondence concerning this article should be addressed to Eleanor Spens, Institute of Cognitive Neuroscience, Alexandra House, 17-19 Queen Square, London WC1N 3AZ. Email: eleanor.spens.20@ucl.ac.uk.

We thank Tim Behrens and Brad Love for useful discussions. Funding support for this work was received from a Wellcome Principal Research Fellowship 'Neural mechanisms of memory & prediction: Finding structure in experience', a Wellcome Collaborative Award 'Organising knowledge for flexible behaviour in the prefrontal-hippocampal circuitry', and an ERC advanced grant NEUROMEM.

Abstract

Human episodic memories are (re)constructed, combining unique features with schema-based predictions, and share neural substrates with imagination. They also show systematic schema-based distortions that increase with consolidation. Here we present a computational model in which hippocampal replay (from an autoassociative network) trains generative models (variational autoencoders) in neocortex to (re)create sensory experiences via latent variable representations. Simulations using large image datasets reflect the effects of memory age and hippocampal lesions and the slow learning of statistical structure in agreement with previous models (Complementary Learning Systems and Multiple Trace Theory), but also explain schema-based distortions, imagination, inference, and continual representation learning in memory. Critically, the model suggests how unique and predictable elements of memories are stored and reconstructed by efficiently combining both hippocampal and neocortical systems, optimising the use of limited hippocampal storage. Finally, the model can be extended to sequential stimuli, including language, and multiple neocortical networks could be trained, including those with latent variable representations in entorhinal, medial prefrontal, and anterolateral temporal cortices. Overall, we believe hippocampal replay training neocortical generative models provides a comprehensive account of memory construction and consolidation.

Keywords— episodic memory, imagination, consolidation, memory distortion

1 Introduction

Episodic memory is thought to be constructive; recall is the (re)construction of a past experience, rather than the retrieval of a copy (Bartlett, 1932; Schacter, 2012). Whilst episodic memory is the autobiographical memory of an experience in its spatiotemporal context, semantic memory is factual knowledge (Tulving, 1985). Episodic and semantic memory are thought to complement each other, with the former rapidly capturing cross-modal experience via long term potentiation in the hippocampus, enabling the latter to learn statistical regularities over multiple experiences in the neocortex (Marr, 1970, 1971; McClelland et al., 1995; Teyler & DiScenna, 1986).

The standard model of systems consolidation is a simple transfer of information from the hippocampus to neocortex (Alvarez & Squire, 1994), whereas other views suggest that episodic and semantic information from the same events can exist in parallel (Nadel & Moscovitch, 1997). However, consolidation does not just change which brain regions support memory traces; it also converts them into a more abstract representation, a process sometimes referred to as semanticisation (Norman et al., 2021; Winocur & Moscovitch, 2011).

Here, we propose that consolidated memory takes the form of a generative model, trained to capture the statistical structure of stored events by learning to reproduce them. 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). The link between scene generation and recall follows from the reconstructive nature of memory, and is supported by evidence from neuropsychology that hippocampal damage 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 evidence from neuroimaging that recall and imagination involve similar neural processes (Addis et al., 2007; Hassabis & Maguire, 2007).

We model consolidation as the training of a generative model by an initial autoassociative encoding of memory, through the mechanism of teacher-student learning (Hinton et al., 2015) during hippocampal replay. Recall after consolidation has occurred is a generative process mediated by schemas, as are other forms of scene construction or imagination. Our model builds on research on the relationship between generative models and consolidation (Káli & Dayan, 2000, 2002), and on the use of variational autoencoders to model the hippocampal formation (HF) (Nagy et al., 2020; van de Ven et al., 2020; Whittington et al., 2020).

More generally, we build on the idea that the overall memory system learns schemas consistent with past experience for the reconstruction of input patterns (Hemmer & Steyvers, 2009). In this view, unpredictable aspects of experience (which cannot be reconstructed by the generative model) need to be stored in detail for further learning, while fully predicted ones do not, consistent with the idea that memory serves prediction of the future (Bein et al., 2021; Biderman et al., 2020; Schacter et al., 2007; Sherman et al., 2022).

We suggest that experience is decomposed into familiar, predictable components and novel, poorly-predicted components. 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. In other words, the autoassociative network can bind both conceptual and sensory aspects of a memory, and recall makes use of the generative network where possible prior to consolidation. This is more efficient in terms of memory storage, 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.

Table 1 shows a summary of features required in a model of memory processing, some existing models of them, and relevant sections of this paper. To expand on each point in more detail:

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

2. The semantic content of a memory becomes independent of the hippocampus over time, as shown by lesion studies (Manns et al., 2003; Squire et al., 2015; Vargha-Khadem et al., 1997), consistent with CLS.

3. However vivid, detailed episodic memory remains dependent on HF (McKenzie & Eichenbaum, 2011), consistent with multiple trace theory (Nadel & Moscovitch, 1997) (but not CLS). Encoding of new episodes also remains hippocampus-dependent.

4. Similar neural circuits are involved in recall, imagination, and episodic future thinking (Addis et al., 2007; Becker & Burgess, 2000; Hassabis & Maguire, 2007), suggesting a common mechanism for episode generation, as modelled in spatial cognition (Bicanski & Burgess, 2018).

5. After consolidation, statistical regularities can be extracted from episodic memories to inform behaviour (Durrant et al., 2011; Richards et al., 2014), and provide better support for 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).

6. However post-consolidation episodic memories are also 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. The entorhinal cortex (EC) is thought to encode latent structures underlying experiences (Constantinescu et al., 2016; Whittington et al., 2020), and other regions of the association cortex, such as the medial prefrontal cortex (mPFC), may compress stimuli to a minimal representation (Mack et al., 2020). The TEM (Whittington et al., 2020) models the extraction of latent structure in this way.

8. Gradual forgetting but not catastrophic forgetting of consolidated memory is observed (Ebbinghaus, 2013); in other words, the system displays continual learning (Hadsell et al., 2020). Several computational models explore how this may be achieved by generative replay (Káli & Dayan, 2002; Shin et al., 2017; Van de Ven & Tolias, 2018).

9. Limited hippocampal capacity must be used efficiently (Benna & Fusi, 2021). Novelty switches the hippocampus into encoding mode (Hasselmo et al., 1996), while new events consistent with existing schemas are consolidated more rapidly (Tse et al., 2007). Activity in the hippocampus can reflect prediction error or mismatch novelty (i.e. the difference between prediction based on the existing 'model' of the world and reality) (Chen et al., 2011; Kumaran & Maguire, 2006).

10. Memory traces in hippocampus bind together a mixture of sensory and conceptual elements, with the latter encoded by concept cells (Quiroga, 2012). Few models explore how this could happen.

Table 1: Key features required in a model of memory processing

No.	Feature	Existing models	Section
1	'One-shot' encoding is followed by gradual consolidation	CLS (McClelland et al., 1995) and most subsequent models	3.1
2	Semantic content of memory becomes hippocampus-independent	CLS (McClelland et al., 1995) and most subsequent models	3.3, 3.7
3	But episodic recollection remains HF-dependent	Multiple trace theory (Nadel & Moscovitch, 1997)	2.1
4	Common mechanism for episode generation	Bicanski and Burgess (2018)	3.6
5	Consolidation promotes inference and generalisation	TEM (Whittington et al., 2020); Kumaran et al. (2016)	3.6, 3.9
6	Consolidation increases schema-based distortions	Nagy et al. (2020)	3.5
7	Memory processing involves gist extraction from experience	TEM (Whittington et al., 2020)	2.1
8	Catastrophic forgetting of consolidated memories does not occur	Káli and Dayan (2004), Van de Ven and Tolias (2018), Singh et al. (2022) and others	2.3, 3.8

No.	Feature	Existing models	Section
9	Novel stimuli encoded preferentially for efficiency	Benna and Fusi (2021), TEM (Whittington et al., 2020)	2.2, 3.4
10	Memory traces bind together sensory and conceptual elements	-	2.2, 3.4

2 The model

2.1 Consolidation as teacher-student learning

Our model simulates how 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 (modelled as one-shot memorisation in a modern Hopfield network; Ramsauer et al., 2020). Then generative networks (modelled by variational autoencoders; Kingma and Welling, 2013, 2019) take over, having been trained on replayed representations from the hippocampal autoassociative network, see Appendix for details. This makes the memory more abstracted, more supportive of generalisation and relational inference, and also more prone to gist-based distortion. The generative networks can be used to reconstruct (for memory) or construct (for imagination) sensory experience via the EC and its return projections to neocortex, or to support semantic memory and inference directly from their latent variable representations, see Figure 1.

Before consolidation, a modern Hopfield network (Ramsauer et al., 2020) - an autoasso-

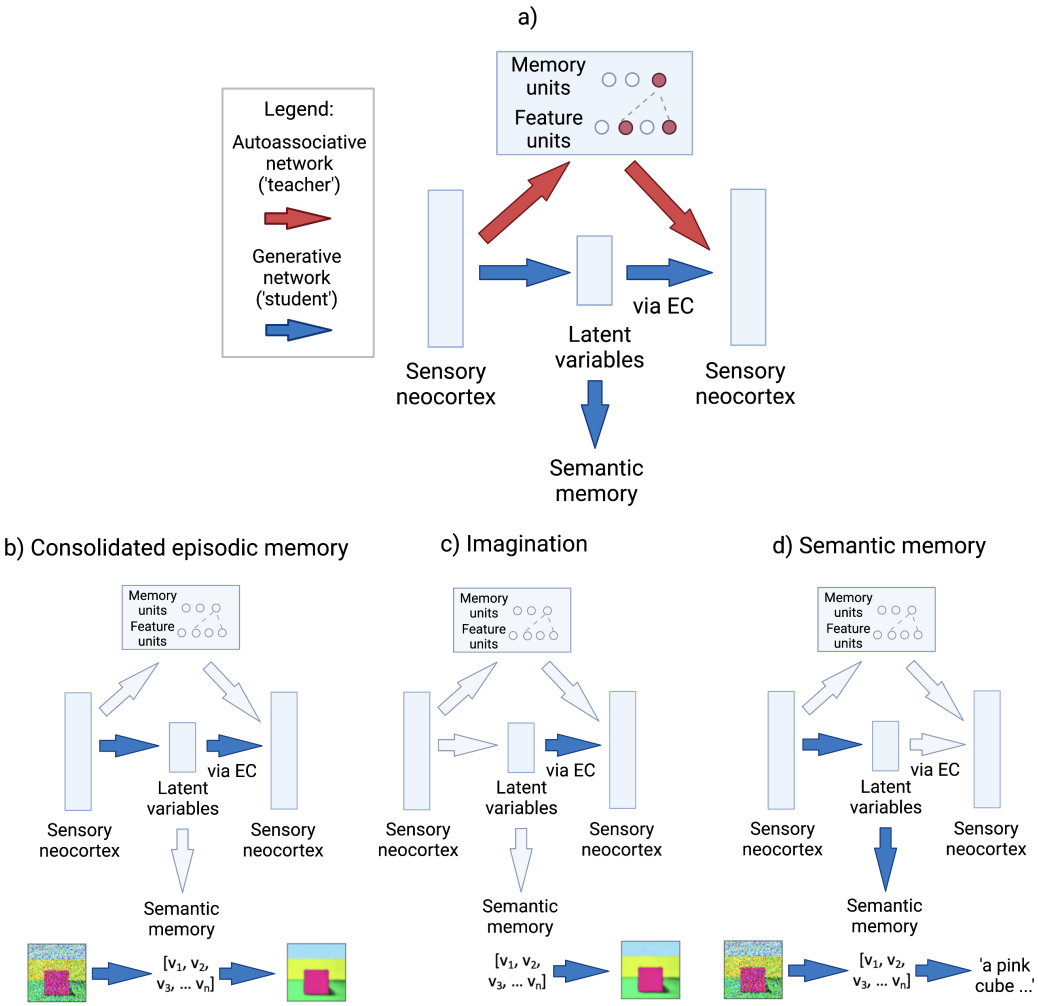

Figure 1: a) Architecture of the basic model: First the hippocampal formation rapidly encodes an event, modelled as one-shot memorisation in an autoassociative network (a modern Hopfield network). Then generative networks are trained on replayed representations from the autoassociative network, learning to reconstruct memories by capturing the statistical structure of experienced events. b) Episodic memory after consolidation: a partial input is mapped to latent variables whose return projections to sensory neocortex via EC then decode these back into a sensory experience. c) Imagination: latent variables are decoded into an experience by the EC and return projections to neocortex. d) Semantic memory: a partial input is mapped to latent variables, which capture the 'key facts' of the scene. (The bottom rows of parts b-d illustrate these functions in a model that has encoded the Shapes3D dataset into latent variables $[v_1, v_2, v_3 \ldots v_n]$.)

ciative network which improves on traditional Hopfield networks - encodes the memory. Two properties of this network are particularly important for our simulation: firstly, memorisation can occur with only one exposure, and secondly, random inputs to the network retrieve attractor states, allowing sampling from the whole set of stored patterns during replay. We consider the biological interpretation of the modern Hopfield network as feature units and memory units suggested by Krotov and Hopfield (2020) (see Appendix).

After consolidation, generative networks also encode the information contained in memories. Specifically, we use variational autoencoders (VAEs), so that the most compressed layer of the autoencoder represents a set of latent variables which can be sampled from (Kingma & Welling, 2013, 2019) (see Appendix for details). We suggest that there are generative networks with latent variables in entorhinal, medial prefrontal, and anterolateral temporal cortices, but in each case return projections to sensory neocortex via entorhinal cortex (EC) are required to decode latent variables into sensory experiences.

In general, reliance on the generative networks increases over time as they learn to reconstruct a particular event, and the hippocampal encoding decays. During perception, the generative model provides an ongoing estimate of novelty from its reconstruction error. Memories that are very consistent with previous experience do not need to be encoded in detail in the autoassociative 'teacher' network (see Bein et al., 2021; Biderman et al., 2020; Schacter et al., 2007; Sherman et al., 2022). (How the model deals with events composed of both predictable and unpredictable elements is addressed in the next section.)

Teacher-student learning (Hinton et al., 2015) has clear relevance to the transfer of memories from one neural network to another during consolidation (Sun et al., 2021). Based on this concept, we use outputs from the autoassociative network to train the generative network in our simulations. To do so, we give random noise as an input to the modern Hopfield

network, then use just the outputs of the network 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). In other words, random inputs to the hippocampus result in the reactivation of memories, and this reactivation results in consolidation. Attributes such as reward salience may also influence which memories are replayed, but this is not modelled here for simplicity (Igata et al., 2021).

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 in the generative network neocortical inputs project to association cortex, which then projects back to neocortex via the hippocampal formation. The association cortex encodes latent variable representations of experience; the entorhinal (EC), medial prefrontal (mPFC), and anterolateral temporal cortices are all prime candidates for this role.

Firstly, the EC is the main route between the hippocampus and neocortex, and is also 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, the mPFC is highly connected to the HF and plays 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). It is thought to encode schemas (Ghosh & Gilboa, 2014), and has been implicated in transitive inference (Koscik & Tranel, 2012) and the integration of memories (Spalding et al., 2018). Furthermore, Mack et al. (2020) suggest the mPFC performs dimensionality

reduction on incoming data, compressing representations to remove irrelevant features as learning progresses. Thirdly, the anterior and lateral temporal cortices associated with semantic memory (Chan et al., 2001) and retrograde amnesia (Bright et al., 2006) are likely to contain latent variable representations capturing semantic structure. This third generative network might correspond to the 'anterior temporal network' associated with semantic dementia in the model proposed by Ranganath and Ritchey (2012), while the first network (between visual and entorhinal cortices) might correspond to their 'posterior medial network', and to the network mapping between visual scenes and allocentric spatial representations in Becker and Burgess (2000), Bicanski and Burgess (2018) and Byrne et al. (2007). Therefore we suggest that these regions are involved in the extraction of underlying structure from experience into several generative networks.

2.2 Episodic memory, prediction error, and concepts

In most models of consolidation, fine-grained sensory representations update coarse-grained conceptual representations, e.g. the image of a particular dog updates the concept of a dog. Modelling hippocampal representations as sensory is a reasonable simplification, which we make in simulations of the 'basic' model presented in Figure 1. However, in reality, memories bind together representations along a spectrum from coarse-grained and conceptual to fine-grained and sensory. Consider the example of a person remembering the previous day out at the beach. The hippocampal encoding of this experience is likely to bind together high-level concepts like 'beach' and 'sea' along with sensory representations like the melody of an unfamiliar song, or sight of a particular sandcastle. This also fits with the evidence for concept cells in hippocampus (Quiroga, 2012).

Furthermore, encoding every sensory detail in hippocampus would clearly be inefficient

(many elements would be redundant, e.g. encoding the presence of sand on a beach is unnecessary when the concept of 'beach' already predicts it). In other words, memories often share an underlying structure; an efficient system should take advantage of shared structure, i.e. the correlations between memories, to encode only what needs to be encoded. In line with these findings, we suggest that predictable elements are encoded as concepts in the autoassociative network, while unpredictable elements are encoded in a more detailed and veridical form.

As shown in Figure 2, suppose someone sees an unfamiliar animal in a forest for the first time. Much of the event might be consistent with an existing forest schema, but certain aspects – say the sight of an unfamiliar animal – would be novel. We propose that during perception, the reconstruction error per memory element of the experience is calculated by the existing generative model. For each memory element, if the reconstruction error is high, it is encoded in the autoassociative network, in fitting with evidence that activity in the hippocampus can reflect prediction error (Chen et al., 2011; Kumaran & Maguire, 2006). In other words, each pattern is split into a predictable component (approximating the VAE's prediction for the pattern), plus an unpredictable component (those parts of the memory with a high prediction error).

But storing only the unpredictable elements in the autoassociative network is not enough; for the system to recall or sample a memory, there must be a way to add the predictable elements 'back in' to the memory without any external stimuli. An obvious suggestion is to store the latent variables in addition to the unpredictable elements, however the latent variables are not stable - they are constantly updated as the generative network learns. If some latent variables are stored in the autoassociative network, and then some time later the VAE's decoder attempts to decode them into an experience, the result might

be very poor, as the decoder's weights have changed in this time; this is a limitation of previous models (Benna & Fusi, 2021). It does not seem plausible for the latent variable representations to be constantly updated either.

We suggest that to make these representations more stable, the autoassociative network stores unpredictable elements, and also concepts corresponding to the predictable elements. Projections from the latent variables can classify attributes with only a small amount of training data (see Section 3.3); we suggest there could be a two-way mapping between latent variables and concepts, which supports categorisation of incoming experience as well as semantic memory. In other words, storing a concept (e.g. the label 'three' in the case of MNIST) is more stable than storing the latent variables (e.g the *vector* corresponding to 'three') directly, as these are constantly updated. This approach is simulated in the 'extended model', see Figure 2.

So in summary, we suggest concepts and unpredictable elements are both encoded in the autoassociative network. This could fit with evidence for a functional gradient along the long axis of the hippocamus; the posterior hippocampus may encode more detailed, more perceptual representations of episodic memories, and the anterior hippocampus may encode less detailed, more conceptual representations (Moscovitch et al., 2016). This produces a much sparser, less correlated vector than storing every element in detail, which would increase the storage capacity of the network, as noted by Benna and Fusi (2021). The concepts can then be decoded back into experience via their latent variables, enabling conceptual and sensory data to be recombined in recall.

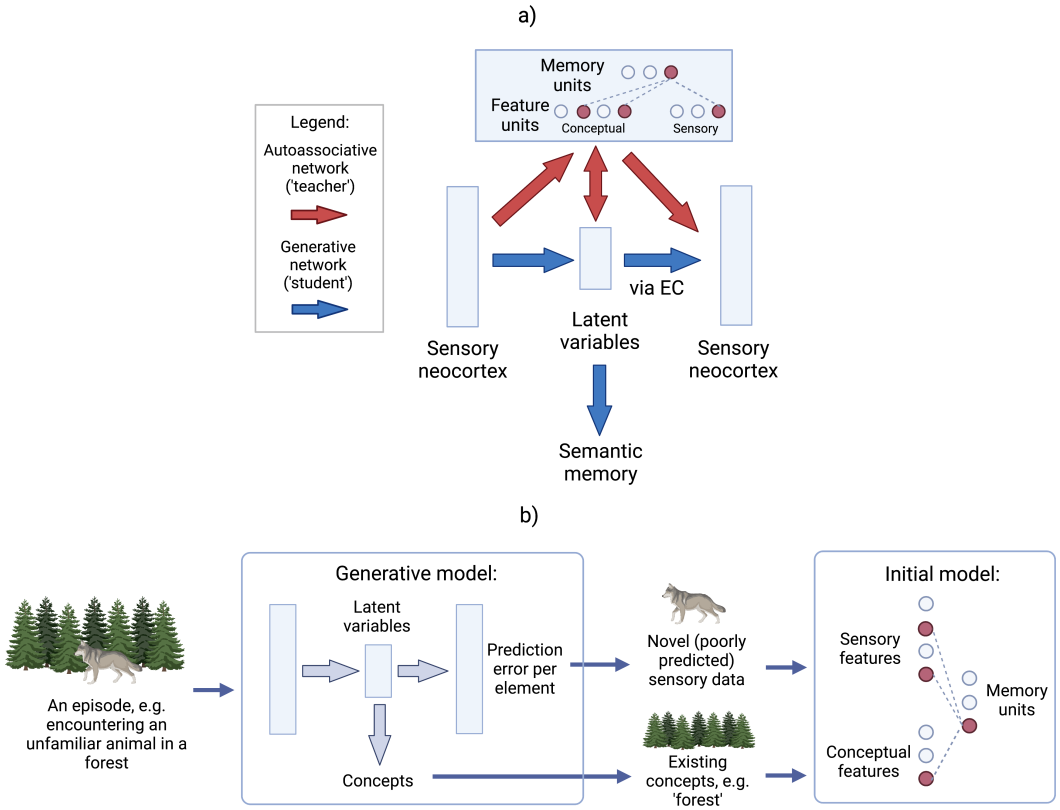

Figure 2: The extended version of the model. a) Each memory is initially encoded as the sum of a predictable and an unpredictable component. The predictable component consists of schemas / concepts captured by the latent variables of the generative network, and the unpredictable component consists of parts of the stimuli that were poorly predicted by the generative network. A Modern Hopfield Network model (in red) can have both conceptual and sensory feature units, where the former may correspond to concept cells. Memories may eventually become fully dependent on the generative model, but 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. b) An illustration of encoding in the extended model.

2.3 Continual representation learning and generative replay

Continual learning is the ability to learn, or memorise, a series of tasks, or items, sequentially, without the occurrence of catastrophic interference (Hadsell et al., 2020). Catastrophic interference or forgetting occurs when newly learned information overwrites previously learned information in a neural network, leading to poor performance on the previously learned task (McCloskey & Cohen, 1989). Gradual forgetting is expected in computational models of memory, but catastrophic forgetting is a much more dramatic decline which is not observed in healthy human memory. During consolidation, hippocampal replay allows schemas in association cortex to incorporate new information in a way that does not destroy previous learning, but the mechanism for this is not understood.

Whilst most research has been on classification tasks, and more recently reinforcement learning ones, continual learning also has clear relevance to unsupervised representation learning. Specifically the generative latent variable model trained by consolidation needs to retain its ability to represent a certain type of episode even when it has not been experienced for months or years. (Catastrophic forgetting is also an issue for models of imagination - if the model cannot encode and decode a type of episode to and from latent variables, it cannot generate similar episodes either.) Catastrophic forgetting is alleviated if unexpected memories are stored preferentially, as in our proposal - when a category of memory begins to decay, examples start to be stored again due to increased prediction error, refreshing the category's representation. However this does not fully solve the problem, as some memories will be of rare or unique categories, for which 'reminders' are received rarely or not at all.

An obvious way to avoid this is to interleave examples from the different classes in the training data, e.g. by shuffling a balanced dataset of examples (as in CLS; McClelland

et al., 1995). But this requires retaining all the training data, which is not biologically plausible; the hippocampus is not thought to store an exact copy of every memory from its entire lifetime. 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).

Generative replay (Shin et al., 2017; van de Ven et al., 2020) is an example of the rehearsal-based category, most often discussed in the context of supervised learning (e.g. the training of a classifier on a set of classes that expands over time). A generative model learns to generate training data, so that sampling from the model generates representative examples from previous phases of training. These are then labelled by the classifier and mixed into the training data for the next phase. Generative replay is a powerful way to avoid catastrophic forgetting in another classifier, providing you have a generative model that does not suffer from catastrophic forgetting itself. If it does, generative replay just moves the problem elsewhere. To avoid this, we suggest that the generative model is additionally trained on its own generated data, reducing catastrophic forgetting in the generative model itself (as in Van de Ven and Tolias, 2018).

3 Results

3.1 Modelling encoding and recall

In the following simulations, images represent events. Results are shown across four separate image datasets: MNIST (LeCun et al., 2010), Fashion-MNIST (Xiao et al., 2017), Shapes3D (Burgess & Kim, 2018), and Symmetric Solids (Murphy et al., 2021). (See Ap-

pendix for further details.) Note that unless stated otherwise, in subsequent simulations we use the simplified version of the model for clarity, in which the autoassociative network's representations are sensory.

Each new event is initially encoded as an autoassociative trace in the hippocampus, which we simulate with a modern Hopfield network. Importantly, encoding is one-shot. We model recall as (re)constructing a memory from a partial input. In the case of image data, this can be conveniently modelled by presenting a network with a partial version of the image.

Firstly, we simulate recall in the autoassociative network. The network memorises a set of images, representing events, as described above. When the network is given a partial input, it retrieves the closest attractor state, i.e. the closest memory (activity propagates through the network, and the values of the nodes are adjusted to minimise the energy function). Even when the network is given random noise, it retrieves memories (see Figure 3).

Secondly, we simulate recall in the generative network. As shown in Figure 4, a generative network was trained on the reactivated memories from the autoassociative network. (One variational autoencoder was trained per dataset from the corresponding modern Hopfield network's outputs.) When presented with a partial version of an item from the training data, the generative network is able to reconstruct the original. It is also able to generate new 'events', as exemplified by the interpolation shown in Figure 5 (see also Section 3.6).

In the current model, 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, we have not

simulated decay, deletion or capacity constraints in the autoassociative memory part of the model. In these simulations, the main cause of the symptoms of forgetting would be interference from new memories in the generative model.

3.2 Memory and prediction error

In our model, memories that are very consistent with previous experiences do not need to be encoded in the hippocampal autoassociative network; they can be recalled by the generative model straight away. Thus the reconstruction error of the generative model is used to assess the novelty of events during perception, and detection of novelty switches the autoassociative network into encoding mode (see also Hasselmo et al., 1996). To start with we simulate the basic model in which each memory is consolidated all in one go (as opposed to the extended model).

We ran the simulations with the system first 'experiencing' the MNIST dataset to represent existing memories. We then exposed the system to new images from three datasets; those with a reconstruction error below a certain threshold were not encoded in the modern Hopfield network, since the generative model could recall them adequately already. Figure 6 shows a histogram of reconstruction errors across three datasets, for a VAE trained on items from the MNIST dataset. Unseen items from the MNIST dataset have a low reconstruction error, but items from the two novel datasets, KMNIST and Fashion-MNIST, have higher reconstruction errors. When a threshold for encoding in the modern Hopfield network is set, a veridical encoding in the autoassociative network is required for a small minority of MNIST images but a majority of KMNIST and Fashion-MNIST images. (See Figure 6.)

Figure 3: Random noise inputs to a modern Hopfield network (upper row) reactivate its memories (lower row). Results are shown across four datasets: a) MNIST, b) Fashion-MNIST, c) Shapes3D (converted to black and white), and d) Symmetric Solids. One modern Hopfield network was used per dataset.

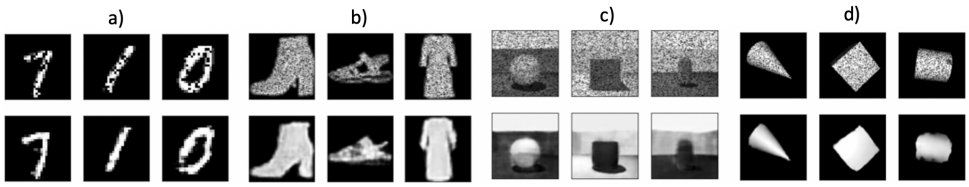

Figure 4: A generative model (a variational autoencoder) can recall images (lower row) from a partial input (upper row), following training on reactivated memories from a modern Hopfield network. Results are shown across a) MNIST, b) Fashion-MNIST, c) Shapes3D (converted to black and white), and d) Symmetric Solids. One generative model was trained per dataset.

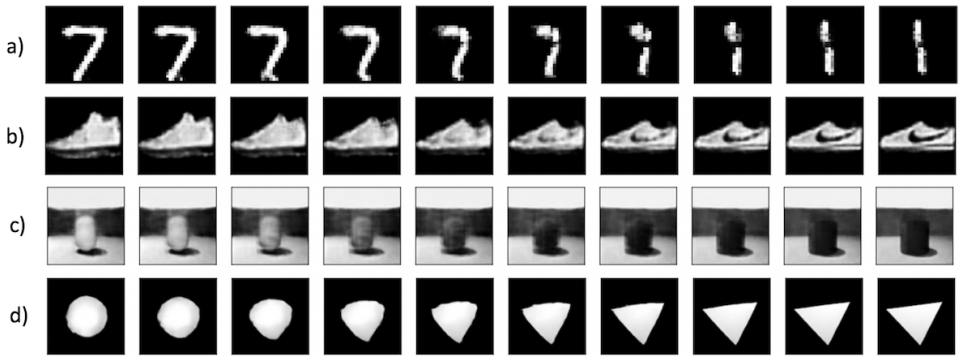

Figure 5: Generating new items with interpolation in latent space. Each row shows points along a line in the latent space between two items from the training data. Results are shown across a) MNIST, b) Fashion-MNIST, c) Shapes3D (converted to black and white), and d) Symmetric Solids.

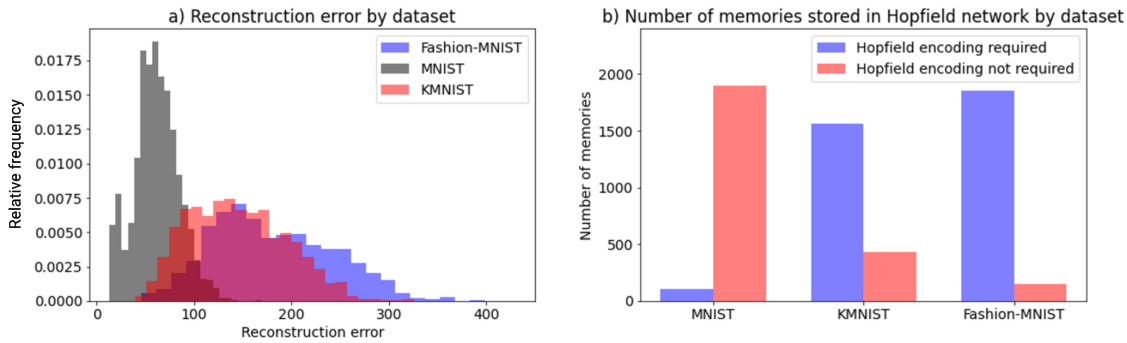

Figure 6: Novel items need to be encoded in the hippocampus more than predictable ones. a) A histogram of reconstruction errors across three datasets, for a VAE trained on items from the MNIST dataset. Reconstruction errors are measured for 2000 unseen images from each dataset. The reconstruction error per image is the sum across all pixels of the absolute difference between true and predicted pixel values, such that the maximum possible error for an image with N pixels is N, and the minimum is 0. (All three datasets have 784 pixels per image.) b) Starting with a VAE trained on items from the MNIST dataset, and simulating novelty-mediated consolidation, how many items are stored in the Hopfield network? The system experiences 2000 new images from each dataset. The threshold for encoding in the modern Hopfield network is set to 100.

3.3 Modelling semantic memory

A key finding in neuropsychology is that existing semantic memory survives when the hippocampus is lesioned or removed (Manns et al., 2003; Squire et al., 2015; Vargha-Khadem et al., 1997). 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 provides strong evidence that semantic memories – thought to be derived from episodic memories - are HF independent. In our model, lesions to the hippocampus would impair retrieval of recent events, while lesions to EC would impair all truly episodic recollection, since EC is required for the generation of sensory experience from latent variables in the generative network. However we will now describe how remote memories could be retrieved in semantic form despite this.

As Figure 1d shows, we propose that semantic memory for an individual event could correspond to a simple projection from the latent variable representation of that event in association cortex. The latent variables encode the key facts about an event, and semantic memory draws on these directly, rather than reconstructing them back into a sensory experience. In other words, the output route via EC is capable of turning latent variables into a sensory experience, but the latent variables themselves could support semantic memory. When the HF (including EC) is removed from the model, semantic memory still survives.

To illustrate the idea that the latent variables capture the key 'facts' of a scene, we trained models to predict attributes of each image, e.g. the digit in the case of MNIST, from its latent vector. Figure 7 shows that semantic 'decoding accuracy' increases as training progresses, thanks to structure developing in the latent space. Decoding accuracy is meas-

ured by training a new support vector classifier on 200 examples at the end of each epoch, and measuring classification accuracy on a held-out test set. (Notably, there is good performance with only a small amount of training data, i.e. few-shot learning is possible by making use of compressed 'semantic' representations.) Visualising the latent space shows that items from the same class are often clustered together, explaining the ease of decoding semantic information from the latent space. The ability to perform meaningful inference also illustrates the way that relevant facts are extracted from the Shapes3d dataset (Section 3.6, Figure 11).

To put this another way, given a good latent variable representation, it is straightforward to convert this to the key facts of a scene. We suggest that this corresponds to the persistence of semantic forms of autobiographical memory even when the hippocampus is damaged.

3.4 Combining predictable conceptual and unpredictable sensory elements in memory

In the extended model, memories stored in the hippocampal autoassociative network combine concepts (as derived from the generative network's latent variables) and unpredictable elements (those with a high reconstruction error in the generative network)(see Figure 2). Here we simulate this with the MNIST dataset, with white squares added to the images to represent novel aspects of an event. The 'concepts' in this scenario are simply the ten digits, whereas in reality there would be many concepts corresponding to any event.

To encode an image, it is first decomposed into its predictable (conceptual) and unpredictable (sensory) components by the generative network, and these are stored in the

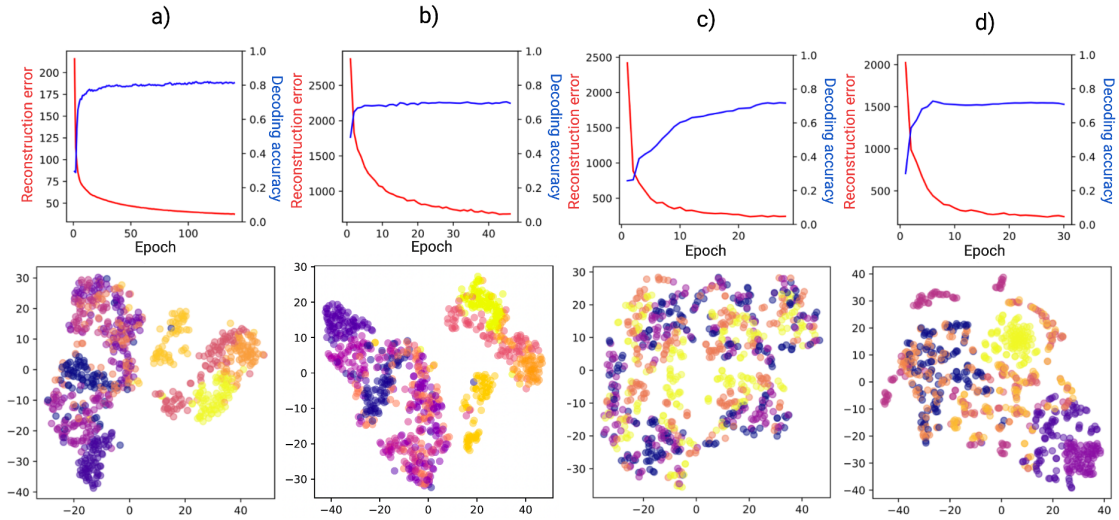

Figure 7: Upper row: reconstruction error (red) and decoding accuracy (blue) during training. The latter is measured by training a new support vector classifier on 200 examples at the end of each epoch. Lower row: the latent space projected into 2D with TSNE (Van der Maaten & Hinton, 2008), colour coded by label. Note that separation of classes is seen in the latent space projection. Results are shown across four datasets: a) MNIST (with ten digits as classes), b) Fashion-MNIST (with ten items of clothing as classes), c) Shapes3D (converted to black and white, with four shapes as classes), and d) Symmetric Solids (with eight shapes as classes).

conceptual and sensory feature units of the autoassociative network respectively. The conceptual component is identified by a simple model mapping latent variables to concepts, as described in Section 3.3.

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 concept's latent variables. (We trained a regression model to map 'concepts' to latent variables, but simply storing the mean latent vectors as 'prototypes' gives similar results.) Next, the predictable and unpredictable elements are recombined, simply by taking the unpredictable element if it is stored (i.e. if it is non-zero), and otherwise taking the element from the 'prototype'. Figure 8a 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 3.5).

How does replay work? When the autoassociative network is given random noise, both the unpredictable elements and the corresponding concept label are retrieved. In Figure 8c, the square images show the unpredictable elements of MNIST images and the rectangles below these display the vector of concept cells.

3.5 Modelling schema-based distortions

The model we propose is compatible with certain semantic, or schema-based, distortions seen in human episodic memory. Gist-based memory distortion increases over time (Bartlett, 1932), and processes correlated with memory consolidation, such as sleep, tend to promote it (Payne et al., 2009). The consolidation mechanism we propose is consistent

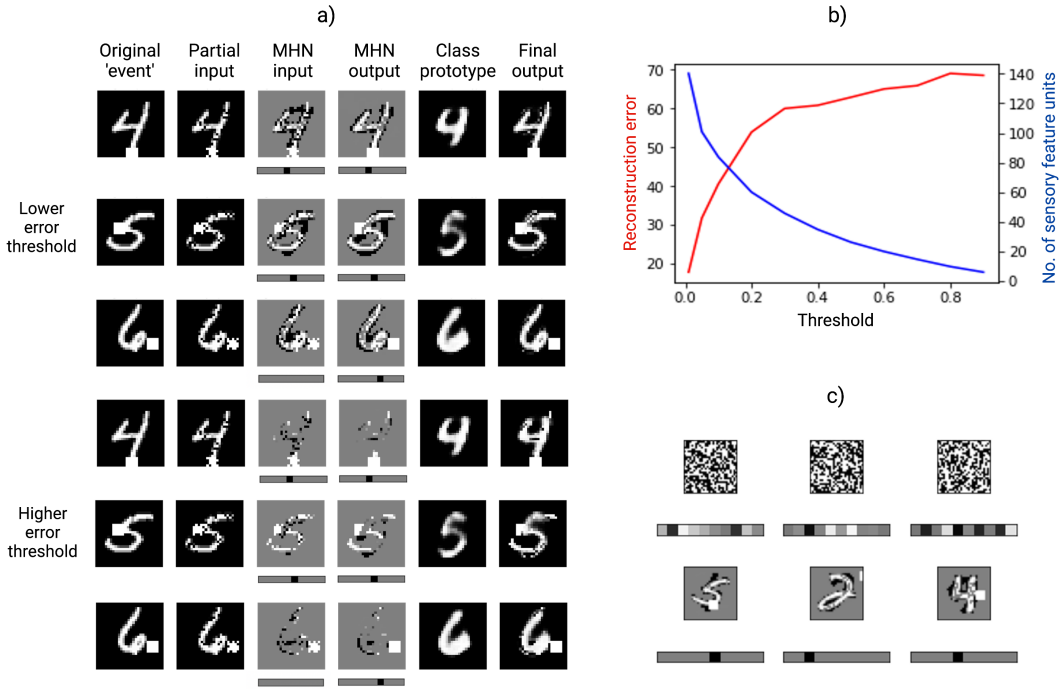

Figure 8: Recall in the extended model. a) The stages of recall are shown from left to right in each row: Firstly, the original 'event' is shown. Secondly, a partial (noisy) input triggers recall. Thirdly, the output and latent variable representations of the generative model identify the unpredictable elements and relevant concept(s) respectively, corresponding to sensory and conceptual features in the autoassociative Modern Hopfield Network (MHN). The square images show the high error elements of MNIST images - these are sparse representations as grey pixels indicate a value of zero in the array (i.e. grey pixels are those elements with a sub-threshold reconstruction error, which do not need to be stored). The rectangles below these display the vector of concept cells, where the concept of digit n is represented by a vector with 1 at element n of the vector. Fourthly, the autoassociative MHN performs pattern completion, producing a memory trace consisting of the unpredictable elements plus the concept(s) (MHN output). Fifthly, the 'prototype' is obtained by decoding the concept's latent variables. Finally, the unpredictable and prototypical components are recombined. A higher error threshold for encoding (bottom three rows) means fewer poorly predicted sensory features are stored in the autoassociative MHN, leading to more prototypical recall, with increased reconstruction error. b) Average reconstruction error and number of sensory features (i.e. pixels) stored in the autoassociative MHN against the error threshold for encoding. c) Replay in the autoassociative network. The autoassociative network can sample memories when random noise is given as an input. As above, the square images show the high error elements of MNIST images and the rectangles below these display the vector of concept cells.

with this; as reliance on the generative model increases, so does the level of distortion. But distortions of this kind can also have a rapid onset, as in the Deese-Roediger-McDermott task (Deese, 1959; Roediger & McDermott, 1995) where semantic intrusions (e.g. 'cold' when recalling a list of words semantically associated with cold) are observed in immediate recall, which is explained by the extended model.

Recall by the generative network distorts memories towards prototypical representations. Figure 9 shows that MNIST digits 'recalled' by a VAE become more prototypical in three ways. Firstly, Figure 9a shows MNIST digits and the VAE's output for each. Recalled pairs from the same class become more similar, i.e. intra-class variation decreases. Secondly, Figure 9b corroborates this by showing quantitatively that the intra-class variation decreases. Thirdly, plotting the pixel space of MNIST digits before and after recall, and the latent space of their encodings, also shows this effect (Figure 9c and 9d). Denser clusters are observed in the pixel space and latent space of recalled digits compared to the originals. In summary, recall with a VAE distorts stimuli towards more prototypical representations, even when no class information is given during training. (In other words, 'prototypes' are attractor states of the network.)

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. When they tried to reproduce the image after a delay, their drawings were distorted to look more like members of the context category. Figure 10b simulates the Carmichael et al. (1932) experiment, considering a recent memory in the extended model. The same image has been encoded with two different externally provided concepts (seven in the upper row, one in the lower row), as opposed to predicted concepts as in Figure 8a. During recall, the encoded concept is retrieved

in the autoassociative network, determining the prototypical digit reconstructed by the generative network. This biases recall towards the class provided as context, mirroring Figure 10a.

More generally, recall is biased towards the of 'mean' of the class soon after encoding in the extended model, due to the influence of the conceptual representations (Figure 8c). This is more pronounced when the error threshold for encoding is high, as there is more reliance on the 'prototype' representations in this case. Figure 8b shows how the threshold for encoding sensory details determines the behaviour of the model when it comes to distortions. 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. Conversely, at a higher error threshold, there is greater dependence on the conceptual features, so the dimension of the memory trace is lower, and hence distortion is higher. (Further research could explore how such a threshold might vary depending on the importance of the stimuli, or the amount of attentional resource available.)

3.6 Modelling imagination and inference

We model imagination as the ability to generate an event that has not been experienced in reality. Specifically, imagination is the decoding of a set of latent variables which do not correspond to a real memory into a sensory experience, i.e. an image in our simulations. Items can either be generated by the generative network from scratch (imagination), or by transforming the latent variable representations of existing items (inference). The former is simulated by randomly sampling from the latent space. The latter is simulated by interpolating between the latent representations of items, or by doing vector arithmetic

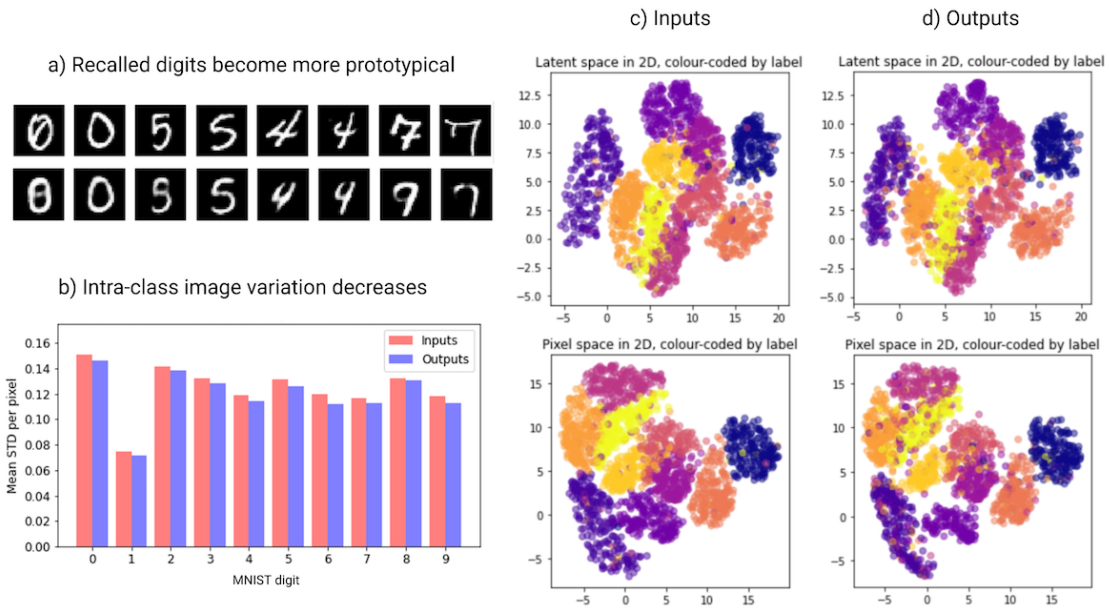

Figure 9: Items recalled by the generative network become more prototypical. a) MNIST digits (upper row) and the VAE's output for each (lower row). Recalled pairs from the same class become more similar. b) The intra-class variation, measured as the mean standard deviation per pixel within each MNIST class, is smaller for the recalled items than for the original inputs. 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.

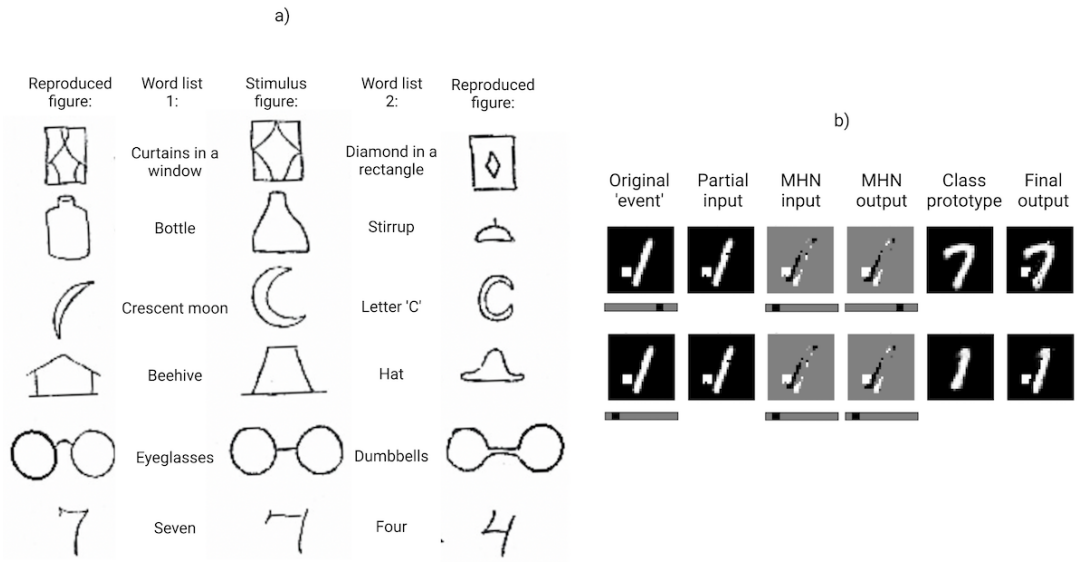

Figure 10: Contextual modulation of memory. a) Adapted version of figure from Carmichael (Carmichael et al., 1932) showing that humans recall the memory of an ambiguous item differently depending on its context, distorting the memory of an image to match a contextual cue at the time of encoding. b) Memory distortions in the extended model, when the original image is encoded along with activation of a given concept unit (7 above, 1 below). Then a partial input is processed by the generative network to produce a predicted conceptual feature (in this case 1) 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 encoding with conceptual unit 7 (top row) distorts recall towards a prototypical 7.

in the latent space. We simulated each of these as follows, using a VAE trained on the Shapes3D dataset (see Appendix for details).

Firstly, we can randomly sample from the latent space. Secondly, one can interpolate between items of the training data in the latent space, as shown in Figure 5 and Figure 11. Thirdly, we can do vector arithmetic in the latent space, as shown in Figure 11. Each equation shows $vector_1 + (vector_2 - vector_3) = result$, i.e. the output image is produced by taking the relation between $vector_2$ and $vector_3$, applying that to $vector_1$, and decoding the result. (For example, in the bottom row, the relation between $vector_2$ and $vector_3$ is to turn a cube into a sphere, so when this is applied to $vector_1$, the cube becomes a sphere but the other attributes stay the same.) This illustrates that the model has learnt some conceptual structure to the data, supporting reasoning tasks of the form 'A is to B as C is to ?'. This could correspond to the association cortex's ability to find relational structure in experience, and use it for inference. It may also reflect the flexible recombination of memories thought to underly episodic future thinking (Schacter et al., 2017).

3.7 Modelling brain damage

As described above, recent episodic memory is impaired following damage to the hippocampal formation, whereas semantic memory – including the semantic content of remote episodes - appears relatively spared. This is consistent with our model, in which the semantic form of a consolidated memory survives damage to the hippocampal formation thanks to latent variable representations in the mPFC or anterolateral temporal cortex (even if those in the EC are lesioned). Whether remote episodic memory survives is debated; some findings suggest that all proper episodic recollection requires the hippocampus and only semantic memory is possible without it (Nadel & Moscovitch, 1997), as in our

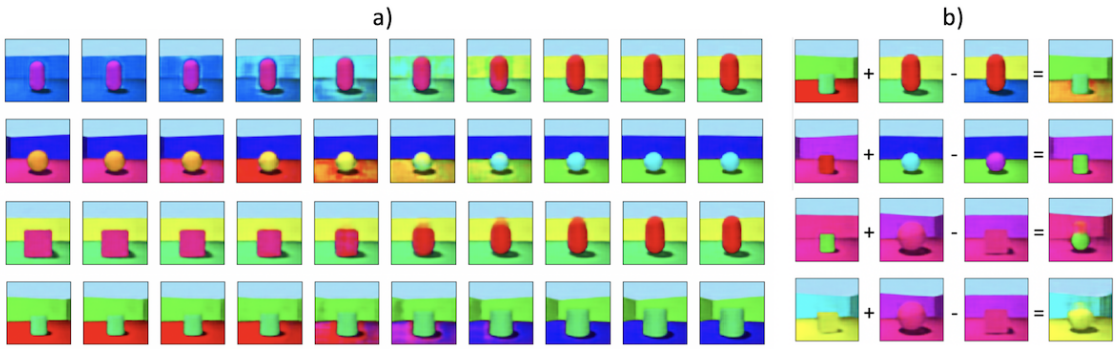

Figure 11: a) Generating new items via interpolation in the latent space. Each row shows items generated from latent variables along a line in the latent space between two real items from the training data. b) Modelling inference as vector arithmetic in the latent space. The left-hand side of each equation consists of three 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 item on the right-hand side is generated. The new item is related to the first item on the left in the same way as the second item is related to the third one. For example (upper row), the new item is related to the first item by the same transformation that takes the third item to the second item (addition of yellow to the floor), or (lower row) the new item is related to the first item by the transformation of cube to sphere.

model, but others dispute this (Squire et al., 2015).

Hippocampal amnesia can be explored computationally by 'lesioning' the parts of the model representing the hippocampus. McClelland et al. (1995) reproduce the temporal gradients observed in hippocampal amnesia in this way, showing that the neocortical network's recall performance improves with repeated exposures to the initial encoding. Therefore when the hippocampal network is damaged, and recall relies solely on the neocortical network, older memories are better preserved. In our model, hippocampal lesions can affect both the autoassociative network and the decoder of the generative network. When these regions are damaged, memories cannot be recalled *episodically* (by generating a sensory experience), but can be recalled *semantically* (which becomes possible after consolidation), in agreement with Nadel and Moscovitch (1997).

Figure 7, showing the decoding accuracy of image attributes from the latent variables as consolidation progresses, demonstrates how semantic recall performance depends on the age of a memory. The semantic projections do not rely on the hippocampus, so when the hippocampus is damaged and episodic recall is impaired, the semantic recall of events is better for older memories. This displays the same temporal gradient as in hippocampal amnesics, and is a consequence of the better latent variable representation that has been learned for older memories. Our model is also consistent with the fact that hippocampal amnesics struggle to encode new episodic memories, known as anterograde amnesia, as the hippocampus is required to initially bind stimuli together. (We did not simulate the decay of the initial encoding over time, as we assumed that all memories were still in the process of being consolidated in this simulation.)

Whilst these findings reflect the general pattern of deficits, there is great variation between patients depending, in part, on which regions of the hippocampal formation are damaged.

Whilst hippocampal damage appears to cause anterograde amnesia in all patients, the extent of retrograde amnesia can vary greatly. While some patients, such as V.C. (Cipolotti et al., 2001), exhibit retrograde amnesia extending all the way back to childhood, others such as R.B. (Zola-Morgan et al., 1986) show very limited retrograde amnesia. The dissociation of retrograde and anterograde amnesia in such cases suggests that the circuits for encoding memories and the circuits for recalling them via the hippocampal formation only overlap partially, and that the circuits for episodic recall of early memories are more damaged in some patients than others. Thus, the greater extent of hippocampal damage in V.C. than R.B. may explain V.C.'s impaired episodic recollection of early memories.

As shown in Figure 1, the amount of damage within the hippocampal formation also affects the resulting deficits in our model. For example, if the autoassociative network is damaged but not the generative network's decoder, the generative network can still perform reconstruction of consolidated memories, allowing episodic recollection to occur. This could explain varying reports of the gradient of retrograde amnesia when assessing episodic recollection (as opposed to semantic memory). Perhaps the generative network's decoder was intact in patients in which the episodic recollection of early memories appears spared (Squire et al., 2015).

Our model also predicts that hippocampal amnesics' schema acquisition would be impaired (as schema learning occurs when the schema-mediated route is trained by the autoassociative network). However, this model does not imply that consolidation is the only mechanism for schema acquisition, so it does not predict that schema acquisition is impossible after hippocampal damage (but it would likely be much slower without the benefit of replay). Anterograde semantic learning deficits consistent with this prediction have been observed (Scoville & Milner, 1957), but there is some conflicting evidence (see e.g. Gardiner et al.,

2008; Spiers et al., 2001).

In semantic dementia, semantic memory is impaired, and, unlike in hippocampal amnesia, remote episodic memory is impaired more than recent episodic memory (Hodges & Graham, 2001). In our model, lesions to the generative networks would affect remote memories more than recent ones, as recent memories can rely more on the hippocampal autoassociative network, therefore the behaviour of the model would be consistent with these data. However the exact effects would depend on the distribution of damage across the various potential generative networks in EC, mPFC and anterolateral temporal cortices. Of these networks, the anterolateral temporal network has been specifically associated with semantic dementia, and the posterior medial network (corresponding to the generative network between sensory areas and EC in our model) with Alzheimer's disease (Ranganath & Ritchey, 2012). In addition, the extended model would predict that the hippocampal encoding is less efficient after the onset of semantic dementia, relying more on sensory features and less on conceptual features than before.

3.8 Modelling continual learning

Catastrophic forgetting occurs when newly acquired knowledge or memory 'overwrites' previously acquired knowledge or memory in a neural network (McCloskey & Cohen, 1989). During consolidation, schemas in association cortex incorporate new information in a way that does not destroy previous learning, but the mechanism for this is not understood. One solution is to continue replaying hippocampal representations of consolidated memories indefinitely, as in CLS (McClelland et al., 1995), but this is inefficient. Another is to try to preserve existing representations by replaying information from within the generative network itself.

We suggest that generative replay prevents catastrophic forgetting in the generative model (Van de Ven & Tolias, 2018), and that this may occur in REM sleep (Singh et al., 2022). To demonstrate this, a VAE (with the same architecture as in the other simulations) was trained for ten epochs on 5000 MNIST images, and the histogram of reconstruction errors for new images from MNIST and Fashion-MNIST datasets was plotted. Unsurprisingly errors were low for MNIST and higher for Fashion-MNIST (Figure 12a). Then the model was trained for ten epochs on 3000 Fashion-MNIST images, leading to catastrophic forgetting of MNIST (Figure 12b). However, interleaving samples generated by the VAE at the end of MNIST training alleviates this and error for the MNIST dataset remains low (Figure 12c). 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 (Figure 12d).

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. For example, Caccia et al. (2021) explore methods for 'shielding the learned representations from drastic adaptation', which is reminiscent of regularisation-based approaches like elastic weight consolidation (Kirkpatrick et al., 2017). However, whether this is at the cost of other abilities - e.g. inferring the relationship between old and new categories - is unclear.

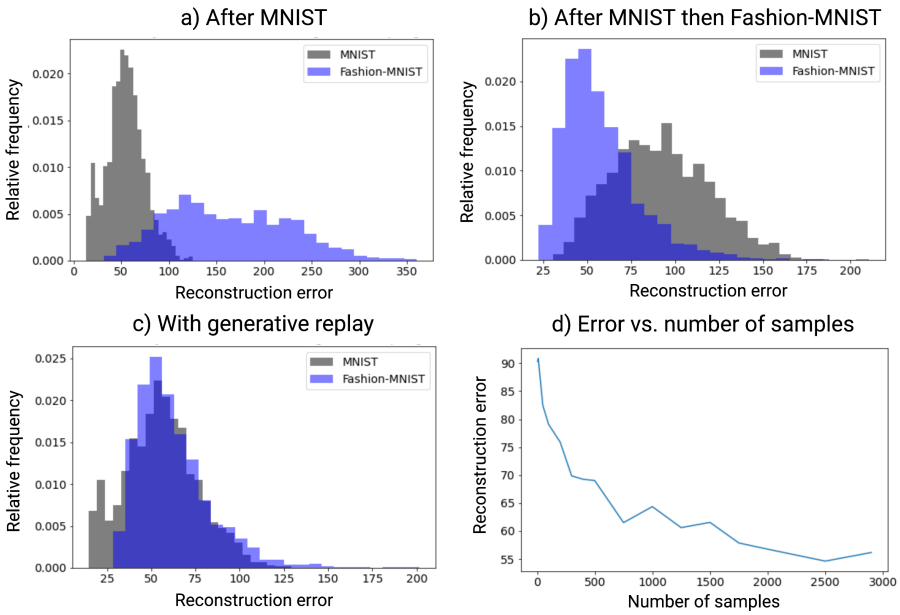

Figure 12: Generative replay reduces catastrophic forgetting. a) Results for a model trained on the MNIST dataset. The reconstruction error per image is the sum across all pixels of the absolute difference between true and predicted pixel values, such that the maximum possible error for an image with n pixels is n, and the minimum is 0. As expected, the error is low for new MNIST items and higher for Fashion-MNIST items. b) Histograms of reconstruction error for a model trained on MNIST then Fashion-MNIST. Catastrophic forgetting of MNIST items occurs. c) Results for a model trained on MNIST (phase one) then Fashion-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) The number of replayed samples from phase one against the mean reconstruction error for the MNIST dataset (using the training procedure in c).

3.9 Modelling memory for sequential structure

As opposed to static 'events' simulated by images, real episodes contain important sequential structure. Here we show that replacing the static VAE with an autoregressive generative network (GPT-2; Radford et al., 2019) enables the system to capture the consolidation and generation of sequential information, including the acquisition of 'semantic' knowledge and consequent distortions to memory.

The GPT-2 architecture can be trained on arbitrary linguistic or non-linguistic sequences; the task is simply to predict the next item in sequences drawn from the training data (see the Appendix for further details). As with the VAEs in previous simulations, training therefore involves learning to reconstruct inputs in a self-supervised manner. Whilst we do not explicitly model the initial storage of sequences in a modern Hopfield network, we note that there are several options for the initial encoding. (Most simply, the total number of feature units could be the product of the number of features at each time step and the number of time steps, so that feature node f_{ij} stores feature i at time step j.)

Firstly, we 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 the reverse – most transitions were random. After listening to a structured sequence for almost seven 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.

We generated a set of sequences using the transition structure in Durrant et al. (2011). These represent sequences stored in an initial bank of memories. The GPT-2 architecture was trained from scratch on the structured sequences only (with the sequences represented as comma-separated strings, e.g. '2,2,3,1,...',). It was trained for three epochs on 1000 sequences of 50 tones each, and at the end of each epoch of the training, the perplexity was calculated for test sets of structured and unstructured sequences. Perplexity indicates how 'surprising' a sequence is to the model's learned probability distribution (i.e. it is a measure of prediction error). In our model, 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 13 shows, the difference in perplexity increases over time as the generative model is trained, in agreement with the finding that consolidation improves ability to recognise structured vs. unstructured sequences.

Secondly, we explore distortions in memory for sequences, adding to evidence that consolidation into a generative model is consistent with the occurrence of memory distortions more broadly. We show that the GPT-2 model mirrors the finding that recalled narratives are distorted based on background semantic knowledge (Bartlett, 1932). In the Bartlett 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).

We fine-tuned GPT-2 on text from 'The War of the Ghosts', i.e. the existing GPT-2 model trained on a large corpus of contemporary documents was then trained on the story for 400 epochs. We explored recall of the story by giving the network the first few words of the story, and inspecting the predicted continuation at varying temperatures. (The temperature parameter determines the 'sharpness' of the distribution from which output tokens are selected. At a lower temperature the distribution is 'sharper', suppressing low probabilities, and the model is more likely to reproduce previously seen sequences.) We found that substitution of contemporary words and confabulated details were observed, as in the human data (see Table 2), and that greater distortion was observed at higher temperatures (see Figure 14).

Thirdly, we consider the ability to perform structural inference after experiencing multiple sequences of observations from a graph with a given structure, i.e., whether learning common relational structure between items allows inference of the relationship between a novel pair of items. 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. The relations in these tasks can be seen as edges in graphs (Whittington et al., 2020). Consolidation improves performance at structural inference (Ellenbogen et al., 2007; Kumaran, 2012).

As in Whittington et al. (2020), we model the spatial task of predicting the next location in a trajectory as the prediction of the next node in a graph. We created a large set of graphs, each one an n-by-n grid of nodes representing a simple spatial environment. Whilst each graph's structure is the same, nodes are labelled with random letters to represent arbitrary

Table 2: Example results from modelling Bartlett (1932) experiment

Original sentence	Recalled sentence
Now canoes came up, and they heard the noise of paddles, and saw one canoe coming up to them.	Now canoes came up, they heard the gunshots, and saw one canoe coming up to them.
And the warriors went on up the river to a town on the other side of Kalama.	And the soldiers went on up the river to a town on the other side of Kalama.
Behold I accompanied the ghosts, and we went to fight.	Behold I accompanied the enemy, and we attacked them.

features at a particular location. Each directed edge, i.e. each possible transition in the graph, is of the type north, south, east, or west. Random walks in the set of graphs were used to train the model (see Figure 15a); these could represent sequences stored in an initial bank of memories. The generative model was trained from scratch on the replayed sequences from these graphs (converted to strings of the form 'node1 E node2 W node3 ...'). It was trained for ten epochs, on 10,000 walks of 50 nodes each. The resulting model's successful performance on inference, and 'imagination' tasks is shown in Figures 15e and 15f respectively.

In summary, by replacing the VAE with other types of generative model such as autoregressive sequence models, our framework applies to sequential and as well as static 'events'.

4 Discussion

We have proposed a model of systems consolidation as the training of a generative neural network, which learns to support episodic memory, and also imagination, semantic memory

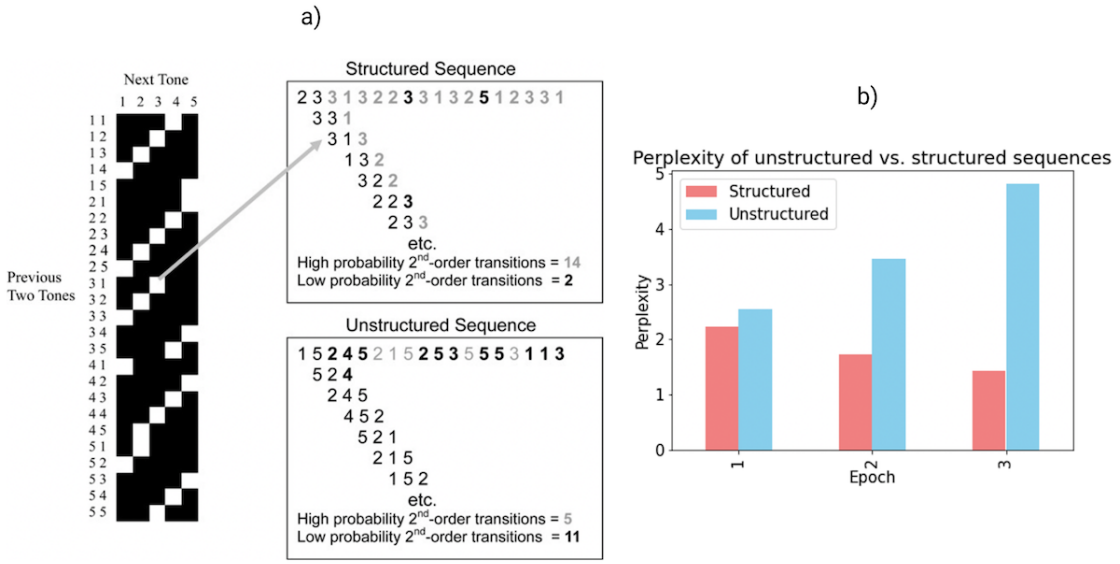

Figure 13: Learning sequential structure. a) Figure One from Durrant et al. (2011), showing the transition structure for structured sequences of tones in the task. In these sequences, the preceding two tones determine the next according to this matrix, except for 10% of transitions which are random to avoid repetition. That is, white squares indicate a 90% probability of a certain tone coming next in the sequence. b) Perplexity over time for structured vs. unstructured sequences, for a generative model trained on structured sequences. The ability to distinguish the two types based on their perplexity increases over time.

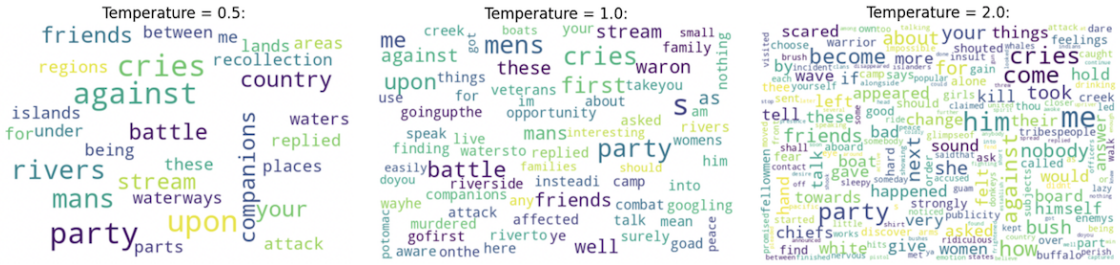

Figure 14: Distortion at different temperatures. These word clouds show terms in the recalled Bartlett stories which do not occur in the original text for a range of temperature values, with terms collated from five trials per temperature. More 'semantic intrusions' are observed at higher temperatures.

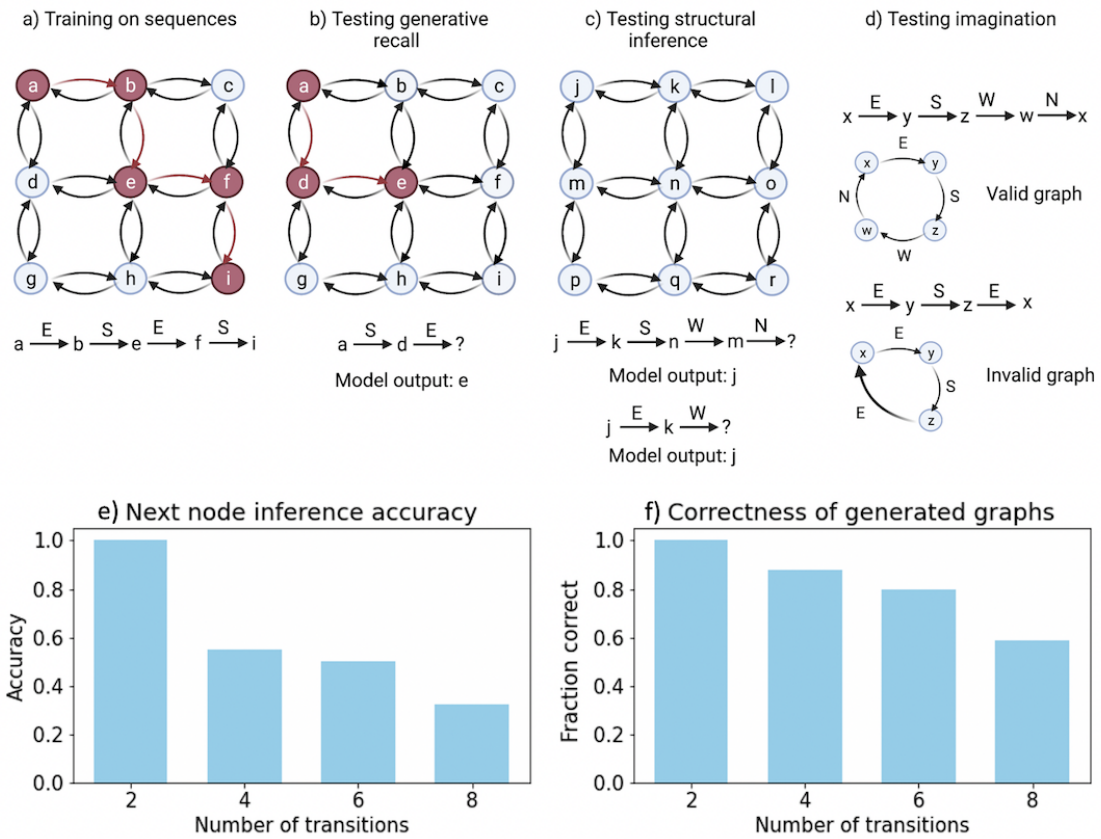

Figure 15: Learning to perform structural inference. a) The training data for an autoregressive generative model of sequences (GPT-2) are sampled from trajectories on a graph. b) Testing generative recall, i.e. the ability to complete unseen sequences from the graph used for training. c) Testing structural inference, i.e. the ability to complete sequences on unseen graphs based on structural similarities with seen graphs (e.g. by inferring a route is a shortcut). d) Testing imagination, modelled as the ability to generate sequences that correspond to a graph with a valid structure. e) The accuracy of GPT-2 trained on graph sequences for next node prediction, where the task is to predict the final node in a cycle, shown as a function of the number of transitions in the cycle. In each case, the next node can only be predicted using structural inference. f) The fraction of 'imagined' sequences which are structurally valid against the number of transitions.

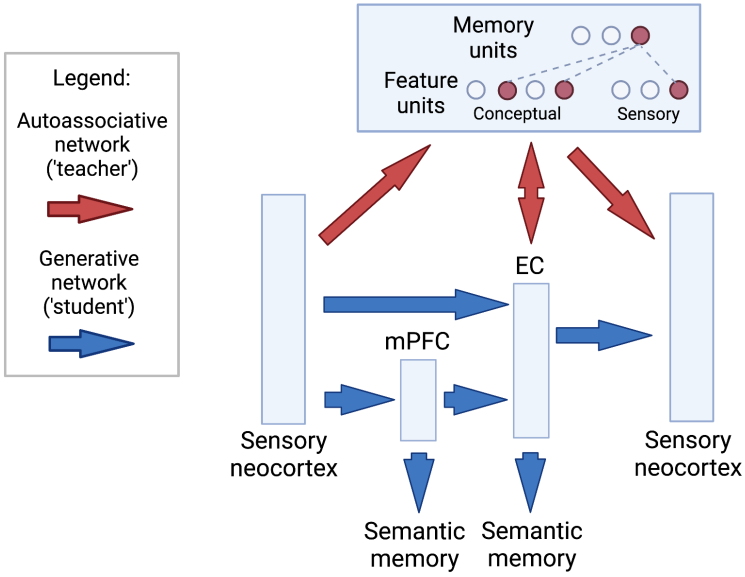

Figure 16: Multiple generative networks can be trained concurrently from a single autoassociative network through consolidation, with different networks optimised for different tasks. We expect there to be networks with latent variables in EC, mPFC, and anterolateral temporal cortices (not shown), each with their own semantic projections. But in all cases, return projections to sensory neocortex are via EC.

and inference. This occurs through teacher-student learning. First the hippocampal 'teacher' rapidly encodes an event, modelled as one-shot memorisation in a modern Hopfield network. This initial encoding 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)(see Figure 2). After exposure to replayed representations from the 'teacher', a generative 'student' network representing the association cortex and EC supports reconstruction of events. 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, in which the model reconstructs memories without the need to store them individually, but also imagination (by sampling from the latent variable distributions), inference (by using the learned statistics of experience to predict the values of unseen variables), and semantic memory. However, the generative network necessarily introduces semanticisation and distortion compared to the initial memory system. Multiple generative networks can be trained in parallel, and we expect this includes networks with latent variables in EC, mPFC, and anterolateral temporal cortex.

We will now revisit the list of findings from Table 1, and review how well the model fits with each one:

1. The initial encoding of memory happens rapidly whereas the consolidated form of memory is acquired more gradually (Marr, 1970, 1971; McClelland et al., 1995). In our model, a memory is encoded in the hippocampal 'teacher' network after a single exposure, and transferred to the generative 'student' network after being replayed repeatedly. Figure 3 shows how random noise can trigger the replay of memories, and Figures 4 and 5 show the behaviour of the generative model trained on these samples.

Linking memory consolidation to teacher-student learning of representations extends the CLS framework (McClelland et al., 1995).

2. Semantic content of memory becomes independent of hippocampus (Squire et al., 2015). Our model is consistent with this, as the latent variable representation constitutes the 'key facts' of an episode (see Figure 7).

3. However truly episodic memory (i.e. rich visuospatial imagery) remains dependent on the hippocampal formation (McKenzie & Eichenbaum, 2011; Nadel & Moscovitch, 1997), as return projections via EC are required to decode the latent variable representations into a sensory experience (see Figure 1).

4. Many forms of episode generation rely on similar neural circuits (Addis et al., 2007; Hassabis et al., 2007; McCormick et al., 2018; Schacter et al., 2017; Spanò et al., 2020). We propose that generative models are a common mechanism underlying episode generation. Figures 5 and 2 show how new items can be generated in our model by transforming existing items or from scratch by sampling, providing a model for imagination, scene construction and episodic future thinking.

5. Post-consolidation episodic memories support relational inference, generalisation, and the extraction of statistical regularities (Durrant et al., 2011; Ellenbogen et al., 2007; Richards et al., 2014). We show that relational inference arises naturally in generative latent variable models (e.g. a given relation can be applied to an item by vector arithmetic in the latent space, as in Figure 11b). Whilst inference in generative models is well understood in the machine learning literature, this has not been explored much in neuroscience and psychology.

6. Post-consolidation episodic memories are also more prone to gist-based distortions

(Bartlett, 1932; Payne et al., 2009). In Figure 10 we show how memory distortions arise from the post-consolidation generative model. This builds on Nagy et al. (2020), but to our knowledge there has been little modelling of the relationship between memory distortion and consolidation.

7. Regions of association cortex like mPFC are involved in schema processing (Ghosh & Gilboa, 2014), and are also thought to compress stimuli to a minimal representation (Mack et al., 2020). Given certain types of cue, they drive episodic recollection in the hippocampus, except for the most recent memories (McCormick et al., 2020). In addition, the EC can also encode the latent structures underlying experiences (Constantinescu et al., 2016; Whittington et al., 2020), and anterolateral temporal cortices are strongly associated with semantic memory (Chan et al., 2001). In our model the association cortex can encode latent variable representations of memory, which drive episodic recollection following consolidation via connections to the hippocampal formation, in agreement with these data.

8. Gradual, not catastrophic, forgetting is observed. In our model, when the generative network is trained on its own self-generated representations, catastrophic forgetting is alleviated (as in van de Ven et al., 2020).

9. Novelty affects memory processing (Hasselmo et al., 1996). In our model, the generative network's reconstruction error indicates novelty. Memories that are very consistent with the existing generative model do not need to be encoded in the autoassociative network (see Figure 6). The generative network is constantly calculating the prediction error (i.e. reconstruction error) of experience (as in Chen et al., 2011; Kumaran and Maguire, 2006). In the simplified version of the model this determines whether the entire memory is encoded in hippocampus, while in the

extended version this determines which elements are encoded as sensory features and which as conceptual features.

10. Memory traces in hippocampus bind together a mixture of sensory and conceptual elements (Quiroga, 2012). In our model, when an experience combines a mixture of familiar and unfamiliar elements, both concepts and poorly-predicted elements are stored in hippocampus, the former as conceptual features linked to the generative network's latent variable representations, and the latter as sensory features linked to sensory cortex.

A key aspect of the model is that multiple generative networks can be trained concurrently from a single autoassociative network through consolidation (Figure 16). This is advantageous because different generative networks may be optimised for different tasks. For example, one might hypothesise that the networks with latent representations in mPFC and anterolateral temporal cortices are more closely linked to language (in humans) than the network with latent representations in EC. For example, mPFC patients show particular deficits in studies in which memories are retrieved from a single generic cue word (e.g. Moscovitch and Melo, 1997). Similarly, a study in which mPFC was found to drive hippocampal activity during recall used schema-related prompts such as 'remember an easter egg hunt' (McCormick et al., 2020). Latent representations in mPFC may also depend more on value (Lin et al., 2016). In addition, we note the association of anterolateral temporal networks with semantic dementia, consistent with a role in semantic output, and the EC to sensory network with Alzheimer's disease, consistent with a role in episodic re-experience (Ranganath & Ritchey, 2012).

How might these differences arise computationally despite a shared training process? Firstly, the generative networks may differ in their structure, e.g. the degree of compression (i.e.

number of latent variables). Secondly, they may have additional training objectives that shape their representations, such as being optimised to predict some additional attribute. As Gluck and Myers (1993) suggest, this results in separation of classes in the latent space. Our approach resembles a model (Schapiro et al., 2017) that learns both individual examples and statistical regularities within the hippocampal formation itself (in the trisynaptic loop and EC-CA1 connections, respectively), but with statistical regularities captured in larger generative networks involving EC, which can reconstruct sensory neocortical representations.

Episodic memory is generally sequential in nature, so in additional simulations we have shown that replacing the static VAE with a sequential generative network (GPT-2; Radford et al., 2019) enables the system to capture the consolidation and generation of sequential information (Figure 15), including the acquisition of statistical structure (Figure 13) and consequent distortions to memory (Table 2). The sequence types tested are 'one dimensional' - at each time step a single variable (i.e. token) is present, although complex structural inference can be learned (Figure 15). However, each time step of experience is 'high dimensional' including many variables. Recent papers demonstrate that autoregressive sequence modelling can be applied to latent representations, such as those from a VAE, enabling the generation of complex sequential data like video (Yan et al., 2021).

This research raises some fundamental questions, such as whether true episodic memory requires event-unique detail, or whether combinations of prototypical predictions can qualify as memory rather than imagination. A related question is whether event-unique detail in memory must come from the hippocampus. In our model, event-unique details can be provided by the generative network, although they remain dependent on the hippocampus for longer. For example, if somebody knows that a particular person was at their 8th

birthday party and gave them 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 to qualify as episodic memory. The increasingly sophisticated generation of images from text using modern generative models (Ramesh et al., 2022) suggests that episode construction from semantic facts is consistent with this framework. Even without semantic inputs, the GPT-2 simulation (see Table 2) provides evidence that generative models can capture specifics - the particular story in the Bartlett experiment was remembered by the model, just with some distortions. (The same probability distribution in the weights supports both accurate 'memory', including specific details, and more obviously gist-based reconstruction at different temperature values.)

How does our model relate to the definition of episodic memories in terms of their unique spatiotemporal context and capacity for re-experiencing (Tulving, 1985)? We interpret re-experience here as the (re)construction of the high-level neural representations consistent with appropriate schema and including event-unique sensory features (see above). 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. The latent variable representations might be supported by grid cells in medial EC, forming 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 cells in lateral EC for temporal context (Bright et al., 2020; Tsao et al., 2018). As with the example of generating events consistent with personal semantic facts (above), events with specific spatial and temporal context could be generated from these latent variable representations.

Our model necessarily simplifies the true nature of mnemonic processing in several ways. The interaction of sensory and conceptual features in hippocampus and latent variables in EC during retrieval could be complex (as hinted at in the Carmichael experiment, Figure 10), with each type of representation contributing to the pattern completion of the other as per the interactions of items and context in the Temporal Context Model (Howard & Kahana, 2002), and might occur iteratively over retrievals from both the hippocampus and generative networks as per Kumaran et al. (2016). In addition, our model distinguishes between 'sensory' and 'conceptual' representations in hippocampus, linked to the sensory neocortex at the input/output of the generative network and to the latent variable layer in the middle respectively. In reality there might be a gradient from detailed sensory representation to coarse-grained conceptual representations in hippocampus, respectively linked to lower or higher neocortical areas. In this case, it is tempting to think that this gradient might map onto the longitudinal axis of the hippocampus, consistent with observed functional gradients along this axis (Strange et al., 2014).

There are several possible directions for future research. Firstly, experimental tests of the model could be designed. One test would be to have participants learn a set of stimuli generated from known latent variables, and use fMRI representational similarity analyses to see if this latent variable representation develops in association cortex over time. If the stimuli also contained slight variation, i.e. they were largely but not entirely described by the latent variables, behavioural tests could assess whether the development of a latent variable representation is correlated with gist-based distortions in memory, and anti-correlated with hippocampal processing of unpredictable elements. Secondly, the relationship between cognitive maps (Behrens et al., 2018) and latent spaces could be explored. For example, is the position of a memory in latent space reflected in place and grid cell firing, as observed for

other conceptual representations (Constantinescu et al., 2016; Nieh et al., 2021)? Thirdly, boundary extension and related phenomena could be modelled as gist-based distortions, for comparison with the data on how specific lesion affect them (Mullally et al., 2012).

Fourthly, the role of sleep in memory processing could be investigated. Singh et al. (2022) suggest that NREM and REM sleep may correspond to replay of hippocampal and neocortical memories respectively, and that this promotes continual learning. The authors observe that hippocampus and neocortex are tightly coupled during NREM but not REM sleep. They suggest that NREM sleep involves the replay of attractors from the hippocampal 'teacher', but REM sleep allows neocortex to replay its own attractors. In other words they suggest that in NREM sleep the hippocampus teaches neocortex its memories, but in REM sleep the neocortex 'revises' its own memories, and they demonstrate that this alternation prevents catastrophic forgetting. Our model contains similar phases of hippocampal replay to train neocortical generative networks and generative replay to support continual learning within the generative network (sections 2.3 and 3.8). Following Singh et al. (2022), we hypothesise that training on 'self-generated' episodes may occur during REM sleep, in contrast to hippocampal replay, which predominantly occurs in NREM sleep.

Fifthly, biological intelligence excels at generalising from only a small number of examples; how generative models may aid generalisation could be explored further. A key advantage of latent variable representations is that they make learning to categorise stimuli easier. Since the latent variables provide a minimal representation that already captures the key attributes of an input, another model can label the input's attributes with little training (see Figure 7). A related advantage of generative models is suggested by Barry and Love (2021), who found that adding samples from generative replay to the training data for a classifier improved its ability to generalise. The intuition behind this finding is clear from

inspecting the latent space of VAEs (e.g. in Figure 7). If one augmented the training data for category learning by generating items from nearby points in the latent space, or by inferring variants (see Figure 11), this could improve performance by exposing the model to more varied examples, which would be especially beneficial if there was little training data. Therefore future research should investigate how generative models in the brain might result in better ability to generalise to new stimuli.

Finally, future work could explore how language relates to our model. Memories include conceptual elements and, in humans, concepts are tied to language. Intuitively language is a compressed representation of experience which could be thought of as a special kind of 'latent variable representation'. In simulations using MNIST digits (see Figure 8) we have modelled the conceptual part of each image as a single concept for simplicity, but the schematic component of experience is more likely to be a combination of concepts. Language is well suited to representing a flexible combination of concepts in this way. Furthermore there are only ten relevant concepts for MNIST images, but for naturalistic images (and events) there are obviously a vast number, with complex relationships between them. Simply storing concepts predicted from latent variables is no longer feasible in this case, however language describing an event could be generated from the latent variables for efficient storage instead.

In summary, our 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, replay, and consolidation, as well as aspects of imagination, structural inference and long-term continual learning.

References

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.

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.

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.

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.

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.

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).

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.

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.

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*.

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.

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.

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.

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.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*.

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.

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.

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.

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

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of neurosciences*, *20*(4), 155.

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.

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature reviews neuroscience*, *6*(2), 119–130.

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.

Gardiner, J. M., Brandt, K. R., Baddeley, A. D., Vargha-Khadem, F., & Mishkin, M. (2008). Charting the acquisition of semantic knowledge in a case of developmental amnesia. *Neuropsychologia*, *46*(11), 2865–2868.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114.

Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, *21*(8), 618–631.

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.

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.

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.

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.

Hemmer, P., & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

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.

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.

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.

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.

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.

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). Over-

coming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

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.

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.

Lin, W.-J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific reports*, *6*(1), 1–10.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature communications*, *11*(1), 1–11.

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.

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., Rosenthal, C. R., Miller, T. D., & Maguire, E. A. (2018). Mind-wandering in people with hippocampal damage. *Journal of Neuroscience*, *38*(11), 2745–2754.

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.

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.

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.

Murphy, K., Esteves, C., Jampani, V., Ramalingam, S., & Makadia, A. (2021). Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, *7*(2), 217–227.

Nagy, D. G., Török, B., & Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, *16*(10), e1008367.

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, 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.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

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.

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.

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597.

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.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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.

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.

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.

Schacter, D. L. (2012). Constructive memory: Past and future. *Dialogues in clinical neuroscience*, *14*(1), 7.

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., 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.

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*.

Singh, D., Norman, K. A., & Schapiro, A. C. (2022). A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *bioRxiv*.

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., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor perspectives in biology*, *7*(8), a021766.

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.

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, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2021). Organizing memories for generalization in complementary learning systems. *BioRxiv*.

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.

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

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. (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 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 Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

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 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.

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.

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.

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society*, *17*(5), 766–780.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

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.

A Appendix

Code for all simulations can be found at https://github.com/ellie-as/generative-consolidation.

Diagrams were created using BioRender.com.

A.1 Datasets

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
Fashion-MNIST (Xiao et al., 2017)	https://www.tensorflow.org/datasets/catalog/fashion_mnist
Shapes3D (Burgess & Kim, 2018)	https://www.tensorflow.org/datasets/catalog/Shapes3D
Symmetric Solids (Murphy et al., 2021)	https://www.tensorflow.org/datasets/catalog/symmetric_solids
KMNIST (Clanuwat et al., 2018)	https://www.tensorflow.org/datasets/catalog/kmnist

A.2 Model details

A.2.1 Variational autoencoders

An autoencoder is a neural network which encodes an input into a shorter vector, and then decodes this compressed representation 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 regions 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, which allow one to sample realistic items. The result is called a variational autoencoder (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 - see Figure 6 of Hou et al. (2017) for an example.

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 and Symmetric Solids datasets (containing images of size 128x128 pixels), and a small VAE was used for the MNIST and Fashion-MNIST datasets (containing images of size 28x28 pixels). In the large model's encoder, five convolutional layers, plus a final pooling layer, gradually decrease the width and height of the representation and increase the depth (as is standard when using convolutional neural networks to encode images). 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, five convolutional layers alternate with up-sampling layers to increase the width

and height of the representation and decrease the depth. The encoder has filters (i.e. convolution windows, or feature detectors) of 4x4 pixels per layer. The convolutional layers one to five have 32, 64, 128, 256 and 512 filters respectively. The decoder has filters of 3x3 pixels per layer. The convolutional layers one to five have 128, 64, 32, 16 and 3 filters respectively (i.e. the final three filters are the three channels of an RGB image). The small model used a reduced architecture, with fewer convolutional layers, for efficiency.

A.2.2 Modern Hopfield networks

A Hopfield network uses a simple Hebbian learning rule to memorise 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 in several ways. Krotov and Hopfield (2016) devise a new energy function involving a polynomial function, and a corresponding update rule to minimise this; the activation of a node flips from -1 to 1 or vice versa if the energy is lower in the flipped state. 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) extend this to memories involving continuous variables and further amend the energy function, enabling 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. To recap, Equation (1), from Krotov and Hopfield (2016), gives the energy of a standard Hopfield Network. During recall, a node's value is updated to the sign of the weighted sum of its inputs; 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. (In these equations, σ gives the state of the network as a vector, and ξ gives a stored pattern.)

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

Equation (2), 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, Equation (2) reduces to Equation (1) 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.

$$(2) \qquad E = -\sum_{\mu} F(\sum_{i} \xi_{\mu i} \sigma_i) = -\sum_{i,j,k} T_{ijk} \sigma_i \sigma_j \sigma_k$$

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. As a result, a modern Hopfield network can be understood as a neural network, like its predecessor. The authors provide two equations for the evolution of the feature neurons and hidden neurons over time. Rather than using discrete time steps as in a classical Hopfield network, time is modelled as continuous. They therefore give a pair of differential equations, in which change to each set of currents is driven by the weighted sum of currents in the other layer. They then define an energy function, chosen 'so that the energy function decreases on the dynamical trajectory'. The energy function has three terms: energy in the feature neurons, energy in the hidden neurons, and energy from the interaction between the two groups. Importantly, the interaction term can be described in terms of two-body synapses, so once again the energy is a function of weights and activations in a neural network.

The authors state that 'the memory patterns ... can be interpreted as the strengths of the synapses connecting feature and memory neurons'. To understand the intuition behind this, suppose we set the weights connecting a particular hidden node with the feature neurons to the values of the pattern to be memorised. Then activating the hidden node results in the pattern being reinstated in the feature neurons. In other words, each hidden node represents a memory, and each memory could be encoded using Hebbian learning. The key point is that the energy does not require a matrix of stored patterns, unlike in earlier formulations of modern Hopfield networks – the patterns are encoded in the weights, and the energy is a function of weights and activations as explained above.

Krotov and Hopfield (2020) show that under different circumstances, their formulation can be simplified to dense associative memory (Krotov & Hopfield, 2016), or modern Hopfield networks (Ramsauer et al., 2020). Having established that modern Hopfield networks increase memory performance and are biologically plausible (in the sense that they involve

only 'two-body synapses', and that memories can be stored as weights), we use them to model the initial learning in the hippocampus.

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 is similar to previous models of the hippocampus as 'indexing', or binding together, a set of memory components (Teyler & DiScenna, 1986). The innovative aspect of modern Hopfield networks is the update rule, which is cleverly designed to guarantee the desired properties. The equation below gives the new state pattern ξ^{new} in terms of the previous state ξ, stored patterns X^T, and inverse temperature β:

$$\xi^{new} = X softmax(\beta X^T \xi) \tag{3}$$

In modern Hopfield networks, the inverse temperature parameter β determines whether individual attractors or metastable states (superpositions of stored attractors) are retrieved. In our simulations we set β very high in order to ensure that only individual 'memories' are recalled.

It should be noted that the modern Hopfield network could be swapped out for other computational models of associative memory, providing they i) are high capacity, ii) can retrieve memories from noise, and iii) are capable of one-shot memorisation.

A.2.3 Training procedure

During the 'active' state in our simulations, images are stored in a continuous modern Hopfield network with a high inverse temperature. Ramsauer et al. (2020) provide an excellent Python implementation of modern Hopfield networks that we use in our code. (Note that with the exception of the continual learning experiments, only one dataset is used at a time.)

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 β).

A VAE was then trained on just the 'replayed' images from the modern Hopfield network, using the Keras API for TensorFlow (Abadi et al., 2016). The stochastic gradient descent method used was the AMSGrad variant of the Adam optimiser, with early stopping enabled. A latent code of length 10, learning rate of 0.001 and KL weighting of 1 were used in the main results. The variational autoencoders were not optimised for performance, as their purpose is illustrative (more data and hyperparameter tuning would be likely to improve reconstruction accuracy).

A.2.4 GPT-2

GPT-2 (Radford et al., 2019) (along with more recent variants like GPT-3; Brown et al., 2020) learns to predict the next item in a text sequence given the preceding items, a procedure known as causal language modelling. The probability of a generated sentence is a product of conditional probabilities for each token given the preceding tokens in the sentence. Tokens are chunks of characters that tend to occur together, typically shorter

than words; the ‘tokeniser’ component of the model transforms the original sequence into a sequence of tokens.

Once trained, GPT-2 outputs predicted probabilities for the next token given the sequence so far. Generating text involves sampling repeatedly from these predicted probabilities, controlled by a number of parameters. (Depending on the choice of these parameters, GPT-2 can be treated as a model of both recall and generation of new content. In particular, the temperature parameter determines the ‘sharpness’ of the distribution from which output tokens are selected. At a lower temperature the distribution is ‘sharper’, suppressing low probabilities, and the model is more likely to reproduce previously seen sequences.)

To simulate the Bartlett experiment (Figure 14 and Table 2), the open-source GPT-2 model (trained on causal language modelling of a large corpus of documents) was then trained on the ’War of the Ghosts’ story with causal language modelling for 400 further epochs.

For the non-language tasks in Figures 13 and 15, we take the simple approach of converting the data into sequences of characters. The procedure for training GPT-2 on non-language sequences of characters differs in two ways from the procedure above. Firstly, a custom tokeniser must be used (the base model’s tokeniser is not applicable as it was fitted on language). Secondly, the models are trained from scratch (as opposed to fine-tuning), because prior training on language was not relevant to the task.

For the Durrant et al. (2011) experiment, the model was trained for three epochs on 1000 sequences of 50 tones each, with the sequences converted to strings of the form ‘1,5,2,3,. . . ’.. For the graph transitions experiment, walks on the graph were converted to strings of the form ‘node1 E node2 W node3 . . . ’. The model was trained for ten epochs, on 10,000 walks of 50 nodes each.