nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01799-z

A generative model of memory construction and consolidation

Received: 30 May 2023

Accepted: 5 December 2023

Published online: 19 January 2024

Check for updates

Eleanor Spens [1] & Neil Burgess [1,2]

Episodic memories are (re)constructed, share neural substrates with imagination, combine unique features with schema-based predictions and show 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) to (re)create sensory experiences from latent variable representations in entorhinal, medial prefrontal and anterolateral temporal cortices via the hippocampal formation. Simulations show effects of memory age and hippocampal lesions in agreement with previous models, but also provide mechanisms for semantic memory, imagination, episodic future thinking, relational inference and schema-based distortions including boundary extension. The model explains how unique sensory and predictable conceptual elements of memories are stored and reconstructed by efficiently combining both hippocampal and neocortical systems, optimizing the use of limited hippocampal storage for new and unusual information. Overall, we believe hippocampal replay training generative models provides a comprehensive account of memory construction, imagination and consolidation.

Episodic memory concerns autobiographical experiences in their spatiotemporal context, whereas semantic memory concerns factual knowledge[1]. The former is thought to rapidly capture multimodal experience via long-term potentiation in the hippocampus, enabling the latter to learn statistical regularities over multiple experiences in the neocortex[2–5]. Crucially, episodic memory is thought to be constructive; recall is the (re)construction of a past experience, rather than the retrieval of a copy[6,7]. But the mechanisms behind episodic (re)construction and its link to semantic memory are not well understood.

Old memories can be preserved after hippocampal damage despite amnesia for recent ones[8], suggesting that memories initially encoded in the hippocampus end up being stored in neocortical areas, an idea known as 'systems consolidation'[9]. The standard model of systems consolidation involves transfer of information from the hippocampus to the neocortex[2–4,10], whereas other views suggest that episodic and semantic information from the same events can exist in parallel[11]. Hippocampal 'replay' of patterns of neural activity during rest[12,13] is thought to play a role in consolidation[14,15]. 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 semanticization[16,17].

Generative models capture the probability distributions underlying data, enabling the generation of realistic new items by sampling from these distributions. Here we propose that consolidated memory takes the form of a generative network, trained to capture the statistical structure of stored events by learning to reproduce them (see also refs. 18,19). As consolidation proceeds, the generative network supports both the recall of 'facts' (semantic memory) and the reconstruction of experience from these 'facts' (episodic memory), in conjunction with additional information from the hippocampus that becomes less necessary as training progresses.

This builds on existing models of spatial cognition in which recall and imagination of scenes involve the same neural circuits[20–22], and is supported by evidence from neuropsychology that damage to the

[1]UCL Institute of Cognitive Neuroscience, University College London, London, UK. [2]UCL Queen Square Institute of Neurology, University College London, London, UK. e-mail: eleanor.spens.20@ucl.ac.uk; n.burgess@ucl.ac.uk

hippocampal formation (HF) leads to deficits in imagination[23], episodic future thinking[24], dreaming[25] and daydreaming[26], as well as by neuroimaging evidence that recall and imagination involve similar neural processes[27,28].

We model consolidation as the training of a generative model by an initial autoassociative encoding of memory through 'teacher–student learning'[29] during hippocampal replay (see also ref. 30). Recall after consolidation has occurred is a generative process mediated by schemas representing common structure across events, as are other forms of scene construction or imagination. Our model builds on: (1) research into the relationship between generative models and consolidation[18,19], (2) the use of variational autoencoders to model the hippocampal formation[31–33] and (3) the view that abstract allocentric latent variables are learned from egocentric sensory representations in spatial cognition[22].

More generally, we build on the idea that the memory system learns schemas which encode 'priors' for the reconstruction of input patterns[34,35]. Unpredictable aspects of experience need to be stored in detail for further learning, while fully predicted aspects do not, consistent with the idea that memory helps to predict the future[36–39]. We suggest that familiar components are encoded in the autoassociative network as concepts (relying on the generative network for reconstruction), while novel components are encoded in greater sensory detail. This is efficient in terms of memory storage[40–42] and reflects the fact that consolidation can be a gradual transition, during which the autoassociative network supports aspects of memory not yet captured by the generative network. In other words, the generative network can reconstruct predictable aspects of an event from the outset on the basis of existing schemas, but as consolidation progresses, the network updates its schemas to reconstruct the event more accurately until the formerly unpredictable details stored in HF are no longer required.

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

1. The initial encoding of memory requires only a single exposure to the event and depends on the HF, while the consolidated form of memory is acquired more gradually[2,3,10], as in the complementary learning systems (CLS) model[4].
2. The semantic content of memories becomes independent of the HF over time[43–45], consistent with CLS.
3. Vivid, detailed episodic memory remains dependent on HF[46], consistent with multiple trace theory[11] (but not with CLS).
4. Similar neural circuits are involved in recall, imagination and episodic future thinking[27,28], suggesting a common mechanism for event generation, as modelled in spatial cognition[22].
5. Consolidation extracts statistical regularities from episodic memories to inform behaviour[47,48], and supports relational inference and generalization[49]. The Tolman–Eichenbaum machine (TEM)[31] simulates this in the domain of multiple tasks with common transition structures (see also ref. 50), while ref. 51 models how both individual examples and statistical regularities could be learned within HF.
6. Post-consolidation episodic memories are more prone to schema-based distortions in which semantic or contextual knowledge influences recall[6,52], consistent with the behaviour of generative models[32].
7. Neural representations in the entorhinal cortex (EC) such as grid cells[53] are thought to encode latent structures underlying experiences[31,54], and other regions of the association cortex, such as the medial prefrontal cortex (mPFC), may compress stimuli to a minimal representation[55].
8. Novelty is thought to promote encoding within HF[56], while more predictable events consistent with existing schemas are consolidated more rapidly[57]. Activity in the hippocampus can reflect prediction error or mismatch novelty[58,59], and novelty is thought to affect the degree of compression of representations in memory[60] to make efficient use of limited HF capacity[42].
9. Memory traces in the hippocampus appear to involve a mixture of sensory and conceptual features, with the latter encoded by concept cells[61], potentially bound together by episode-specific neurons[62]. Few models explore how this could happen.

Consolidation as the training of a generative model

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; then, generative networks gradually take over after being trained on replayed representations from the hippocampus. This makes the memory more abstracted, more supportive of generalization and relational inference, but also more prone to gist-based distortion. The generative networks can be used to reconstruct (for memory) or construct (for imagination) sensory experience, or to support semantic memory and relational inference directly from their latent variable representations (see Fig. 1).

Before consolidation, the hippocampal autoassociative network encodes the memory. A modern Hopfield network (MHN)[63] is used, which can be interpreted such that the feature units activated by an event are bound together by a memory unit[64] (see Methods and Supplementary Information). Teacher–student learning[29] allows transfer of memories from one neural network to another during consolidation[30]. Accordingly, we use outputs from the autoassociative network to train the generative network: random inputs to the hippocampus result in the reactivation of memories, and this reactivation results in consolidation. After consolidation, generative networks encode the information contained in memories. Reliance on the generative networks increases over time as they learn to reconstruct a particular event.

Specifically, the generative networks are implemented as variational autoencoders (VAEs), which are autoencoders with special properties such that the most compressed layer represents a set of latent variables, which can be sampled from to generate realistic new examples corresponding to the training dataset[65,66]. 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 Methods). The VAE's encoder 'encodes' sensory experience as latent variables, while its decoder 'decodes' latent variables back to sensory experience. In psychological terms, after training on a class of stimuli, VAEs can reconstruct such stimuli from a partial input according to the schema for that class, and generate novel stimuli consistent with the schema. (Our use of VAEs is illustrative, and we would expect a range of other generative latent variable models, such as predictive coding networks[67–69], to show similar behaviour.) See Methods and Supplementary Information for further details.

Generative networks capture the probability distributions underlying events, or 'schemas'. In other words, here 'schemas' are rules or priors (expected probability distributions) for reconstructing a certain type of stimulus (for example, the schema for an office predicts the presence of co-occurring objects such as desks and chairs, facilitating episode generation), whereas concepts represent categories but not necessarily how to reconstruct them. However, schemas and concepts are closely related, and their meanings can overlap, with conflicting definitions in the psychology literature[70,71].

During perception, the generative model provides an ongoing estimate of novelty from its reconstruction error (also known as 'prediction error', the difference between input and output representations). Aspects of an event that are consistent with previous experience (that is, with low reconstruction error) do not need to be encoded in detail in the autoassociative 'teacher' network[36–39]. Once the generative network's reconstruction error is sufficiently low, the hippocampal trace is unnecessary, freeing up capacity for new encodings. However, we have

https://doi.org/10.1038/s41562-023-01799-z

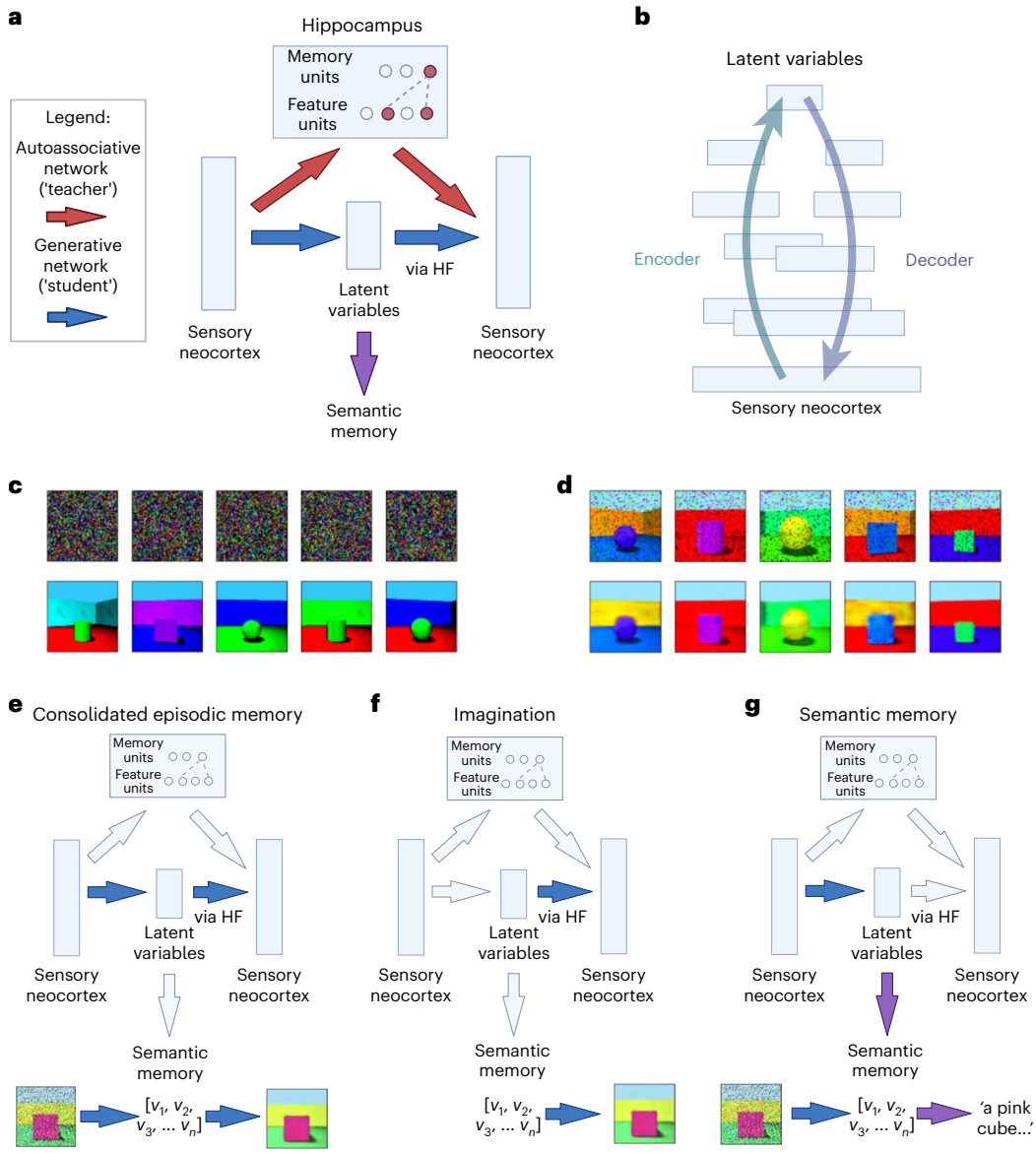

Fig. 1 | Architecture of the basic model. a, First, the hippocampus rapidly encodes an event, modelled as one-shot memorization in an autoassociative network (an MHN). Then, generative networks are trained on replayed representations from the autoassociative network, learning to reconstruct memories by capturing the statistical structure of experienced events. **b**, A more detailed schematic of the generative network to indicate the multiple layers of, and overlap between, the encoder and decoder (where layers closer to the sensory neocortex overlap more). The generation of a sensory experience, for example, visual imagery, requires the decoder to the sensory neocortex via HF. **c**, Random noise inputs to the MHN (top row) reactivate its memories (bottom row) after 10,000 items from the Shapes3D dataset are encoded, with five examples shown. **d**, The generative model (a variational autoencoder) can recall images (bottom row) from a partial input (top row), following training on 10,000 replayed memories sampled from the MHN. **e**, Episodic memory after consolidation: a partial input is mapped to latent variables whose return projections to the sensory neocortex via HF then decode these back into a sensory experience. **f**, Imagination: latent variables are decoded into an experience via HF and return projections to the neocortex. **g**, Semantic memory: a partial input is mapped to latent variables, which capture the 'key facts' of the scene. The bottom rows of **e**–**g** illustrate these functions in a model that has encoded the Shapes3D dataset into latent variables ($v_1, v_2, v_3, \ldots, v_n$). Diagrams were created using BioRender.com.

not simulated decay, deletion or capacity constraints in the autoassociative memory part of the model.

Combining conceptual and sensory features in episodic memory

Consolidation is often considered in terms of fine-grained sensory representations updating coarse-grained conceptual representations, for example, the sight of a particular dog updating the concept of a dog. Modelling hippocampal representations as sensory-like is a reasonable simplification, which we make in simulations of the 'basic' model in Fig. 1. However, memories probably bind together representations along a spectrum from coarse-grained and conceptual to fine-grained and sensory. For example, the hippocampal encoding of a day at the beach is likely to bind together coarse-grained concepts such as 'beach' and 'sea' along with sensory representations such as the melody of an unfamiliar song or the sight of a particular sandcastle, consistent with the evidence for concept cells in the hippocampus[61]. (This also fits with the observation that ambiguous images 'flip' between interpretations in perception but are stable when held in memory[72], reflecting how the conceptual content of memories constrains recall.)

https://doi.org/10.1038/s41562-023-01799-z

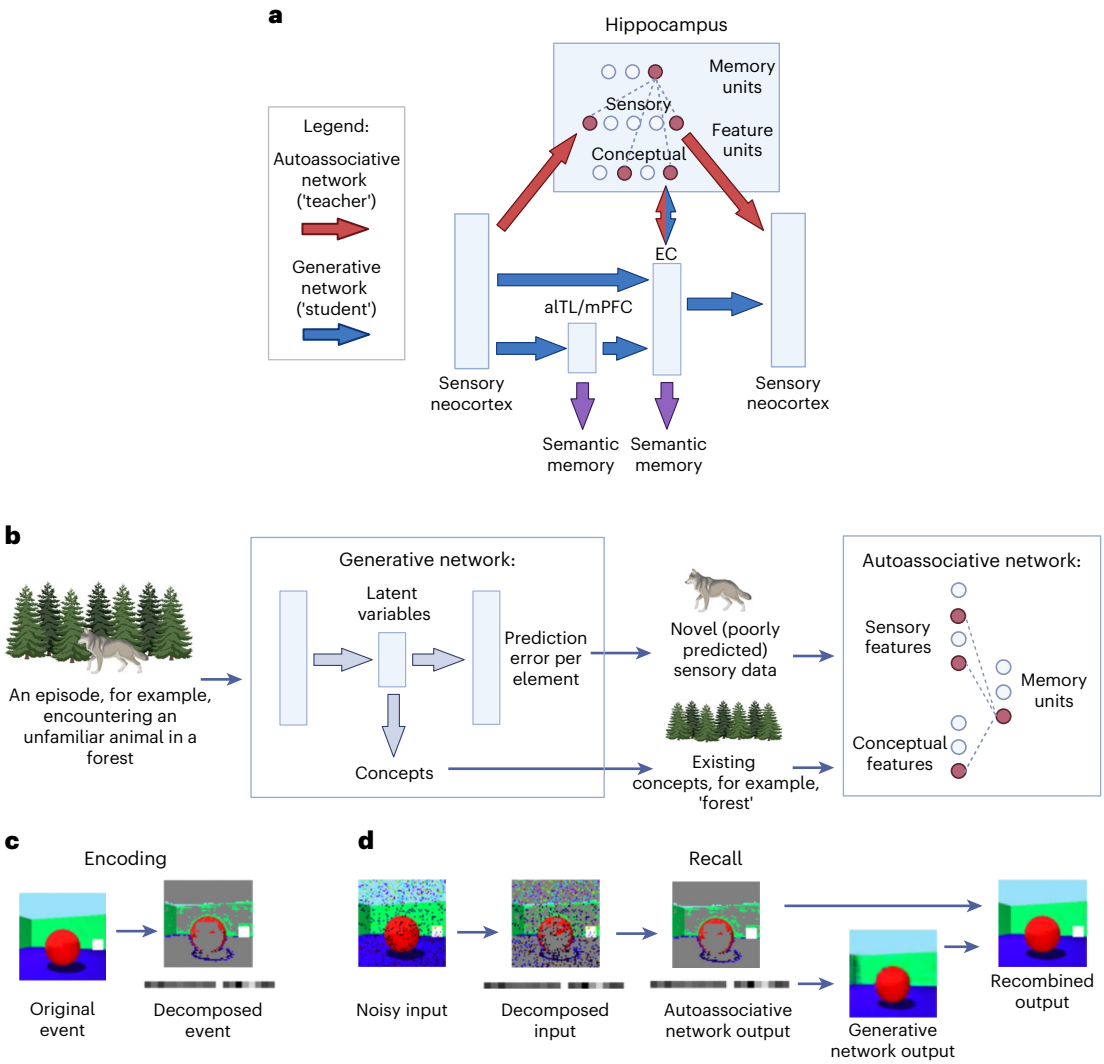

Fig. 2 | Architecture of the extended model. a, Each scene is initially encoded as a combination of predictable conceptual features related to the latent variables of the generative network and unpredictable sensory features that were poorly predicted by the generative network. An MHN (in red) encodes both sensory and conceptual features (with connections to the sensory neocortex and latent variables in EC, respectively), binding them together via memory units. Memories may eventually be learned by the generative model (in blue), but consolidation can be a prolonged process, during which time the generative network provides schemas for reconstruction and the autoassociative network supports new or detailed information not yet captured by these schemas. Multiple generative networks can be trained concurrently, with different networks optimized for different tasks. This includes networks with latent variables in EC, mPFC and alTL, each with its own semantic projections. However, in all cases, return projections to the sensory neocortex are via HF. **b**, An illustration of encoding in the extended model. **c**, Encoding 'scenes' from the Shapes3D dataset, with each 'scene' decomposed into unpredicted sensory features (top) and conceptual features linked to the generative network's latent variables (bottom). Novel features (white squares overlaid on the image with varying opacity) are added to each 'scene'. **d**, Recalling 'scenes' (with novel features) from the Shapes3D dataset. First, the input is decomposed; then, the MHN performs pattern completion on both sensory and conceptual features. The conceptual features (which in these simulations are simply the generative network's latent variables) are then decoded into a schema-based prediction, onto which any stored sensory features are overwritten. Diagrams were created using BioRender.com.

Furthermore, encoding every sensory detail in the hippocampus would be inefficient (elements already predicted by conceptual representations being redundant); an efficient system should take advantage of shared structure across memories to encode only what is necessary[40,41]. Accordingly, we suggest that predictable elements are encoded as conceptual features linked to the generative latent variable representation, while unpredictable elements are encoded in a more detailed and veridical form as sensory features.

Suppose someone sees an unfamiliar animal in the forest (Fig. 2b). Much of the event might be consistent with an existing forest schema, but the unfamiliar animal would be novel. In the extended model (Fig. 2 and section 'Combining conceptual and unpredictable sensory features'), the reconstruction error per element of the experience is calculated by the generative model during perception, and elements with high reconstruction error are encoded in the autoassociative network as sensory features, along with conceptual features linked to the generative model's latent variable representation. In other words, each pattern is split into a predictable component (approximating the generative network's prediction for the pattern), plus an unpredictable component (elements with high prediction error). This produces a sparser vector than storing every element in detail, increasing the capacity of the network[42].

Neural substrates of the model

Which brain regions do the components of this model represent? The autoassociative network involves the hippocampus binding together

https://doi.org/10.1038/s41562-023-01799-z

the constituents of a memory in the neocortex, whereas the generative network involves neocortical inputs projecting to latent variable representations in the higher association cortex, which then project back to the neocortex via the HF. The entorhinal (EC), medial prefrontal cortex (mPFC) and anterolateral temporal lobe (alTL) are all prime candidates for the site of latent variable representations.

First, the EC is the main route between the hippocampus and the neocortex, and is where grid cells, which are thought to be a latent variable representation of spatial or relational structure[31,54], are most often observed[73]. Second, mPFC and its connections to HF play a crucial role in episodic memory processing[70,74–78], are thought to encode schemas[57,71], are implicated in transitive inference[79] and the integration of memories[80], and perform dimensionality reduction by compressing irrelevant features[55]. Third, the anterior and lateral temporal cortices associated with semantic memory[81] and retrograde amnesia[82] probably contain latent variable representations capturing semantic structure. This might correspond to the 'anterior temporal network' associated with semantic dementia[83], while the first network (between sensory and entorhinal cortices) might correspond to the 'posterior medial network'[83], and to the network mapping between visual scenes and allocentric spatial representations[20–22].

Which regions constitute the generative network's decoder? The decoder converts latent variable representations in the higher association cortex back to sensory neocortical representations via HF. Patients with damage to the hippocampus proper but not the EC can generate simple scenes (or fragments thereof), but an intact hippocampus is required for more coherent imagery of complex ones[23]. We hypothesize that conceptual units in the hippocampus proper help to generate complex, conceptually coherent scenes (perhaps through a recurrent 'clean up' mechanism), but that an intact EC and its return pathway to the sensory neocortex (the ventral visual stream for images) can still decode representations to some extent in their absence.

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

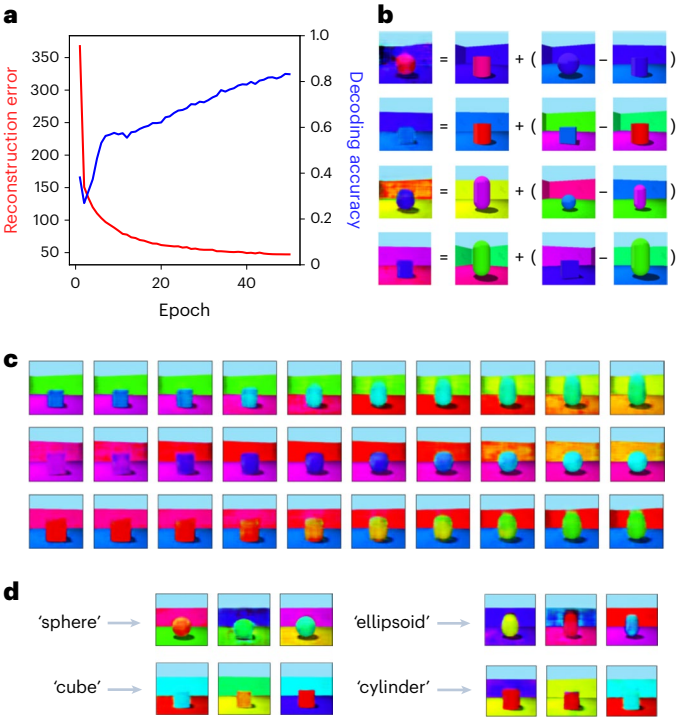

Fig. 3 | Learning, relational inference and imagination in the generative model. a, Reconstruction error (red) and decoding accuracy (blue) improve during training of the generative model. Decoding accuracy refers to the performance of a support vector classifier trained to output the central object's shape from the latent variables, using 200 examples at the end of each epoch of generative model training. An epoch is one presentation of the training set of 10,000 samples from the hippocampus. **b**, Relational inference as vector arithmetic in the latent space. The three items on the right of each equation are items from the training data. Their latent variable representations are combined as vectors according to the equation, giving the latent variable representation from which the first item is generated. The pair in brackets describes a relation which is applied to the second item to produce the first. In the top row, the object shape changes from a cylinder to a sphere. In the second, the object shape changes from a cylinder to a cube, and the object colour from red to blue. In the third and fourth, more complex transitions change the object colour and shape, wall colour and angle. **c**, Imagining new items via interpolation in latent space. Each row shows points along a line in the latent space between two items from the training data, decoded into images by the generative network's decoder. **d**, Imagining new items from a category. Samples from each of the shape categories of the support vector classifier in **a** are shown.

Results

Modelling encoding and recall

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

We model recall as (re)constructing a scene from a partial input. First, we simulate encoding and replay in the autoassociative network. The network memorizes a set of scenes, representing events, as described above. When the network is given a partial input, it retrieves the closest stored memory. Even when the network is given random noise, it retrieves stored memories (see Fig. 1c). Second, we simulate recall in the generative network trained on reactivated memories from the autoassociative network, which is able to reconstruct the original image when presented with a partial version of an item from the training data (Fig. 1d).

In the basic model (Fig. 1a), the prediction error could be calculated for each event so that only the unpredictable events are stored in the hippocampus, as the predictable ones can already be retrieved by the generative network (however, this is not simulated explicitly). In the extended model (Fig. 2 and section 'Combining conceptual and unpredictable sensory features'), prediction error is calculated for each element of an event, determining which sensory details are stored.

Modelling semantic memory

Existing semantic memory survives when the hippocampus is lesioned[43–45], and hippocampal amnesics can describe remote memories more successfully than recent ones[8,84], even if they might not recall them 'episodically'[11]. This temporal gradient indicates that the semantic component of memories becomes HF-independent. In the model, EC lesions impair all truly episodic recollection since the return projections from the HF are required for the generation of sensory experiences. Here we describe how remote memories could be retrieved 'in semantic form' despite lesions including the hippocampus and the EC.

The latent variable representation of an event in the generative network encodes the key facts about the event and can drive semantic memory directly without decoding the representation back into a sensory experience (Fig. 1g). The output route via HF is necessary for turning latent variable representations in mPFC or alTL into a sensory

https://doi.org/10.1038/s41562-023-01799-z

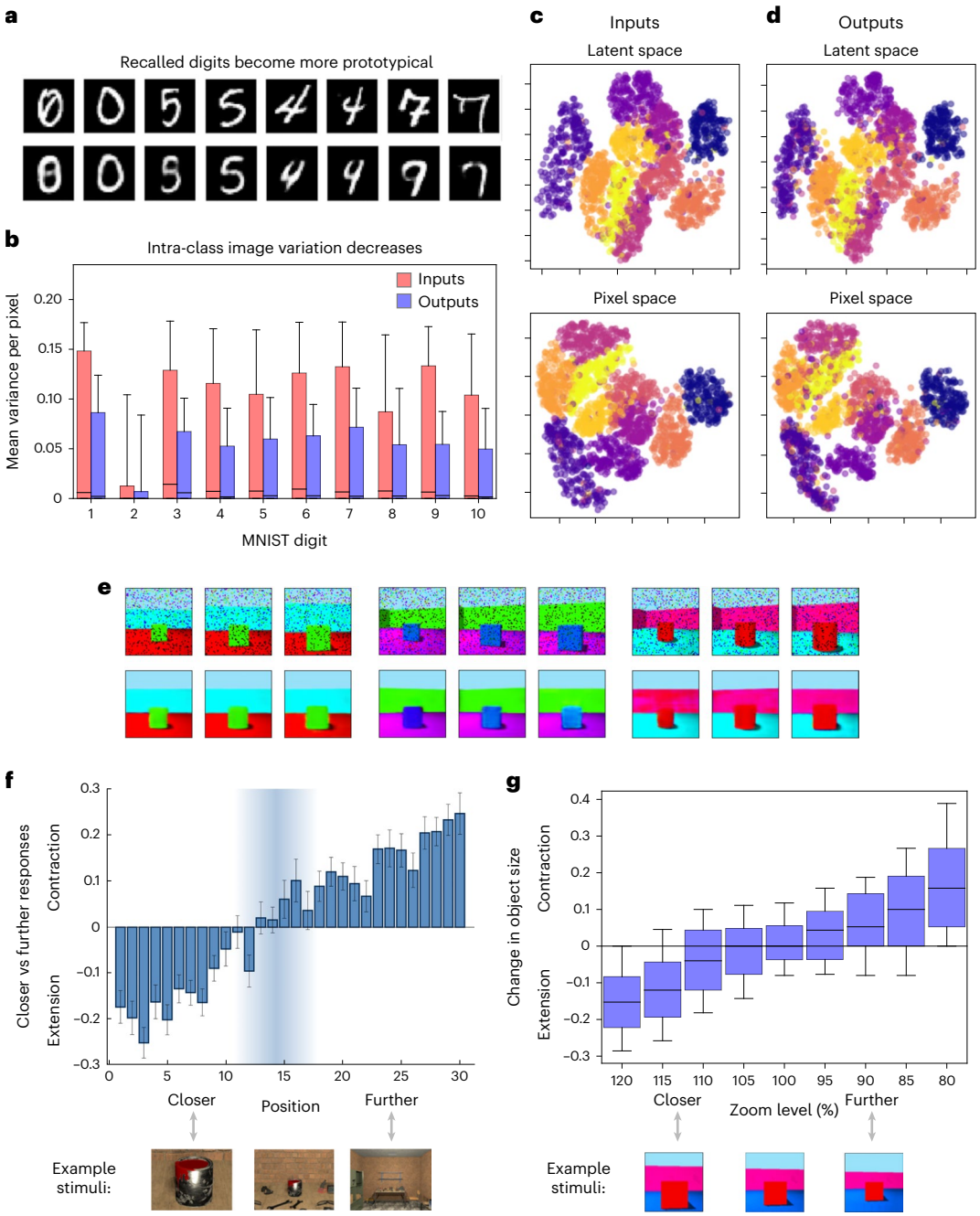

Fig. 4 | Generative network shows schema-based distortions. a, MNIST digits (top) and the VAE's output for each (bottom). Recalled pairs from the same class become more similar. A total of 10,000 items from the MNIST dataset were encoded in the MHN, and 10,000 replayed samples were used to train the VAE. **b**, The variation within each MNIST class is smaller for the recalled items than for the original inputs. For each of the 10 classes, the variance per pixel is calculated across 500 images, and the 784 pixel variances are then plotted for each class before and after recall. In each boxplot, the box gives the interquartile range, its central line gives the median, and its whiskers extend to the 10th and 90th percentiles of the data. **c**,**d**, The pixel spaces of MNIST digits (bottom row) and the latent space of their encodings (top 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[146] and colour-coded by class. **e**, Examples of boundary extension and contraction. Top row: the noisy input images (from a held-out test set), with an atypically 'zoomed out' or 'zoomed in' view (by 80% and 120% on the left and right, respectively) for three original images. Bottom row: the predicted images for each input image, which are distorted towards the 'typical view' in each case. **f**, Adapted figure from ref. 92, showing the distribution of boundary extension vs contraction as a function of the viewpoint of an image. Specifically, the values are the average of 'closer' vs 'further' judgements, assigned −1 and 1, respectively, of an identical stimulus image in comparison with the remembered image (with 900 trials per position). Error bars give the standard error of the mean. Example stimuli are shown at the bottom. **g**, In our model, the VAE increases the estimated size of the central object in atypically 'zoomed out' views compared with the training data, and decreases it in atypically 'zoomed in' views, as in ref. 92. Two hundred images are used at each 'zoom level'. See **b** for a description of boxplot elements.

experience, but the latent variables themselves could support semantic retrieval. Thus, when the HF (including the EC) is removed, the model can still support retrieval of semantic information (see section 'Modelling brain damage' for details). To show this, we trained models to predict attributes of each image from its latent vector. Figure 3a shows that semantic 'decoding accuracy' increases as training progresses, reflecting the learning of semantic structure as a by-product of learning to reconstruct the sensory input patterns ($r_s(48) = 0.997$, $P < 0.001$,

https://doi.org/10.1038/s41562-023-01799-z

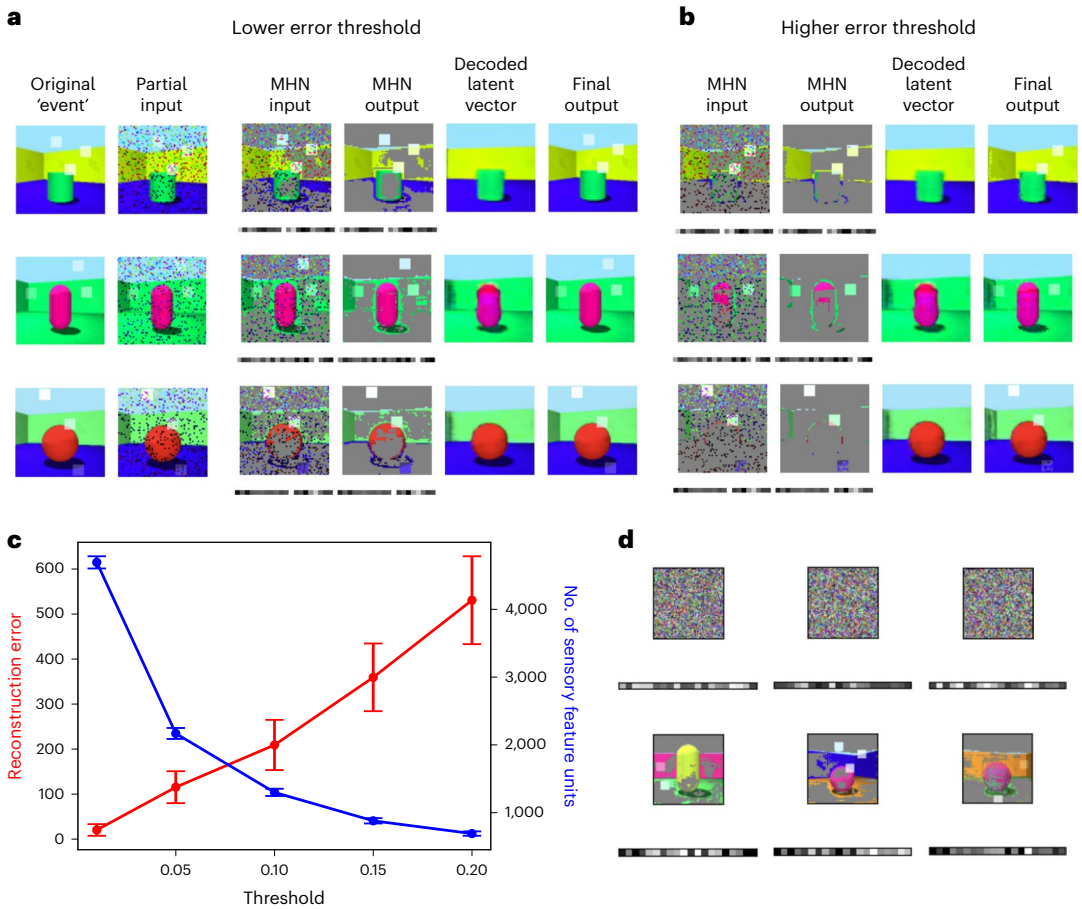

Fig. 5 | Retrieval dependence on reconstruction error threshold and replay in the extended model. a, The stages of recall are shown from left to right (see Fig. 2d), where each row represents an example scene. Each scene consists of a standard Shapes3D image with the addition of novel features (several white squares overlaid on the image with varying opacity). **b**, Repeating this process with a higher error threshold for encoding (with the same events and partial inputs) means fewer poorly predicted sensory features are stored in the autoassociative MHN, leading to more prototypical recall with increased reconstruction error. **c**, Average reconstruction error and number of sensory features (that is, pixels) stored in the autoassociative MHN against the error threshold for encoding. One hundred images are tested and error bars give the s.e.m. **d**, Replay in the extended model. The autoassociative network retrieves memories when random noise is given as input, as shown for three example inputs (upper row). As above, the square images show the poorly predicted sensory features and the rectangles below these display the latent variable representations (lower row).

95% confidence interval (CI) = 0.987, 1.000). While semantic memory is much more complex than simple classification, richer 'semantic' outputs such as verbal descriptions can also be decoded from latent variable representations of images[85,86].

Imagination, episodic future thinking and relational inference

Here we model the generation of events that have not been experienced from the generative network's latent variables. Events can be generated either by external specification of latent variables (imagination) or by transforming the latent variable representations of specific events (relational inference). The former is simulated by sampling from categories in the latent space then decoding the results (Fig. 3d). The latter is simulated by interpolating between the latent representations of events (Fig. 3c) or by doing vector arithmetic in the latent space (Fig. 3b). This illustrates that the model has learnt some conceptual structure to the data, supporting reasoning tasks of the form 'what is to *A* as *B* is to *C*?', and provides a model for the flexible recombination of memories thought to underlie episodic future thinking[24].

Modelling schema-based distortions

The schema-based distortions observed in human episodic memory increase over time[6] and with sleep[52], suggesting an association with consolidation. Recall by the generative network distorts memories towards prototypical representations. Figure 4a–d shows that handwritten digits from the MNIST dataset[87] 'recalled' by a VAE become more prototypical (MNIST is used for this because each image has a single category). Recalled pairs from the same class become more similar, that is, intra-class variation decreases (paired samples *t*-test $t(7,839) = 60.523$, $P < 0.001$, Cohen's $d = -0.684$, 95% CI = 0.021, 0.022). The pixel space of MNIST digits before and after recall and the latent space of their encodings also show this effect. In summary, recall with a generative network distorts stimuli towards more prototypical representations even when no class information is given during training. As reliance on the generative model increases, so does the level of distortion.

Boundary extension and contraction exemplify this phenomenon. Boundary extension is the tendency to remember a wider field of view than was observed[88], while boundary contraction is the opposite[89]. Unusually close-up views appear to cause boundary extension, and unusually far away ones boundary contraction[89], although this is debated[90,91]. We modelled this by giving the generative network a range of new scenes that were artificially 'zoomed in' or 'zoomed out' compared with those in its training set; its reconstructions are distorted towards the 'typical view' (Fig. 4e), as in human data. Figure 4g shows the change in the object size in memory quantitatively, mirroring the

https://doi.org/10.1038/s41562-023-01799-z

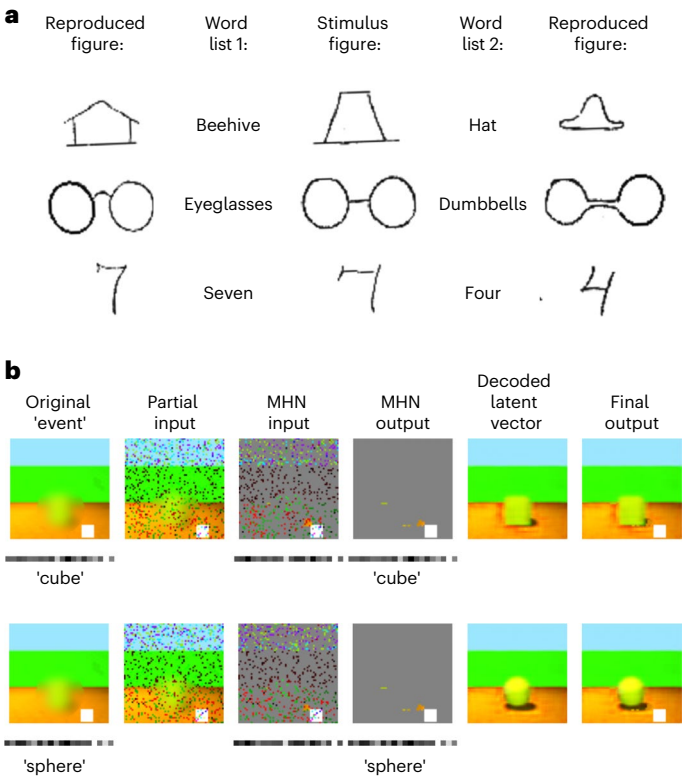

Fig. 6 | Schema-based distortions: effects of conceptual context in the extended model. a, Adapted figure from ref. 95 showing that recall of an ambiguous item (stimulus figure, centre) depends on its context at encoding (word from list 1, left; or list 2, right), as shown by drawing from memory (reproduced figure, far left and far right). **b**, Memory distortions in the extended model, when the original scene (containing an ambiguous blurred shape) is encoded with a given concept (cube, top; sphere, bottom), represented by the latent variables for that class. Then, a partial input is processed by the generative network to produce predicted conceptual features and the sensory features not predicted by the prototype for that concept (in this case, a white square) for input to the autoassociative MHN. However, pattern completion in the MHN reproduces the originally encoded sensory and conceptual features (cube, top; sphere, bottom), and these are recombined to produce the final output, which is distorted towards the encoded conceptual context.

findings in ref. 92 (Fig. 4f). (Note that the measure of boundary extension vs contraction used by ref. 92 is produced by averaging 'closer' vs 'further' judgements of an identical stimulus image in comparison with the remembered image, rather than the drawing-based measure we use, but the two measures are significantly correlated[89].)

Combining conceptual and unpredictable sensory features

In the extended model, memories stored in the hippocampal autoassociative network combine conceptual features (derived from the generative network's latent variables) and unpredictable sensory features (those with a high reconstruction error during encoding) (Fig. 2). In these simulations, the conceptual features are simply a one-to-one copy of latent variable representations. (Since latent variable representations are not stable as the generative network learns, concepts derived from latent variables seem more likely to be stored than the latent variables themselves, so this is a simplification; see section 'Extended model' for further details.)

Figure 5a,b shows the stages of recall in the extended model after encoding with a lower or higher prediction error threshold. After decomposing the input into its predictable (conceptual) and unpredictable (sensory) features, the autoassociative network performs pattern completion on the combined representation. The prototypical (that is, predicted) image corresponding to the retrieved conceptual features must then be obtained by decoding the associated latent variable representation into an experience via the return projections to the sensory neocortex. Next, the predictable and unpredictable elements are recombined, simply by overwriting the prototypical prediction with any unpredictable elements, via the connections from the sensory features to the sensory neocortex. The extended model is therefore able to exploit the generative network to reconstruct the predictable aspects of the event from its latent variables, storing only those sensory details that were poorly predicted in the autoassociative network. Equally, as the generative network improves, sensory features stored in the hippocampus may no longer differ significantly from the initial schematic reconstruction in the sensory neocortex, signalling that the hippocampal representation is no longer needed.

Schema-based distortions in the extended model

The schema-based distortions shown in the basic model result from the generative network and increase with dependence on it, but memory distortions can also have a rapid onset[93,94]. In the extended model, even immediate recall involves a combination of conceptual and sensory features, and the presence of conceptual features induces distortions before consolidation of that specific memory.

In general, recall is biased towards the 'mean' of the class soon after encoding due to the influence of the conceptual representations (Fig. 5a,b). This is more pronounced when the error threshold for encoding is high, as there is more reliance on the 'prototypical' representations, resulting in the recall of fewer novel features. At a lower error threshold, more sensory detail is encoded, that is, the dimension of the memory trace is higher ($r_s(3) = -1$, $P < 0.001$). This results in a lower reconstruction error ($r_s(3) = 1$, $P < 0.001$), indicating lower distortion but at the expense of efficiency.

External context further distorts memory. Reference [95] asked participants to reproduce ambiguous sketches. A context was established by telling the participants that they would see images from a certain category. After a delay, drawings from memory were distorted to look more like members of the context category. Figure 6b shows the result of encoding the same ambiguous image with two different externally provided concepts (a cube in the top row, a sphere in the bottom row), represented by the latent variables for each concept, as opposed to the latent variables predicted by the image itself as in Fig. 5a,b. During recall, the encoded concept is retrieved in the autoassociative network, determining the prototypical scene reconstructed by the generative network. This biases recall towards the class provided as context, mirroring Fig. 6a.

We also simulate the Deese–Roediger–McDermott (DRM) task[93,94] in the extended model to demonstrate its applicability to non-image stimuli. In the DRM task, participants are shown lists of words that are semantically related to 'lure words' not present in the list; there is a robust finding that false recognition and recall of the lure words occur[93,94]. In the extended model, gist-based semantic intrusions arise as a consequence of learning the co-occurrence statistics of words. First, the VAE is trained to reconstruct the sets of words in simple stories[96] converted to vectors of word counts, representing background knowledge. The system then encodes the experimental lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent representation of each word list (respectively analogous to the stimulus-unique pixels and the VAE's latent representation of each image in Fig. 5). As in the human data, lure words are often but not always recalled when the system is presented with 'id_n' (Fig. 7a), since the latent variable representations that generate the words in the list also tend to generate the lure word. The system also forgets some words and produces additional semantic intrusions. In addition, the chance of recalling the lure word is higher for longer lists, as in human

https://doi.org/10.1038/s41562-023-01799-z

data from ref. 97, as more related words provide a stronger 'prior' for the lure (Fig. 7b) ($r_s(10) = 0.998$, $P < 0.001$, 95% CI = 0.982, 1.000).

Modelling brain damage

Recent episodic memory is impaired following damage to the HF, whereas semantic memory, including the semantic content of remote episodes, appears relatively spared. In the model, the semantic form of a consolidated memory survives damage to the HF due to latent variable representations in the mPFC or the alTL (even if those in the EC are lesioned); Fig. 3a demonstrates how semantic recall performance improves with the age of a memory, reflecting the temporal gradient of retrograde amnesia (see section 'Modelling semantic memory'). However, these semantic 'facts' cannot be used to generate an experience 'episodically' without the generative network's decoder, in agreement with multiple trace theory[11].

The extent of retrograde amnesia can vary greatly depending in part on which regions of the HF are damaged[98,99]. The dissociation of retrograde and anterograde amnesia in some cases suggests that the circuits for encoding memories and the circuits for recalling them via the HF only overlap partially[99]. For example, if the autoassociative network is damaged but not the generative network's decoder, the generative network can still perform reconstruction of fully consolidated memories. This could explain varying reports of the gradient of retrograde amnesia when assessing episodic recollection (as opposed to semantic memory), if the generative network's decoder is intact in patients showing spared episodic recollection of early memories[45]. Note that the location of damage within the generative network's decoder also affects the resulting deficit in our model. In particular, patients with damage restricted to the hippocampus proper can (re) construct simple scenes but not more complex ones[23].

Our model also shows the characteristic anterograde amnesia after hippocampal damage, as the hippocampus is required to initially bind features together and support off-line training of the generative model. Anterograde semantic learning would also be impaired by hippocampal damage (as the generative network is trained by hippocampal replay). While hippocampal replay need not be the only mechanism for schema acquisition, it would probably be much slower without the benefit of replay. However, semantic learning over short timescales may be relatively unimpaired, as it is less dependent on extracting regularities from long-term memory[100].

In semantic dementia, semantic memory is impaired, and remote episodic memory is impaired more than recent episodic memory[101]. This would be consistent with lesions to the generative network, as recent memories can rely more on the hippocampal autoassociative network. However, the exact effects would depend on the distribution of damage across the various potential generative networks in the EC, mPFC and alTL. Of these, the alTL network is associated with semantic dementia, and the posterior medial network (corresponding to the generative network between the sensory areas and the EC) with Alzheimer's disease[83].

Finally, neuropsychological evidence suggests a distinction between familiarity and recollection, and furthermore a partial dissociation between different tests of familiarity; patients with selective hippocampal damage can exhibit recognition memory deficits in a simple 'yes/no' task with similar foils, but not in a 'forced choice' variant involving choosing the more familiar stimulus from a set[102]. This is consistent with the idea that lower prediction error in the neocortical generative network indicates familiarity, but retrieval of unique details from the hippocampus is required for more definitive recognition memory.

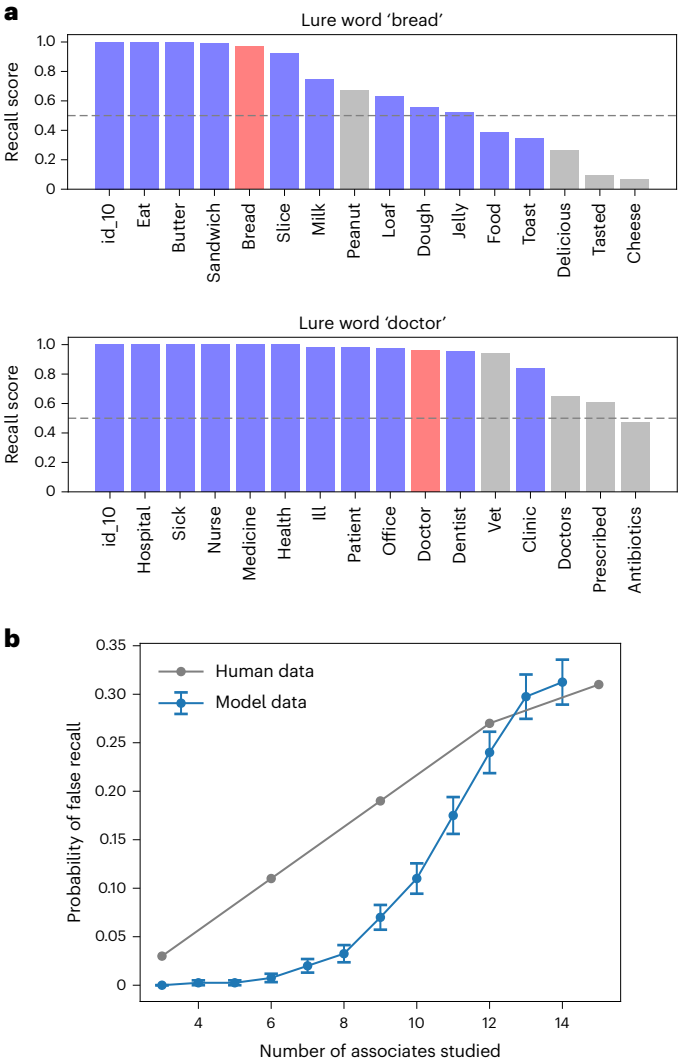

Fig. 7 | Modelling the DRM task. a, First, the VAE is trained to reconstruct simple stories[96] converted to vectors of word counts, representing background knowledge. The system then encodes the lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent variable representation of the word list. In each plot, recalled stimuli when the system is presented with 'id_n' are shown, with output scores treated as probabilities so that words with a score >0.5 (above dashed lines) are recalled. Words from the stimulus list are shown in blue, and lures in red. See Fig. 1 of Supplementary Information for results for the remaining DRM lists. **b**, The chance of recalling the lure word is higher when longer lists are encoded (blue). Each measurement is averaged across 400 trials (20 random subsets of each of the 20 DRM lists), and error bars give the s.e.m. This qualitatively resembles human data from ref. 97 (grey).

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 and inference. This occurs through teacher–student learning. The hippocampal 'teacher' rapidly encodes an event, which may combine unpredictable sensory elements (with connections to and from the sensory cortex) and predictable conceptual elements (with connections to and from latent variable representations in the generative network). After exposure to replayed representations from the 'teacher', the generative 'student' network supports reconstruction of events by forming a schematic representation in the sensory neocortex from latent variables via the HF, with unpredictable sensory elements added from the hippocampus.

In contrast to the relatively veridical initial encoding, the generative model learns to capture the probability distributions underlying experiences, or 'schemas'. This enables not just efficient recall, reconstructing memories without the need to store them individually, but also imagination (by sampling from the latent variable distributions) and inference (by using the learned statistics of experience to predict

the values of unseen variables). In addition, semantic memory (that is, factual knowledge) develops as a by-product of learning to predict sensory experience. As the generative model becomes more accurate, the need to store and retrieve unpredicted details in the hippocampus decreases (producing a gradient of retrograde amnesia in cases of hippocampal damage). However, the generative network necessarily introduces distortion compared to the initial memory system. Multiple generative networks can be trained in parallel, and we expect this to include networks with latent variables in the EC, mPFC and alTL.

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

1. Gradual consolidation follows one-shot encoding: A memory is encoded in the hippocampal 'teacher' network after a single exposure, and transferred to the generative 'student' network after being replayed repeatedly (Fig. 1c,d).
2. Semantic memory becomes hippocampus-independent: The latent variable representations learned by the generative networks constitute the 'key facts' of an episode, supporting semantic memory (Fig. 3a).
3. Episodic memory remains hippocampus-dependent: Return projections to the sensory neocortex via the HF are required to decode the latent variable representations into a sensory experience (Fig. 1). (EC is required for even simple (re)construction, while the hippocampus proper helps to generate complex conceptually coherent scenes and retrieves unpredictable details that are not yet consolidated into the generative network; see section 'Neural substrates of the model'.)
4. Shared substrate for episode generation: Generative models are a common mechanism for episode generation. Familiar scenes can be reconstructed and new ones can be generated by sampling or transforming existing latent variable representations (Fig. 3b–d), providing a model for imagination, scene construction and episodic future thinking.
5. Consolidation promotes inference and generalization: Relational inference corresponds to vector arithmetic applied to the generative network's latent variables (Fig. 3b).
6. Episodic memories are distorted: We show how memory distortions arise from the generative network (Figs. 4, 6 and 7). This extends the model of ref. 32 to relate memory distortion to consolidation.
7. Association cortex encodes latent structure: Latent variable representations in the EC, mPFC, and alTL provide schemas for episodic recollection and imagination (via HF) and for semantic retrieval and inference.
8. Prediction error affects memory processing: The generative network is constantly calculating the reconstruction error of experiences[58,59]. Events that are consistent with the existing generative model require less encoding in the autoassociative hippocampal network (Fig. 5).
9. Episodic memories include conceptual features: When an experience combines a mixture of familiar and unfamiliar elements, both concepts and poorly predicted sensory elements are stored in the hippocampus via association to a specific memory unit.

Our model can be seen as an update to the complementary learning systems (CLS)[4] framework to better account for points 3 to 9 above, reconciling the development of semantic representations in the neocortex (as in CLS) with the continued dependence on the hippocampal formation for episodic recall (as in multiple trace theory[11]). Furthermore, it provides a unified view of: (1) episode generation, (2) how episodic memories change over time and exhibit distortions and (3) how semantic and episodic information are combined in memory. We build on previous work exploring the role of generative networks in consolidation[18,19], as models of the hippocampal formation[31–33], as priors for episodic memory[35] and as models of spatial cognition[22].

A key aspect of the model is that multiple generative networks can be trained concurrently from a single autoassociative network (Fig. 2a) and may be optimized for different tasks. Thus, the latent representations in the mPFC and the alTL may be more closely linked to value or language than those in the EC[103,104]. These differences may arise from differences in network structure (for example, the degree of compression) or from additional training objectives that shape their representations[105] (for example, the generative network with latent variables in the mPFC might be trained to predict task-relevant value in addition to the EC representations). We expect the generative networks to overlap closer to their sensory inputs/outputs, where general-purpose features are more useful, and diverge as the representations become more abstract (or task-specific if there are additional training objectives)[106]. This may involve a primary VAE with latent variables in the EC, with additional pathways from the higher sensory cortex to the EC routed via latent variables in the mPFC or the alTL.

Our model raises some fundamental questions: Does true episodic memory require event-unique detail, and does this require the hippocampus? Or can prototypical predictions qualify as memory rather than imagination? In the model, event-unique details are initially provided by the hippocampus but can also be provided by the generative network. For example, if you know that someone attended your 8th birthday party and gave you a particular gift, these personal semantic facts need not be hippocampal-dependent but could generate a scene with the right event-specific details, which would seem like episodic memory. The increasingly sophisticated generation of images from text using generative models[107] suggests that episode construction from semantic facts is computationally plausible.

Episodic memories are defined by their unique spatiotemporal context[1]. In the model, spatial and temporal context correspond to conceptual features captured by place[108,109] or time[110,111] cells in the hippocampus and might be linked to latent variable representations formed in the EC, such as grid cells in the medial EC, which form an efficient basis for locations in real[31,112,113] or cognitive spaces[31,54], or temporal context representations in the lateral EC[114,115]. Events with specific spatial and temporal context can be generated from these latent variable representations, as has been modelled in detail for space[20–22].

More generally, this work builds on the spatial cognition literature, in which place and head direction cells act as latent variables in a generative model[20–22], allowing the generation of a scene from a specific viewpoint. References [20–22] explore how egocentric sensory representations could be transformed into allocentric latent variables before storage in the medial temporal lobe and conversely, how egocentric representations could be reconstructed from allocentric ones to support imagery. The latent representations learned through consolidation in our model correspond loosely to the allocentric representations, and the sensory representations produced by HF to the egocentric ones; only egocentric and sensory representations are directly experienced, whereas allocentric and semantic representations are useful abstractions that can also be exploited for efficient hippocampal encoding.

Our model simplifies the true nature of mnemonic processing in several ways. First, the interaction of sensory and conceptual features in the hippocampus and latent variables in the EC during retrieval could be more complex, with each type of representation contributing to pattern completion of the other as in interactions between items and contextual representations in the Temporal Context Model[116], and might iterate over retrievals from both hippocampal and generative networks[50]. Second, our model distinguishes between 'sensory' and 'conceptual' representations in the hippocampus, respectively linked to the sensory neocortex at the input/output of the generative network and to the latent variable layer in the middle. In reality, a gradient of levels of representation in the hippocampus is more likely, from detailed sensory representations to coarse-grained conceptual ones, respectively linked to lower or higher neocortical areas[117], and might map onto

the observed functional gradients along the longitudinal axis of the hippocampus[118]. Third, our generative network uses back-propagation of the prediction error between output and input patterns to learn. Generative networks with more plausible (if less efficient) learning rules exist[67–69], which have the advantage of producing a prediction error signal at each layer (between top–down prediction and bottom–up recognition), potentially allowing learning of concepts and exceptions at all levels of description. Fourth, considering consolidation as a continual lifelong process rather than during encoding of a single dataset introduces new complexities; these include the instability of latent representations and the prevention of catastrophic forgetting of already consolidated memories as new memories are assimilated into the generative network. The model could be extended to address this, for example, by using replay from the generative network as well as from the hippocampal network, which could reduce catastrophic forgetting and stabilize latent variable representations in both networks[33,119,120], building on previous research on sleep and learning[121]. Fifth, we model semantic memory as prediction of categorical information for an 'event', but future work should model more complex semantic knowledge, for example, by decoding language from latent representations of multimodal stimuli[85,86]. In particular, the relationship between semantic memory for specific 'events' and the broader 'web' of general knowledge should be considered.

Episodic memories contain important sequential structure not modelled by our encoding and reconstruction of simple scenes. Future work could expand the model's scope to sequential information as follows. A range of stimuli could be represented as sequences of arbitrary symbols (including language, spatial trajectories and transitions on a graph). A heteroassociative variant of an MHN, which is better suited to sequential data, could be used to store such stimuli. Specifically, the interpretation of an MHN that we use[64] can capture sequential information if the projections from feature units to memory units correspond to the current state, but the projections from memory units back to feature units correspond to the next state so that one state retrieves the next[122–124]. With certain modifications based on previous work involving the role of temporal context in memory[116,125], asymmetric MHNs can store sequences with complex repetitions and temporal correlations, such as language. We could then implement the student model as a sequential generative network trained to predict the next input during sequential replay (for example, GPT-2 (ref. 126)). Such networks capture relational structure, developing grid-like latent representations in spatial tasks[31], and learn the gist of narratives. The sequential model could also be applied to phenomena such as event segmentation[127] and memory distortions in narratives[6]. (Note that for more complex sequential data such as videos, pattern completion of both the current stimulus and the next stimulus would be required, potentially needing a combination of autoassociative and heteroassociative connectivity in the hippocampal network.)

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

Second, the model makes multiple predictions about the effects of brain damage. Just as boundary extension is reduced in patients with damage to the HF[128] or the vmPFC[129], we predict that other biases towards the 'canonical view' would be attenuated in such patients; for example, healthy controls would distort images with an atypical viewing angle towards a more typical angle in memory, but this would be reduced in, for example, hippocampal patients. Similarly, ambiguous images such as the duck/rabbit drawing 'flip' between interpretations in perception but are stable when held in imagery[72], presumably due to maintained hippocampal conceptual representations. We predict that this conceptual stability in imagery would also be reduced in such patients. This could also extend to non-scene stimuli: if the ref. 95 task were tested with both healthy controls and patients with damage to the generative decoder, we would predict reduced contextual distortion in the latter. Furthermore, patients with an inaccurate generative model, for example, due to semantic dementia, might rely more on sensory features to compensate. (Note that the pattern of deficits would depend on both the nature of the priors encoded in the generative network and the error threshold for encoding. In some cases, damage to the generative network could produce atypical 'priors' rather than suppressing them. Thus, if the generative network is inaccurate but the error threshold for encoding is high, atypical distortions will be observed rather than a reduction in conceptual distortions.)

Third, the model suggests that the error threshold for encoding could vary depending on the importance of the stimuli or the amount of attentional resource available. For example, emotional salience could lower this threshold, with traumatic memories being encoded in greater sensory detail and with less contextual coherence[130,131]. Equally, conditions such as autism spectrum disorder, which are potentially attributable to hypo-priors[132], might be associated with a lower prediction error threshold for veridical storage (and thus reduced conceptual influence on memory and increased sensory detail). In addition, reality monitoring deficits would change the perceived prediction error relative to reality, leading to atypical memory storage (for example, a reduced ability to compensate for prediction errors by storing sensory details).

Fourth, biological intelligence excels at generalizing from only a small number of examples. The model predicts that learning to generalize rapidly benefits from having a generative model that can create new examples, for example, by inferring variants (as in Fig. 3b) (see also ref. 133). Finally, the model suggests a link between latent spaces and cognitive maps[134]. For example, one might predict that the position of a memory in latent space is reflected in place and grid cell firing, as observed for other conceptual representations[54,134,135].

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, its relationship to schemas, and consolidation, as well as aspects of imagination, inference and semantic memory.

Methods

Data

In the simulations, images represent events (except for the DRM[93,94] task stimuli). The Shapes3D dataset[136] was used throughout, except for the use of MNIST[87] to explore certain distortions. Note that one MHN was used per dataset, and one generative model was trained per dataset from the corresponding MHN's outputs.

Basic model

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

The generative networks used are variational autoencoders, a type of autoencoder with special properties such that randomly sampling values for the latent variables in the model's 'bottleneck' layer generates valid stimuli[65]. Figure 3 of Supplementary Information, adapted

from ref. 137, shows how directions in the latent space can correspond to meaningful transformations. While most diagrams show the VAE's input and output layers in the sensory neocortex as separated (in line with conventions for visualizing neural networks), it is important to note that the input and output layers are in fact the same, as shown in Fig. 1b. There may be considerable overlap between the encoder and decoder, especially closer to the sensory neocortex, but we did not model this explicitly. The autoassociative model is an MHN, with the property that even random input values will retrieve one of the stored patterns via pattern completion. Specifically, we considered the biological interpretation of the MHN as feature units and memory units suggested by ref. 64 (see Supplementary Information for details).

We modelled consolidation as teacher–student learning, where the autoassociative network is the 'teacher' and the generative network is the 'student' trained on replayed representations from the 'teacher'. We gave random noise (consisting of uniformly sampled values in each channel for each pixel) as an input to the MHN, then used 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 the sensory cortex.) The noise input to the autoassociative network could potentially represent random activation during sleep[138–140]. Attributes such as reward salience might also influence which memories are replayed but were not modelled here[141].

During the encoding state in our simulations, images were stored in a continuous MHN with high inverse temperature, β, set to 20 (higher values of β produce attractor states corresponding to individual memories, while lower values of β make metastable states more likely). Reference [63] provides an excellent Python implementation of MHNs that we used in our code. During the 'rest' state, random noise was given as an input N times to the MHN, retrieving N attractor states from the network. (The distribution of retrieved attractor states was not tested but was approximately random, and very few spurious attractors were observed with sufficiently high inverse temperature.) In the main simulations, 10,000 items from the Shapes3D dataset were encoded in the MHN, and 10,000 replayed states were used to train the VAE (that is, N is 10,000). (Rather than replaying new samples from the MHN at each epoch of the VAE's training, a single set of samples was used for efficiency and simplicity.)

A VAE was then trained on the 'replayed' images from the MHN, using the Keras API for TensorFlow[142]. The loss function (that is, the error minimized through training) is the sum of two terms, the reconstruction error and the Kullback–Leibler divergence[65]; the former encourages accurate reconstruction, while the latter (which measures the divergence between the latent variables and a Gaussian distribution) encourages a latent space one can sample from. Specifically, the reconstruction loss in our model is a mean absolute error loss. (Note that the terms reconstruction error and prediction error are used interchangeably throughout.)

The stochastic gradient descent method used was the AMSGrad variant of the Adam optimizer with early stopping enabled, for a maximum of 50 epochs (where an epoch is a complete pass through the training set). A latent variable vector length of 20, learning rate of 0.001 and Kullback–Leibler weighting of 1 were used in the main results. The variational autoencoders were not optimized for performance, as their purpose is illustrative (more data and hyperparameter tuning would be likely to improve reconstruction accuracy). Architectural choices within the VAE were not principled but were based on successful architectures for similar stimuli in the literature. See Supplementary Information for details of the VAE's architecture. The VAEs were trained using gradient descent and back-propagation as usual; while this method is biologically implausible due to its non-local nature, more plausible learning algorithms might be feasible[143].

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

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

While we used only one modality at a time (imagery for the majority of simulations, text for the DRM task), our model is compatible with the multimodal nature of experience, as multimodal inputs to VAEs are possible, which result in a multimodal latent space[144]. This could reflect the multimodal nature of concept cells in the hippocampus[61].

Modelling semantic memory

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

Decoding accuracy was measured by training a support vector machine to classify the central object's shape from the network's latent variables, using 200 examples at the end of each epoch and measuring classification accuracy on a held-out test set. (Notably, there was good performance with only a small amount of training data when decoding the latent variables, compared with decoding alternative representations such as the sensory input or intermediate layer activations, that is, few-shot learning is possible by making use of compressed 'semantic' representations. See Fig. 2 of Supplementary Information.)

Modelling imagination and inference

In the generative network, new items can either be generated from externally specified (or randomly sampled) latent variables (imagination), or by transforming the latent variable representations of specific events (relational inference). The former was simulated by sampling from categories in the latent space, then decoding the results (Fig. 3d). The latter was simulated by interpolating between the latent representations of events (Fig. 3c) or by doing vector arithmetic in the latent space (Fig. 3b).

Examples of the four different object shapes were generated by Monte Carlo sampling for simplicity, that is, samples from the latent space were classified by the semantic decoding classifier, and examples that activate each category are displayed. (Note that there are many alternative ways to do this, for example, by extracting the decision boundaries from the classifier and sampling within the region corresponding to each class.) Generating imagined scenes from more naturalistic inputs, for example, natural language descriptions, would require a much more sophisticated text to the latent space model, but recent machine learning advances suggest that this is possible[145].

To demonstrate interpolation, each row of Fig. 3c shows items generated from latent variables along a line in the latent space between two real items from the training data. To demonstrate vector arithmetic, each equation in Fig. 3b shows 'result = vector_A + (vector_B − vector_C)' (reflecting relational inference problems of the form 'what is to A as B is to C?'), where the result is produced by taking the relation between vector_B and vector_C, applying that to vector_A and decoding the result. In other words, the three items on the right of each equation in Fig. 3b are real items from the training data. Their latent variable representations are combined as vectors according to the equation shown, giving the latent variable representation from which the first item is generated. Thus, the pair in brackets describes a relation that is applied to the first item on the right to produce the new item on the left of the equation.

Modelling schema-based distortions

Items recalled by the generative network become more prototypical, a form of schema-based distortion. This can be shown simply in the basic

model, using the MNIST digits dataset[87] to exemplify ten clearly defined classes of items (Fig. 4). To show this quantitatively, we calculated the intra-class variation, measured as the mean variance per pixel, within each MNIST class before and after recall, for 5,000 images from the test set. As expected, the intra-class variation was smaller for the recalled items than for the original inputs. (See Supplementary Information for details of the model architecture.)

To visualize this, we projected the pixel and latent spaces before and after recall (of 2,000 images from the MNIST test set) into two dimensions (2D) with uniform manifold approximation and projection (UMAP)[146], a dimensionality reduction method, and colour-coded them by class (Fig. 4c,d). The pixel space of MNIST digits (bottom row) and the latent space of their encodings (top row) showed more compact clusters for the generative network's outputs (Fig. 4d) than for its inputs (Fig. 4c).

Modelling boundary extension and contraction

Boundary extension is the tendency to remember a wider field of view than was observed for certain stimuli[88], while boundary contraction is the tendency to remember a narrower one[89]. Whether boundaries are extended or contracted seems to depend on the perceived distance of the central object, with unusually close-up (that is, 'object-oriented') views causing boundary extension, and unusually far away (that is, 'scene-oriented') views causing boundary contraction[89].

We tested boundary extension and contraction in the basic model by giving it a range of artificially 'zoomed in' or 'zoomed out' images, adapted from Shapes3D scenes not seen during training, and observing the outputs. The 'zoomed in' view was produced by removing n pixels from the margin. The 'zoomed out' view was produced by extrapolating the pixels at the margin outwards by n additional pixels. (In both cases, the new images were then resized to the standard size.) The zoom level is the ratio of the central object size in the output image to the size in the original image, given as a percentage; for example, an image with a zoom level of 80% or a ratio of 0.8 was produced by adding a margin so that the object size is 80% of the original size. As the Shapes3D images are of width and height 64, the number of pixels to add or remove was calculated as 'margin = (32/ratio) − 32'.

In Fig. 4g, the change in object size between the noisy input and output was estimated as follows: first the image was converted to a few colours by k-means clustering of pixels. Then, the colour of the central object was determined by finding the predominant colour in a particular central region of the image. A 1D array of pixels corresponding to a vertical line at the horizontal midpoint of the image was processed to identify the fraction of pixels of the central object colour. This enabled us to calculate the change in object size, which we plotted against the degree of 'zoom'. (For this object size estimation approach to work, we filtered the Shapes3D dataset to images where the object colour was different from both the wall and floor colour, and additionally to cubes to minimize shadow.)

Note that the measure of boundary extension vs contraction displayed in Fig. 4f, reproduced from ref. 92, was not based on the degree of distortion, but was produced by averaging 'closer' vs 'further' judgements of an identical stimulus image in comparison to the remembered image. This differs from our measure in Fig. 4g, which instead corresponds to the drawing-based measure in ref. 89; however, these measures have been shown to be correlated[89].

Figure 4e shows a few examples of boundary extension and contraction. In the left- and right-hand images of each set, the margin n is chosen such that the central object is 80% and 120% of the original size, respectively.

Extended model

The extended model was designed to capture the fact that memory traces in the hippocampus bind together a mixture of sensory and conceptual elements, with the latter encoded by concept cells[61], and the fact that schemas shape the reconstruction of memories even before consolidation, as shown by the rapid onset of schema-based distortions[93,94].

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

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

How did recall work before full consolidation? After decomposing the input into its predictable (conceptual) and unpredictable (sensory) components, as described above, the autoassociative network could retrieve a memory. The image corresponding to the conceptual component was then obtained by decoding the stored latent variables. Next, the predictable and unpredictable elements were recombined, simply by overwriting the initial schematic reconstruction in the sensory neocortex with any stored (that is, non-zero) sensory features in the hippocampus. Figure 5a,b shows this process. The lower the error threshold for encoding sensory details, the more information was stored in the autoassociative network, reducing the reconstruction error of recall (see also section 'Modelling schema-based distortions').

How did replay work? When the autoassociative network was given random noise, both the unpredictable elements and the corresponding latent variables were retrieved. In Fig. 5d, the square images show the unpredictable elements of MNIST images and the rectangles below these display the vector of latent variables. (As the generative model improves, the presence of hippocampal sensory features that no longer differ from the initial reconstruction indicates that the hippocampal representation is no longer needed, but this was not simulated explicitly.)

We note that the latent variable representation is not stable as the generative network learns. If some latent variables are stored in the autoassociative network while the VAE continues to change, the quality of the VAE's reconstruction will gradually worsen; this is also a feature of previous models[42]. Some degree of degradation may reflect forgetting, but consolidation can be a prolonged process and hippocampal representations can persist in this time. Therefore, we think that concepts derived from latent variables are more likely to be stored than the latent variables themselves, promoting the stability of hippocampal representations. (For example, in humans, language provides a set of relatively persistent concepts, stabilized by the need to communicate.) Projections from the latent variables can classify attributes with only a small amount of training data (see section 'Modelling semantic memory'); we suggest that there could be a two-way mapping between latent variables and concepts, which supports categorization of incoming experience as well as semantic memory. However, for simplicity, the conceptual features were simply a one-to-one copy of latent variable

representations in these simulations. It may also be possible to stabilize the latent variable representations by reducing catastrophic forgetting in the generative network, for example, by using generative as well as hippocampal replay[33,119,120], with the generative network trained on its own self-generated representations in addition to new memories. This builds on previous research suggesting that certain stages of sleep are optimized to preserve remote memories, while others consolidate new ones[121]. This could reduce interference of new learning with remote memories in the generative network, as well as make hippocampal representations in the extended model more stable.

Modelling schema-based distortions in the extended model

Carmichael experiment. We demonstrated the contextual modulation of memory (as in ref. 95) in the extended model by manipulating the conceptual component of an 'event'. To model an external conceptual context being encoded, the original image was stored in the autoassociative network along with activation of a given concept (a cube or a sphere), represented as the latent variables for that class. While in most simulations the latent variables stored in the MHN were simply the output of the VAE's encoder, here an external context activated the conceptual representation, consistent with activity in the EC, mPFC or alTL driven by extrinsic factors.

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

DRM experiment. The DRM task is a classic way to measure gist-based memory distortion[93,94]. Here we demonstrated the rapid onset of semantic intrusions in the extended model, coming about as a consequence of learning the co-occurrence statistics of words in a text dataset representing 'background knowledge'. This followed on from previous work showing that VAEs produce semantic intrusions[32].

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

The generative network was pre-trained on a set of word lists extracted from simple stories[96], representing learning from replayed memories before the DRM stimuli (although replay was not simulated explicitly). Words occurring in <0.05% or >10% of documents were discarded to keep the vocabulary to a manageable size of 4,206 words (this meant that some rarer words in the DRM lists were removed). The word lists were converted to vectors of word counts of length 4,206, in which the value at index i of the vector for a given list indicated the count of word i in the document. As these representations ignore word order, a sequential model was not required (however, this prevented exploring the effect of list position on recall).

Specifically, the variational autoencoder used for this simulation consisted of an input layer followed by a dropout layer[147] projecting to 300 latent variables (sampled from representations of the mean and log variance vectors as usual), and then to an output layer with a sigmoid activation so that predictions were between 0 and 1, with L1 regularization to promote sparsity in this layer. As above, this was implemented using the Keras API for the TensorFlow library[142,148], with the VAE trained to reconstruct input vectors in the usual way.

Following pre-training of the generative network, the system encoded the DRM stimuli, with each of the 20 word lists represented as vectors of word counts. One important detail was the addition of a term, given by 'id_n' for the nth document in the corpus, representing the unique spatiotemporal context of each word list. (Note that this is a highly simplified representation of the spatiotemporal context[116] for illustration.) This enabled recall to be modelled by presenting the network with the 'id_n' term, and seeing which terms were retrieved.

In the extended model, the latent representation of the word list was encoded in the MHN as the conceptual component, while the unique 'id_n' terms were encoded veridically (as vectors of word counts of length 4,226—the original vocabulary size plus the 20 new 'id_n' terms—with 1 at 'id_n' and 0 elsewhere). The sparse vector representing the unexpected 'id_n' term is analogous to the sparse arrays of poorly predicted pixels in the main simulations of the extended model.

When the MHN was given 'id_n' as an input, it retrieved the hippocampal trace consisting of 'id_n' together with the latent representation of the word list. The latent representation was then decoded to produce the outputs shown in Fig. 7a (a dashed line shows the threshold for recall, interpreting the output as a probability so that words with an output >0.5 are recalled). As in the human data, lure words were often but not always recalled. The system also forgot some words and produced additional semantic intrusions, for example, 'vet' in the case of the 'doctor' list.

To test the effect of varying the number of associates, as in ref. 97, subsets of the DRM lists were encoded in the way described above. Specifically, to test the probability of lure recall with n associates studied, n items from each DRM list were encoded. For each list, this was repeated for 20 randomly sampled combinations of n items. Once again, recall was tested by giving the system 'id_n' as an input.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The following datasets (all covered by the Creative Commons Attribution 4.0 License) were used in the simulations:
MNIST[88]: https://www.tensorflow.org/datasets/catalog/mnist
Shapes3D[137]: https://www.tensorflow.org/datasets/catalog/shapes3d
ROCStories[97]: https://cs.rochester.edu/nlp/rocstories

Code availability

Code for all simulations can be found at https://github.com/ellie-as/generative-memory. Some diagrams were created using BioRender.com.

References

1. Tulving, E. How many memory systems are there? *Am. Psychol.* **40**, 385–398 (1985).
2. Marr, D. A theory for cerebral neocortex. *Proc. R. Soc. Lond. B* **176**, 161–234 (1970).
3. Marr, D. Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond. B* **262**, 23–81 (1971).
4. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
5. Teyler, T. J. & DiScenna, P. The hippocampal memory indexing theory. *Behav. Neurosci.* **100**, 147–154 (1986).
6. Bartlett, F. C. *Remembering: A Study In Experimental and Social Psychology* (Cambridge Univ. Press, 1932).
7. Schacter, D. L. Constructive memory: past and future. *Dialogues Clin. Neurosci.* **14**, 7–18 (2012).
8. Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).
9. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).

10. Alvarez, P. & Squire, L. R. Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl Acad. Sci. USA* **91**, 7041–7045 (1994).
11. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
12. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
13. Diba, K. & Buzsáki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242 (2007).
14. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009).
15. Ego-Stengel, V. & Wilson, M. A. Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus* **20**, 1–10 (2010).
16. Winocur, G. & Moscovitch, M. Memory transformation and systems consolidation. *J. Int. Neuropsychol. Soc.* **17**, 766–780 (2011).
17. Norman, Y., Raccah, O., Liu, S., Parvizi, J. & Malach, R. Hippocampal ripples and their coordinated dialogue with the default mode network during recent and remote recollection. *Neuron* **109**, 2767–2780 (2021).
18. Káli, S. & Dayan, P. Hippocampally-dependent consolidation in a hierarchical model of neocortex. *Adv. Neural Inf. Process. Syst.* **13**, 24–30 (2000).
19. Káli, S. & Dayan, P. Replay, repair and consolidation. *Adv. Neural Inf. Process. Syst.* **15**, 19–26 (2002).
20. Becker, S. & Burgess, N. Modelling spatial recall, mental imagery and neglect. *Adv. Neural Inf. Process. Syst.* **13**, 96–102 (2000).
21. Byrne, P., Becker, S. & Burgess, N. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychol. Rev.* **114**, 340–375 (2007).
22. Bicanski, A. & Burgess, N. A neural-level model of spatial memory and imagery. *Elife* **7**, e33752 (2018).
23. Hassabis, D., Kumaran, D., Vann, S. D. & Maguire, E. A. Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl Acad. Sci. USA* **104**, 1726–1731 (2007).
24. Schacter, D. L., Benoit, R. G. & Szpunar, K. K. Episodic future thinking: mechanisms and functions. *Curr. Opin. Behav. Sci.* **17**, 41–50 (2017).
25. Spanó, G. et al. Dreaming with hippocampal damage. *Elife* **9**, e56211 (2020).
26. McCormick, C., Rosenthal, C. R., Miller, T. D. & Maguire, E. A. Mind-wandering in people with hippocampal damage. *J. Neurosci.* **38**, 2745–2754 (2018).
27. Addis, D. R., Wong, A. T. & Schacter, D. L. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia* **45**, 1363–1377 (2007).
28. Hassabis, D. & Maguire, E. A. Deconstructing episodic memory with construction. *Trends Cogn. Sci.* **11**, 299–306 (2007).
29. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://arxiv.org/abs/1503.02531 (2015).
30. Sun, W., Advani, M., Spruston, N., Saxe, A. & Fitzgerald, J. E. Organizing memories for generalization in complementary learning systems. *Nat. Neurosci.* **26**, 1438–1448 (2023).
31. Whittington, J. C. R. et al. The Tolman–Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263 (2020).
32. Nagy, D. G., Török, B. & Orbán, G. Optimal forgetting: semantic compression of episodic memories. *PLoS Comput. Biol.* **16**, e1008367 (2020).
33. van de Ven, G. M., Siegelmann, H. T. & Tolias, A. S. Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **11**, 4069 (2020).
34. Hemmer, P. & Steyvers, M. A Bayesian account of reconstructive memory. *Top. Cogn. Sci.* **1**, 189–202 (2009).
35. Fayyaz, Z. et al. A model of semantic completion in generative episodic memory. *Neural Comput.* **34**, 1841–1870 (2022).
36. Schacter, D. L., Addis, D. R. & Buckner, R. L. Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* **8**, 657–661 (2007).
37. Biderman, N., Bakkour, A. & Shohamy, D. What are memories for? The hippocampus bridges past experience with future decisions. *Trends Cogn. Sci.* **24**, 542–556 (2020).
38. Bein, O., Plotkin, N. A. & Davachi, L. Mnemonic prediction errors promote detailed memories. *Learn. Mem.* **28**, 422–434 (2021).
39. Sherman, B. E. et al. Temporal dynamics of competition between statistical learning and episodic memory in intracranial recordings of human visual cortex. *J. Neurosci.* **42**, 9053–9068 (2022).
40. Barlow, H. B. et al. in *Sensory Communication* (ed. Rosenblith, W. A.) 217–233 (MIT Press, 2013).
41. Barlow, H. B. Unsupervised learning. *Neural Comput.* **1**, 295–311 (1989).
42. Benna, M. K. & Fusi, S. Place cells may simply be memory cells: memory compression leads to spatial tuning and history dependence. *Proc. Natl Acad. Sci. USA* **118**, e2018422118 (2021).
43. Vargha-Khadem, F. et al. Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**, 376–380 (1997).
44. Manns, J. R., Hopkins, R. O. & Squire, L. R. Semantic memory and the human hippocampus. *Neuron* **38**, 127–133 (2003).
45. Squire, L. R., Genzel, L., Wixted, J. T. & Morris, R. G. Memory consolidation. *Cold Spring Harb. Perspect. Biol.* **7**, a021766 (2015).
46. McKenzie, S. & Eichenbaum, H. Consolidation and reconsolidation: two lives of memories? *Neuron* **71**, 224–233 (2011).
47. Durrant, S. J., Taylor, C., Cairney, S. & Lewis, P. A. Sleep-dependent consolidation of statistical learning. *Neuropsychologia* **49**, 1322–1331 (2011).
48. Richards, B. A. et al. Patterns across multiple memories are identified over time. *Nat. Neurosci.* **17**, 981–986 (2014).
49. Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D. & Walker, M. P. Human relational memory requires time and sleep. *Proc. Natl Acad. Sci. USA* **104**, 7723–7728 (2007).
50. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
51. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160049 (2017).
52. Payne, J. D. et al. The role of sleep in false memory formation. *Neurobiol. Learn. Mem.* **92**, 327–334 (2009).
53. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
54. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
55. Mack, M. L., Preston, A. R. & Love, B. C. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* **11**, 46 (2020).

56. Hasselmo, M. E., Wyble, B. P. & Wallenstein, G. V. Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6**, 693–708 (1996).
57. Tse, D. et al. Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
58. Kumaran, D. & Maguire, E. A. An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS Biol.* **4**, e424 (2006).
59. Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H. & Wagner, A. D. Associative retrieval processes in the human medial temporal lobe: hippocampal retrieval success and CA1 mismatch detection. *Learn. Mem.* **18**, 523–528 (2011).
60. Hedayati, S., O'Donnell, R. E. & Wyble, B. A model of working memory for latent representations. *Nat. Hum. Behav.* **6**, 709–719 (2022).
61. Quiroga, R. Q. Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
62. Kolibius, L. D. et al. Hippocampal neurons code individual episodic memories in humans. *Nat. Hum. Behav.* **7**, 1968–1979 (2023).
63. Ramsauer, H. et al. Hopfield networks is all you need. in *International Conference on Learning Representations* (2021).
64. Krotov, D. & Hopfield, J. Large associative memory problem in neurobiology and machine learning. in *International Conference on Learning Representations* (2021).
65. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
66. Kingma, D. P. & Welling, M. An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**, 307–392 (2019).
67. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The Helmholtz machine. *Neural Comput.* **7**, 889–904 (1995).
68. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
69. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
70. Gilboa, A. & Marlatte, H. Neurobiology of schemas and schema-mediated memory. *Trends Cogn. Sci.* **21**, 618–631 (2017).
71. Ghosh, V. E. & Gilboa, A. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* **53**, 104–114 (2014).
72. Chambers, D. & Reisberg, D. Can mental images be ambiguous? *J. Exp. Psychol. Hum. Percept. Perform.* **11**, 317–328 (1985).
73. Moser, E. I., Kropff, E. & Moser, M. B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. of Neurosci.* **31**, 69–89 (2008).
74. Takashima, A. et al. Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proc. Natl Acad. Sci. USA* **103**, 756–761 (2006).
75. Gais, S. et al. Sleep transforms the cerebral trace of declarative memories. *Proc. Natl Acad. Sci. USA* **104**, 18778–18783 (2007).
76. Frankland, P. W. & Bontempi, B. The organization of recent and remote memories. *Nat. Rev. Neurosci.* **6**, 119–130 (2005).
77. van Kesteren, M. T. R., Fernández, G., Norris, D. G. & Hermans, E. J. Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proc. Natl Acad. Sci. USA* **107**, 7550–7555 (2010).
78. Benchenane, K. et al. Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. *Neuron* **66**, 921–936 (2010).
79. Koscik, T. R. & Tranel, D. The human ventromedial prefrontal cortex is critical for transitive inference. *J. Cogn. Neurosci.* **24**, 1191–1204 (2012).
80. Spalding, K. N. et al. Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *J. Neurosci.* **38**, 3767–3775 (2018).
81. Chan, D. et al. Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann. Neurol.* **49**, 433–442 (2001).
82. Bright, P. et al. Retrograde amnesia in patients with hippocampal, medial temporal, temporal lobe, or frontal pathology. *Learn. Mem.* **13**, 545–557 (2006).
83. Ranganath, C. & Ritchey, M. Two cortical systems for memory-guided behaviour. *Nat. Rev. Neurosci.* **13**, 713–726 (2012).
84. Spiers, H. J., Maguire, E. A. & Burgess, N. Hippocampal amnesia. *Neurocase* **7**, 357–382 (2001).
85. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164 (2015).
86. Mokady, R., Hertz, A. H. & Bermano, A. H. ClipCap: CLIP prefix for image captioning. Preprint at https://arxiv.org/abs/2111.09734 (2021).
87. LeCun, Y., Cortes, C. & Burges, C. J. *MNIST Handwritten Digit Database* (AT&T Labs, 2010).
88. Intraub, H. & Richardson, M. Wide-angle memories of close-up scenes. *J. Exp. Psychol. Learn. Mem. Cogn.* **15**, 179–187 (1989).
89. Bainbridge, W. A. & Baker, C. I. Boundaries extend and contract in scene memory depending on image properties. *Curr. Biol.* **30**, 537–543 (2020).
90. Intraub, H. Searching for boundary extension. *Curr. Biol.* **30**, R1463–R1464 (2020).
91. Bainbridge, W. A. & Baker, C. I. Reply to Intraub. *Curr. Biol.* **30**, R1465–R1466 (2020).
92. Park, J., Josephs, E. L. & Konkle, T. Systematic transition from boundary extension to contraction along an object-to-scene continuum. *J. Vis.* https://doi.org/10.1167/jov.21.9.2124 (2021).
93. Deese, J. On the prediction of occurrence of particular verbal intrusions in immediate recall. *J. Exp. Psychol.* **58**, 17–22 (1959).
94. Roediger, H. L. & McDermott, K. B. Creating false memories: remembering words not presented in lists. *J. Exp. Psychol. Lear. Mem. Cogn.* **21**, 803–814 (1995).
95. Carmichael, L., Hogan, H. P. & Walter, A. A. An experimental study of the effect of language on the reproduction of visually perceived form. *J. Exp. Psychol.* **15**, 73–86 (1932).
96. Mostafazadeh, N. et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. in *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Knight, K. et al.) 839–849 (2016).
97. Robinson, K. J. & Roediger, H. L. Associative processes in false recall and false recognition. *Psychol. Sci.* **8**, 231–237 (1997).
98. Cipolotti, L. et al. Long-term retrograde amnesia… the crucial role of the hippocampus. *Neuropsychologia* **39**, 151–172 (2001).
99. Zola-Morgan, S., Squire, L. R. & Amaral, D. G. Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. *J. Neurosci.* **6**, 2950–2967 (1986).
100. Knowlton, B. J., Squire, L. R. & Gluck, M. A. Probabilistic classification learning in amnesia. *Learn. Mem.* **1**, 106–120 (1994).
101. Hodges, J. R. & Graham, K. S. Episodic memory: insights from semantic dementia. *Phil. Trans. R. Soc. Lond. B* **356**, 1423–1434 (2001).
102. Migo, E., Montaldi, D., Norman, K. A., Quamme, J. & Mayes, A. The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Q. J.Exp. Psychol.* **62**, 1198–1215 (2009).
103. Moscovitch, M. & Melo, B. Strategic retrieval and the frontal lobes: evidence from confabulation and amnesia. *Neuropsychologia* **35**, 1017–1034 (1997).

104. Lin, W. J., Horner, A. J. & Burgess, N. Ventromedial prefrontal cortex, adding value to autobiographical memories. *Sci. Rep.* **6**, 28630 (2016).
105. Gluck, M. A. & Myers, C. E. Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* **3**, 491–516 (1993).
106. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. Preprint at https://arxiv.org/abs/1506.06579 (2015).
107. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/abs/2204.06125 (2022).
108. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
109. Ekstrom, A. D. et al. Human hippocampal theta activity during virtual navigation. *Hippocampus* **15**, 881–889 (2005).
110. Eichenbaum, H. Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* **15**, 732–744 (2014).
111. Umbach, G. et al. Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proc. Natl Acad. Sci. USA* **117**, 28463–28474 (2020).
112. Dordek, Y., Soudry, D., Meir, R. & Derdikman, D. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *Elife* **5**, e10094 (2016).
113. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
114. Tsao, A. et al. Integrating time from experience in the lateral entorhinal cortex. *Nature* **561**, 57–62 (2018).
115. Bright, I. M. et al. A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proc. Natl Acad. Sci. USA* **117**, 20274–20283 (2020).
116. Howard, M. W. & Kahana, M. J. A distributed representation of temporal context. *J. Math. Psychol.* **46**, 269–299 (2002).
117. Moscovitch, M., Cabeza, R., Winocur, G. & Nadel, L. Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annu. Review Psychol.* **67**, 105–134 (2016).
118. Strange, B. A., Witter, M. P., Lein, E. S. & Moser, E. I. Functional organization of the hippocampal longitudinal axis. *Nat. Rev. Neurosci.* **15**, 655–669 (2014).
119. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal–neocortical interactions. *Nat. Neurosci.* **7**, 286–294 (2004).
120. van de Ven, G. M. & Tolias, A. S. Generative replay with feedback connections as a general strategy for continual learning. Preprint at https://arxiv.org/abs/1809.10635 (2018).
121. Singh, D., Norman, K. A. & Schapiro, A. C. A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proc. Natl Acad. Sci. USA* **119**, e2123432119 (2022).
122. Millidge, B., Salvatori, T., Song, Y., Lukasiewicz, T. & Bogacz, R. Universal Hopfield networks: a general framework for single-shot associative memory models. in *International Conference on Machine Learning* 15561–15583 (PMLR, 2022).
123. Chaudhry, H. T., Zavatone-Veth, J. A., Krotov, D. & Pehlevan, C. Long sequence Hopfield memory. Preprint at https://arxiv.org/abs/2306.04532 (2023).
124. Tang, M., Barron, H. & Bogacz, R. Sequential memory with temporal predictive coding. *Adv. Neural Inf. Process. Syst.* **27** (2023).
125. Burgess, N. & Hitch, G. J. Memory for serial order: a network model of the phonological loop and its timing. *Psychol. Rev.* **106**, 551–581 (1999).
126. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
127. Bird, C. M. How do we remember events? *Curr. Opin. Behav. Sci.* **32**, 120–125 (2020).
128. Mullally, S. L., Intraub, H. & Maguire, E. A. Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. *Curr. Biol.* **22**, 261–268 (2012).
129. De Luca, F. et al. Boundary extension is attenuated in patients with ventromedial prefrontal cortex damage. *Cortex* **108**, 1–12 (2018).
130. Van Der Kolk, B. A., Burbridge, J. A. & Suzuki, J. The psychobiology of traumatic memory. Clinical implications of neuroimaging studies. *Ann. N. Y. Acad. Sci.* **821**, 99–113 (1997).
131. Bisby, J. A., Burgess, N. & Brewin, C. R. Reduced memory coherence for negative events and its relationship to posttraumatic stress disorder. *Curr. Dir. Psychol. Sci.* **29**, 267–272 (2020).
132. Pellicano, E. & Burr, D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn. Sci.* **16**, 504–510 (2012).
133. Barry, D. N. & Love, B. C. A neural network account of memory replay and knowledge consolidation. *Cereb. Cortex.* **33**, 83–95 (2022).
134. Behrens, T. E. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
135. Nieh, E. H. et al. Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).
136. Burgess, C. & Kim, H. 3D Shapes dataset. *GitHub* https://github.com/google-deepmind/3d-shapes (2018).
137. Hou, X., Shen, L., Sun, K. & Qiu, G. Deep feature consistent variational autoencoder. in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1133–1141 (IEEE, 2017).
138. Stella, F., Baracskay, P., O'Neill, J. & Csicsvari, J. Hippocampal reactivation of random trajectories resembling Brownian diffusion. *Neuron* **102**, 450–461 (2019).
139. González, O. C., Sokolov, Y., Krishnan, G. P., Delanois, J. E. & Bazhenov, M. Can sleep protect memories from catastrophic forgetting? *Elife* **9**, e51005 (2020).
140. Pezzulo, G., Zorzi, M. & Corbetta, M. The secret life of predictive brains: what's spontaneous activity for? *Trends Cogn. Sci.* **25**, 730–743 (2021).
141. Igata, H., Ikegaya, Y. & Sasaki, T. Prioritized experience replays on a hippocampal predictive map for learning. *Proc. Natl Acad. Sci. USA* **118**, e2011266118 (2021).
142. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)* 265–283 (USENIX Assoc., 2016).
143. Whittington, J. C. R. & Bogacz, R. Theories of error back-propagation in the brain. *Trends Cogn. Sci.* **23**, 235–250 (2019).
144. Khattar, D., Goud, J. S., Gupta, M. & Varma, V. Mvae: multimodal variational autoencoder for fake news detection. in *The World Wide Web Conference* 2915–2921 (ACM, 2019).
145. Ramesh, A. et al. Zero-shot text-to-image generation. in *International Conference on Machine Learning* 8821–8831 (PMLR, 2021).
146. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
147. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
148. Chollet, F. et al. *Keras Documentation* (GitHub, 2015).

Acknowledgements
We thank T. Behrens, B. Love and D. Bush for useful discussions, and K. Norman for constructive comments on an earlier version.
Funding support for this work was received from a Wellcome Principal Research Fellowship 'Neural mechanisms of memory and prediction: Finding structure in experience' (222457/Z/21/Z) (N.B.), a Wellcome Collaborative Award 'Organising knowledge for flexible behaviour in the prefrontal-hippocampal circuitry' (214314/Z/18/Z) (N.B.), and an ERC advanced grant NEUROMEM (N.B.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
E.S. and N.B. designed the research and wrote the paper. E.S. performed the computational modelling.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01799-z.

Correspondence and requests for materials should be addressed to Eleanor Spens or Neil Burgess.

Peer review information *Nature Human Behaviour* thanks Gido van de Ven and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

A Supplementary information

A.1 Supplementary results

Figure 1 shows results for the 18 remaining Deese-Roediger-McDermott task word lists not shown in Figure 7. As in the human data, lure words are often but not always recalled when the model is presented with 'id_n'. The model also forgets some words, and produces additional semantic intrusions. See Methods for further details.

Figure 2 shows that latent representations support few-shot learning better than intermediate representations extracted from the encoder or the 'sensory' image features. Decoding accuracy is measured by training a support vector machine to classify the central object's shape, varying the input features and the amount of data, and evaluating the resulting model on a held-out test set. The intermediate features tested are the outputs of four convolutional layers in the encoder, flattened to one-dimensional vectors.

A.2 Further model details

A.2.1 Variational autoencoders

The generative networks used in the model are 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.[2,3] 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 8b for an example from Hou et al.[4]

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.[5] Applying such a layer to the output of a preceding convolutional layer has the effect of finding higher-level features in the stacked feature maps, i.e. if the first convolutional layer learns to identify simple features such as lines at different orientations, the second convolutional layer might learn features consisting of combinations of lines.

A large VAE was used for the Shapes3D dataset (containing RGB images of size 64x64 pixels), and a small VAE was used for the MNIST dataset (containing greyscale images of size 28x28 pixels). In the large model's encoder, four convolutional layers gradually decrease the width and height of the representation and increase the depth (as is standard when using convolutional neural networks to encode images), followed by a pooling layer and dense layers to represent the mean and log variance of the latent representation. In addition, a dropout layer immediately after the input layer is added to improve the denoising abilities of the model.[6] In the decoder, four convolutional layers alternate with up-sampling layers to increase the width and height of the representation and decrease the depth. The smaller VAE used for the MNIST simulations has a latent dimension of 20,

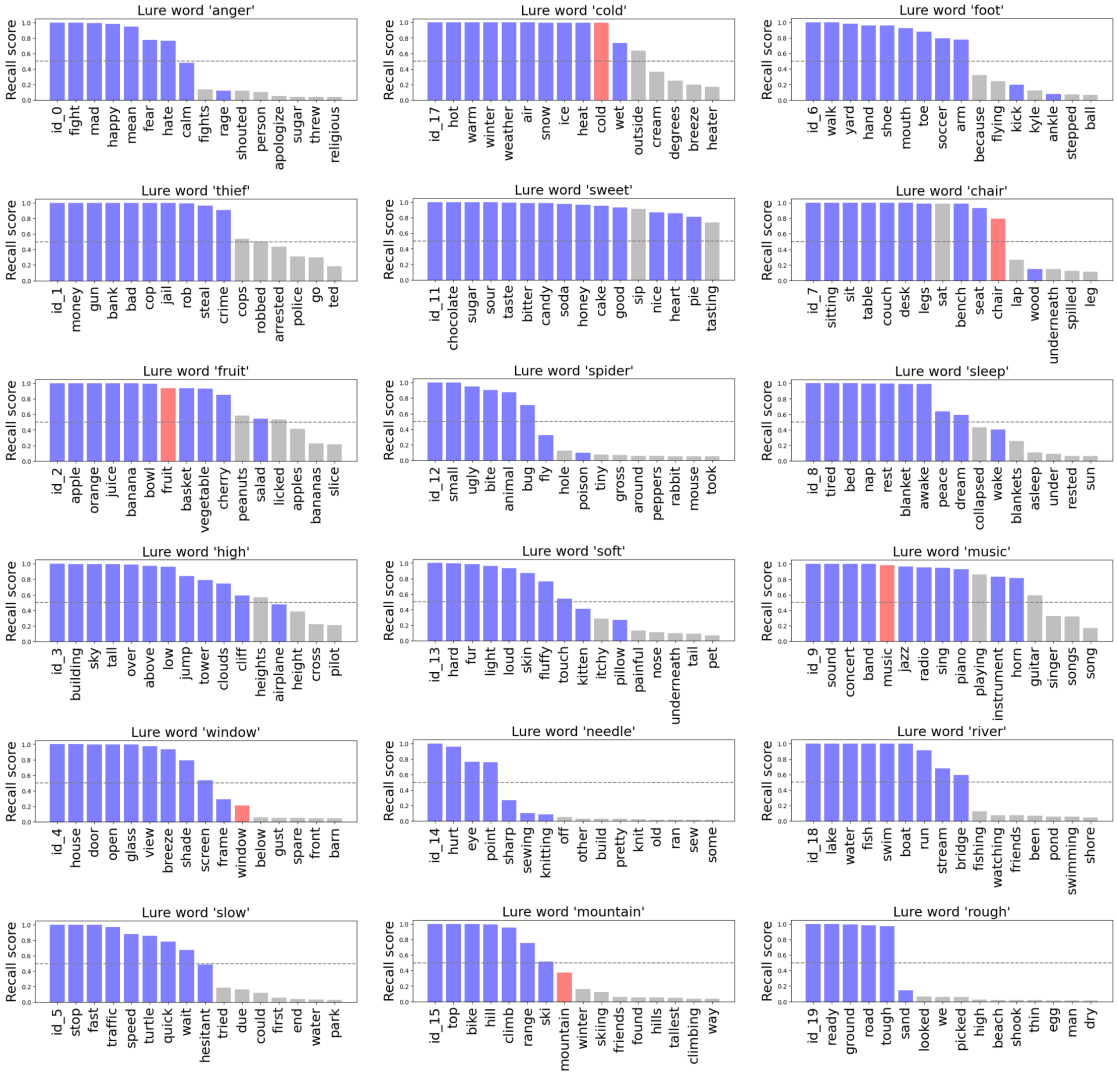

Figure 1: Additional results for the Deese-Roediger-McDermott task. In the extended model, gist-based semantic intrusions arise as a consequence of learning the co-occurrence statistics of words. First the VAE is trained to reconstruct simple stories[1] converted to vectors of word counts, representing background knowledge. The system then encodes the lists as the combination of an 'id_n' term capturing unique spatiotemporal context, and the VAE's latent representation of the word list. In each plot, recalled stimuli when the system is presented with 'id_n' are shown, with output scores treated as probabilities so that words with a score of above 0.5 are recalled. Words from the stimulus list are shown in blue, and lures in red.

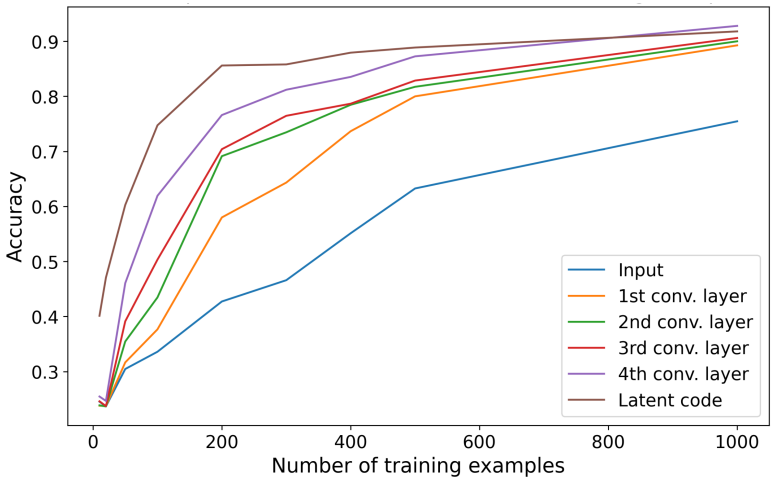

Figure 2: Latent representations support few-shot category learning. The accuracy of an object shape classifier on a held-out test set is shown for different amounts of training data, with different layers of the VAE as input features. The classifier is a simple support vector machine as in Figure 2a.

and a reduced architecture with fewer convolutional layers for efficiency (specifically, there are two convolutional layers in the encoder and two transposed convolutional layers in the decoder).

The following list describes the sequence of operations within the large VAE's encoder network, using the layer names from the TensorFlow Keras API[7] (see also Figure 3):

1. Input layer for arrays of shape (n, 64, 64, 3), representing n 64x64 RGB images
2. Dropout layer with a dropout rate of 0.2 (during training, dropout randomly sets a fraction of the input units to 0 at each step, reducing overfitting and encouraging robustness)
3. Conv2D layer with 32 filters (i.e. convolutional windows, or feature detectors) and kernel size of 4 (i.e. windows of 4x4 pixels)
4. Batch normalisation layer (batch normalisation is a common technique which computes the mean and variance of each feature in a mini-batch and uses them to normalise the activations)
5. LeakyReLU activation layer (LeakyReLU is an activation function that is a variant of the Rectified Linear Unit, ReLU)
6. Conv2D layer with 64 filters and kernel size of 4
7. Batch normalisation layer
8. LeakyReLU activation layer
9. Conv2D layer with 128 filters and kernel size of 4
10. Batch normalisation layer
11. LeakyReLU activation layer

12. Conv2D layer with 256 filters and kernel size of 4
13. Batch normalisation layer
14. LeakyReLU activation layer
15. Global average pooling 2D layer
16. Dense layer to produce the mean of the latent vector
17. Dense layer to produce the log variance of the latent vector (in parallel with the layer above)
18. Custom sampling layer that samples from the latent space, with the mean and log variance layers as inputs

The same information for the decoder network is as follows:

1. Input layer for arrays of shape (n, latent_dimension), where latent_dimension is 20 in these results, representing n latent vectors
2. Dense layer that expands the latent space to a size of 4096
3. Reshape layer to reshape the input to a 4x4x256 tensor
4. Upsampling2D layer with a 2x2 upsampling factor
5. Conv2D layer with 128 filters and kernel size of 3
6. Batch normalisation layer
7. LeakyReLU activation layer
8. Upsampling2D layer with a 2x2 upsampling factor
9. Conv2D layer with 64 filters and kernel size of 3
10. Batch normalisation layer
11. LeakyReLU activation layer
12. Upsampling2D layer with a 2x2 upsampling factor
13. Conv2D layer with 32 filters and kernel size of 3
14. Batch normalisation layer
15. LeakyReLU activation layer
16. Upsampling2D layer with a 2x2 upsampling factor
17. Conv2D layer with 3 filters and kernel size of 3

A.2.2 Modern Hopfield networks

A Hopfield network uses a simple Hebbian learning rule to memorise patterns after a single exposure.[8] 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.[9] It therefore seems unlikely that classical Hopfield networks are a good model of hippocampal memory encoding – even if we assume that

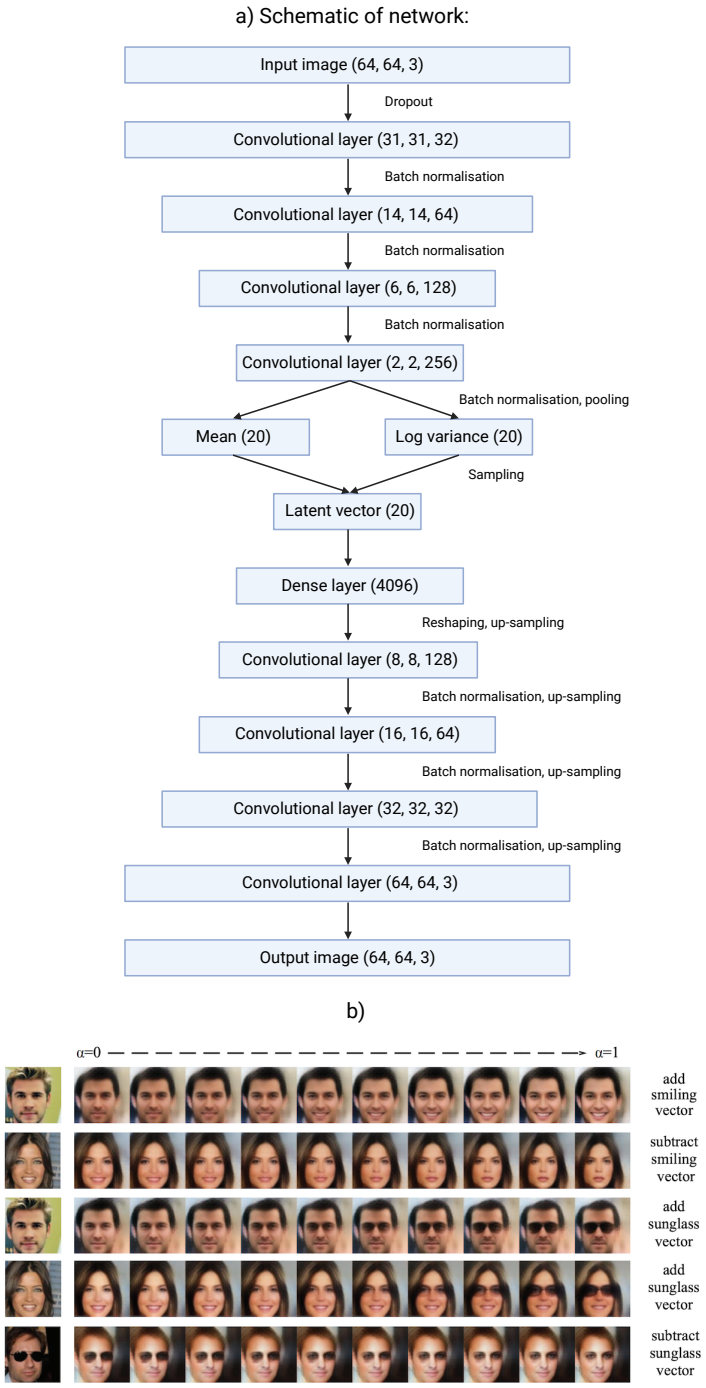

Figure 3: Additional model details. a) Variational autoencoder architecture. Trainable layers (plus the input, output, and sampled latent vector) are shown in boxes, along with the dimensions of their outputs, and non-trainable operations such as activation functions, batch normalisation, and upsampling are shown as annotations. See the SI for more details. b) Figure adapted from Hou et al.[4] with permission, showing the effect of adding and subtracting a proportion α of various different vectors in the latent space of their VAE. (Diagrams were created using BioRender.com.)

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[10] 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. Ref.[11] develops 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.[9] 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) gives the energy of a standard Hopfield Network.[10] 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) gives the energy of a dense Hopfield network.[12] 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[12] 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[12] 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[12] show that under different circumstances, their formulation can be simplified to dense associative memory,[10] or modern Hopfield networks.[9] 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 ref.[12] 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.[13] 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.

References

1. Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., ... Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 839-849.
2. Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
3. Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691.

4. Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. 2017 IEEE winter conference on applications of computer vision (WACV), 1133-1141.

5. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551. MIT Press.

6. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014, JMLR. org.

7. Abadi, M. et al. TensorFlow: A System for Large-Scale Machine Learning in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (2016), 265–283.

8. 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.

9. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., ... Sandve, G. K. (2020). Hopfield networks is all you need. arXiv preprint arXiv:2008.02217.

10. Krotov, D., and Hopfield, J. J. (2016). Dense associative memory for pattern recognition. Advances in neural information processing systems, 29.

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

12. Krotov, D., and Hopfield, J. (2020). Large associative memory problem in neurobiology and machine learning. arXiv preprint arXiv:2008.06996.

13. T. J. Teyler and P. DiScenna, *The hippocampal memory indexing theory*, *Behavioral neuroscience*, **100**(2), 147 (1986), American Psychological Association.