Interplay of episodic and semantic memory arises from adaptive compression

David G. Nagy[1,2], Gergo Orban[3,4,*] & Charley M. Wu[1,2,*]

[1] Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany
[2] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3] CEU Centre for Cognitive Computations, Central European University, Budapest, Hungary
[4] HUN-REN Wigner Research Centre for Physics, Budapest, Hungary
[*] Equal Contributions

Abstract

Sensory experiences are encoded as memories — not as verbatim copies, but through interpretation and transformation. Rate Distortion Theory (RDT) frames this process as lossy compression, aligning with numerous experimental findings. Despite its successes, RDT has a glaring problem: it assumes environmental regularities are known and unchanging, dismissing surprising experiences as noise. However, the brain's model of environmental regularities (semantic memory) is continually learned and refined, with surprising events playing a pivotal role. In this Perspective, we highlight the relevance of this challenge for structure learning and argue that adaptively learned compression fosters characteristic curriculum sensitivity, which has been a recent focus of learning research. We suggest this process provides novel insights into the role of episodic memory in preserving experiences in a relatively raw format for later interpretation. Our Perspective offers a normative framework for the interplay between semantic and episodic memory, encompassing memory distortions, curriculum effects, and prioritised replay.

1. Introduction

Over a century of research has revealed that human memory, far from storing a verbatim copy of sensory experience, is prone to distortions or even the creation of entirely false recollections[1]. Memory can be strikingly inaccurate even for frequently

encountered stimuli such as coins[2], traffic signs[3], corporate logos[4], or icons from popular culture[5]. Rather than being random, many of these memory distortions and biases are systematic[6] and remarkably pervasive. A particularly salient example is the Mandela effect, named after the widespread false memory that Nelson Mandela died in prison during the 1980s, when in fact he was released and later became the President of South Africa. Using a visual analogue (known as the Visual Mandela effect), a recent study demonstrated that a majority of people falsely recognize manipulated versions of visual cultural iconography, such as a monocle-wearing version of the Monopoly Man, even when presented alongside the original[5].

The extent of these inaccuracies might seem surprising and could be perceived as fundamental flaws of human memory. However, they have become widely recognised as reflections of how the primary purpose of memory is not merely the accurate recall of past experience, but rather to support other cognitive functions[6,7], such as prediction, generalisation, decision-making and creativity. For example, many of these errors fall under the category of *gist-based distortions*, where the essential meaning (or "gist") of an experience is retained instead of superficial details[8,9]. This process of gist extraction can be considered to prioritise information most relevant for anticipating future events and guiding behaviour, by interpreting experience in light of prior knowledge and expectations[10–16]. Yet, this leaves the question: what computational principles underlie the way that multiple memory systems (e.g., *semantic* and *episodic memory*) encode past experiences in service of these cognitive goals?

An emerging normative perspective on this question is via compression (Box 1) – specifically, the mathematical framework of Rate Distortion Theory (RDT), which originated in the 1950s as an extension of information theory[17,18]. RDT asks the question of how to optimally encode an input so that it fits within the available capacity budget (the *rate*), while also taking the goals of the system into account. This characterises a fundamental trade-off, whereby reducing *distortion* (measured between the input and the reconstructed memory trace; Fig 1b), results in a corresponding increase in the required *rate* of information (Fig. 1c). An intuitive example is how a streaming video appears degraded when the connection is unstable, with better compression algorithms achieving higher image fidelity for a given connection speed.

A key insight in RDT is that consistent regularities in the environment can be exploited to remove redundant information[17] and thereby use resources more efficiently. This idea has been influential in neuroscience since the 1960's under the rubric of the Efficient Coding Hypothesis[19–21]. However, when resources are insufficient for perfect reconstruction (i.e., lossless compression), RDT allows even further compression by strategically discarding information (i.e., lossy compression) and later trying to reconstruct them based on known regularities. However, this reconstructive process often introduces distortions that make the recalled stimuli more aligned with previously observed regularities.

Applied to human memory, these regularities are a form of pre-existing knowledge typically thought to belong to the domain of *semantic memory*. They can be formalised as an internal *generative model* of the environment, enabling the interpretation and prediction of ongoing experience[14,22–24]. This process of compression using semantic memory introduces distortions in the encoding-decoding process – such as adding a monocle on the Monopoly Man (Fig. 1d). This prediction aligns with gist-based distortions and, more broadly, with early theories of memory distortions, which attribute such errors to the influence of pre-existing knowledge structures known as memory schemas[10]. Although classical compression algorithms produce qualitatively different memory distortions than humans (e.g., blocky compression artefacts in Fig. 1e), recent advances in machine learning, particularly in applying deep generative models[25–27] to compression[28–30] have enabled RDT-based models to capture memory phenomena in complex domains, such as human drawings, text, and natural images[11,12,31]. These findings demonstrate that RDT can serve as a unifying framework for parsimoniously explaining how prior knowledge affects sensory experiences, with characteristic patterns of memory distortions[11–14,32] (Box 1).

In contrast to semantic memory, which retains general knowledge, *episodic memory* is a different representational format that retains traces of specific events and sensory experience in a relatively raw form[33,34]. However, the normative role of episodic memory, specifically its tendency to maintain rich details relative to what is directly relevant to behavioural objectives, has historically been seen as more elusive and the subject of numerous proposals[34–40]. Along this line, recent work applying RDT to memory distortions builds on the distinction between memory systems by arguing that

semantic memory provides the encoding framework for the efficient compression of episodes[11,13,14].

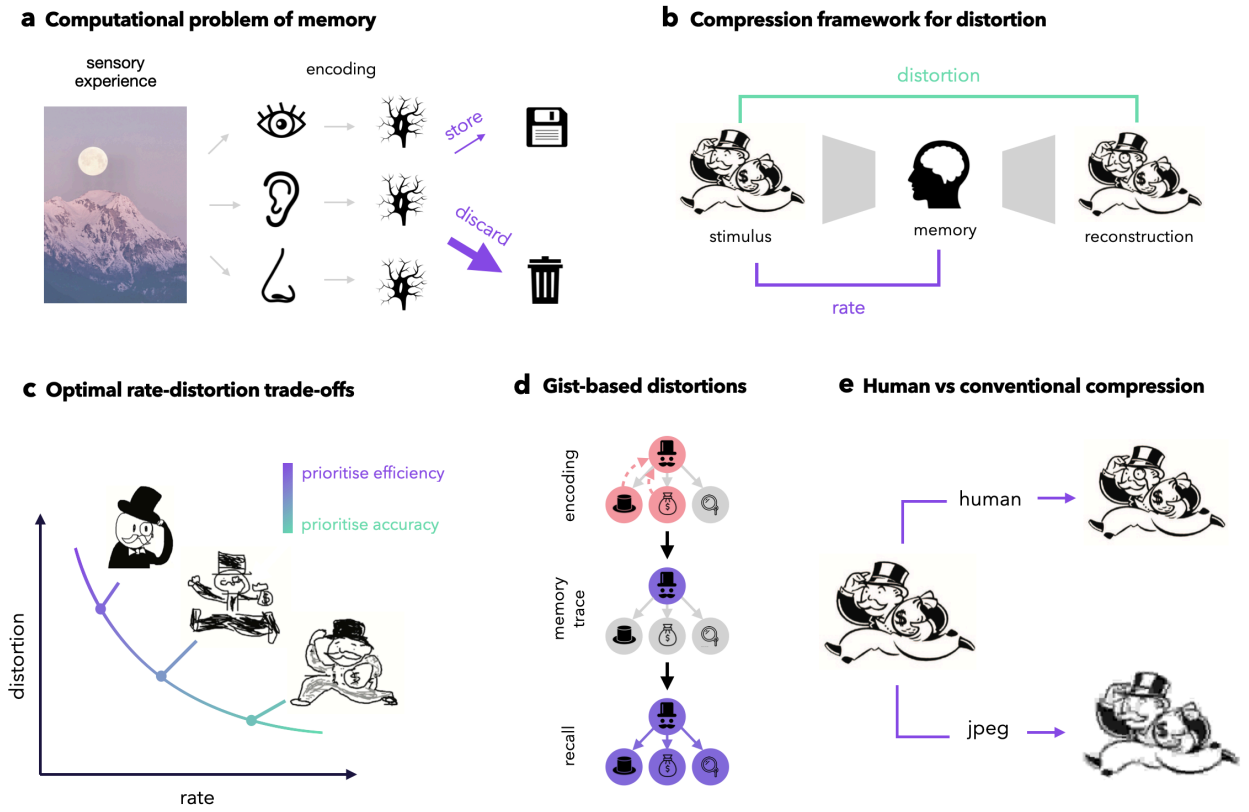


Fig 1. Compression in memory. a. The core computational challenge of memory is determining how to adapt the brain based on incoming experiences. Viewed as a lossy compression process, the critical question is which information to retain and what can be safely discarded to conserve memory resources. **b.** RDT characterises compression algorithms using two key measures: distortion, which quantifies how much the reconstructed experience differs from the original, and rate, which measures the average amount of information preserved in the memory trace. **c.** According to RDT, there is no single optimal encoding strategy, but rather a continuum of trade-offs between rate and distortion. Higher rates allow for more accurate recall, whereas lower rates result in greater distortions. Drawings reprinted from Prasad and Bainbridge (2023)[5]. **d.** A compressed representation relies on knowledge of prior regularities to fill in information missing from the memory trace, leading to gist-based distortions. **e.** Compression artefacts in human memory differ qualitatively from those produced by classical algorithms, such as JPEG.

In this Perspective, we point out that despite its successes in describing how pre-existing knowledge affects the encoding and reconstruction of sensory experience, RDT as a normative framework for human memory (Box 1) suffers from a glaring issue. While providing a unifying account of a wide variety of memory distortions and biases, RDT

neglects a key challenge for memory, namely, the need to learn and update an internal generative model on the basis of continually accumulating experiences. To address this, we begin by proposing an augmented framing of the computational problem of memory as iteratively learned compression and offer a resolution through the combination of semantic and episodic memory systems. Our proposal suggests that the relative richness of episodic memories is due to their role in supporting the online learning of causal structure under resource constraints. We then review literature on curriculum sensitivity in human learning, supporting the view. This allows us to contrast our predictions with Complementary Learning Systems[37,41] (CLS), which provides an alternative account of the interaction between episodic and semantic memory. Next, we turn to the question of what is stored in episodic memory and interpret recent theoretical and empirical investigations into memory prioritisation and experience replay under the light of our framework. Finally, we plot trajectories for future research with a specific focus on how the brain might balance the opposing goals of conserving memory resources and maintaining an ability to learn.

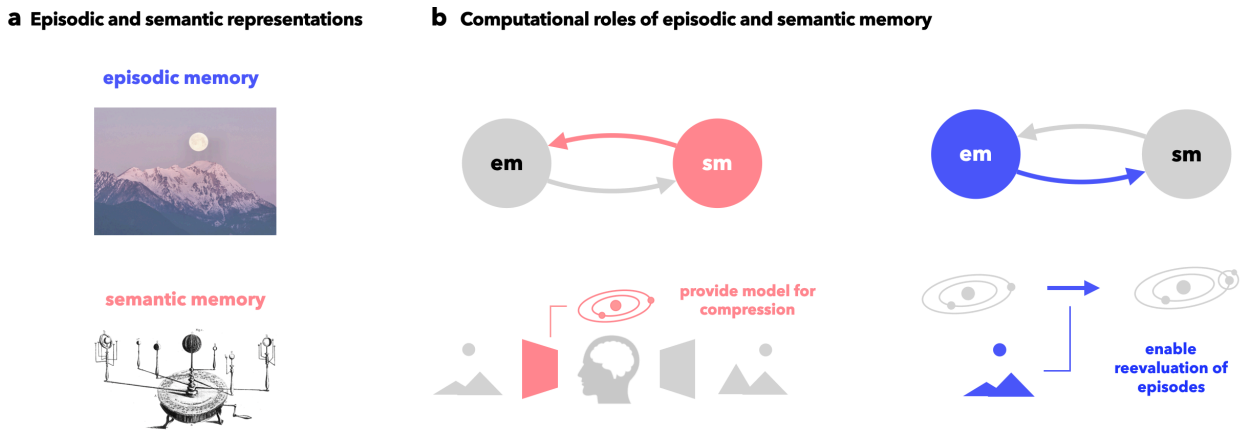


Fig 2. Interactions between semantic and episodic memory. a. We conceptualise episodic memory as retaining traces of individual experiences, and semantic memory as a simplified internal model of the environment, formalised as a probabilistic generative model over experiences. **b.** Our framework offers a perspective on the bidirectional interactions between episodic and semantic memory systems. In these interactions, in line with recent RDT accounts, semantic memory facilitates the efficient compression of episodic memories by providing the statistical model required for compression. However, we argue that the iterative refinement of the model through learning critically relies on encoding surprising and novel episodes in a less compressed format, to preserve them for later re-evaluation, ensuring they are preserved for re-evaluation if the current model proves to be inaccurate or incomplete.

2. The computational problem of memory

While recent research has demonstrated that RDT can serve as a unifying explanation for a wide variety of memory phenomena[11–14,31], there is a fundamental problem: RDT assumes a known and unchanging set of environmental regularities, abstracted into an internal generative model (see Box 1). In contrast, the brain must construct this generative model (semantic memory) over a lifetime, adapting it constantly in light of new experiences; As Barlow himself pointed out as a limitation of his Efficient Coding framework, "what is redundant today was not necessarily redundant yesterday"[20]. This divergence in assumptions also leads to a divergence in predictions: for RDT, since the generative model is assumed to be correct, the only available interpretation for surprising aspects of experience is that they are the result of coincidence or noise, and unlikely to recur. Thus, these surprising aspects are the first to be forgotten when resources are limited. In stark contrast with this prediction, humans tend to recall surprising, novel, and incongruent information with high episodic accuracy[42–46].

We propose that the normative computational problem relevant to human memory is not merely what is considered by RDT, namely, how to efficiently compress experiences under a known generative model. Rather, we need to consider two additional factors: First, the internal generative model is not given but needs to be learned. Second, this learning must proceed in an online, iterative manner, where the model is used for encoding the same experiences that also serve as the basis for updating it. These constraints present a delicate issue for the compression perspective, since during the course of optimising the rate distortion trade-off, an incorrect model discards the very information required for updating it.

To see the inherent challenge in this augmented computational problem, consider the following illustrative example. Imagine learning how to brew good coffee with an unfamiliar machine (e.g., a stovetop moka pot), by figuring out how different variables affect the taste based on trial and error. Each "episode" of brewing a cup (Fig. 3a) involves both relevant (e.g., bean type or grind setting) and irrelevant variables (e.g., the weather or background music). In this situation, we might aim to create a semantic model of coffee brewing, by observing how the relevant input variables affect the taste and capturing this relation in a parametric model. According to normative theories of learning, this can be achieved without specifically remembering any individual episodes.

Rather, all relevant information can be captured by iteratively updating the parameters of the semantic model and then discarding the raw experiences (Fig. 3b top).

Now, imagine that after many brewing episodes, we have figured out a configuration of variables that consistently produces tasty coffee. Yet today, it tastes inexplicably terrible. If we had direct access to all past episodes, we could readily determine the cause: while all features deemed relevant were identical to those during past successes, this time the water added to the pot was too cold, causing the coffee grounds to burn while being heated to a boil. Since water temperature was a factor we previously considered irrelevant, its value in past episodes has been discarded. Thus, we are left surprised, with no clear indication of what went wrong or how to adjust for the next attempt. We might misattribute the failure to coincidence, or worse still, the wrong variable, leading to erroneous parameter updates.

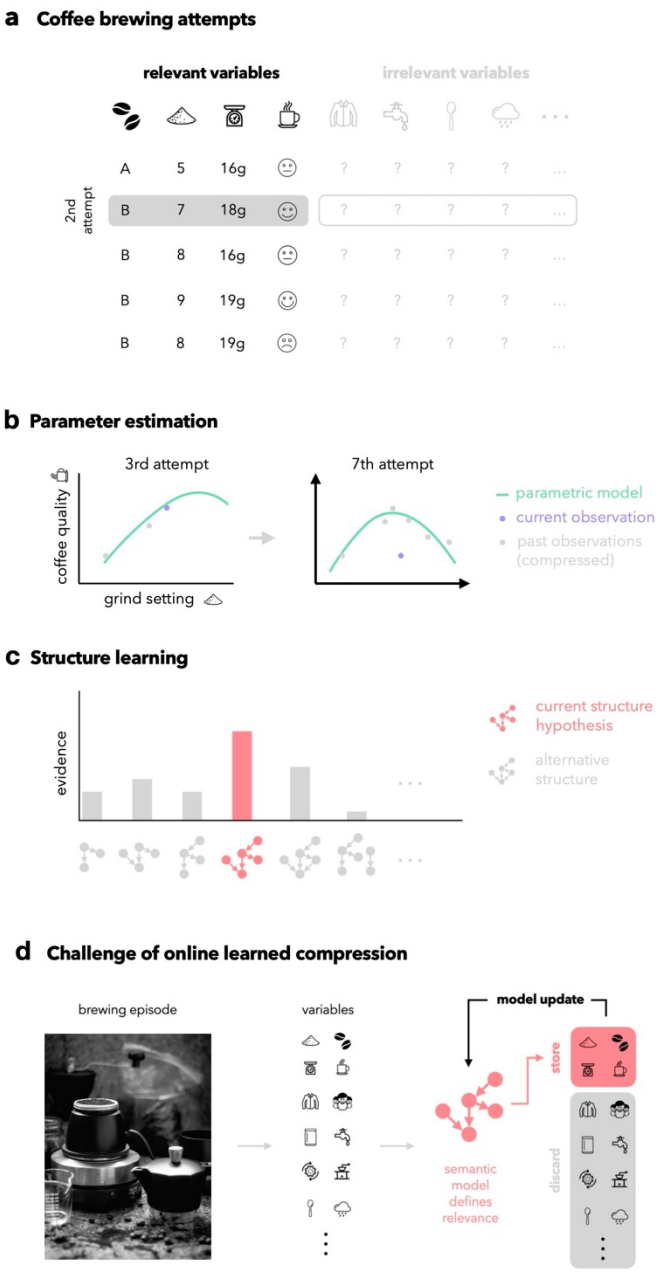


Fig 3. Online structure learning in the coffee brewing example. a. Each episode consists of values for relevant (black) and irrelevant (grey) variables, where for novel situations, initial relevance judgments might be made on the basis of general regularities through generalisation. In online learning, the information content of irrelevant variables is not retained in the parameters, preventing learning about the regularities governed by them. **b.** Iterative learning involves parameter adjustments, such as how grind setting influences coffee quality. Previous data points (grey, shown for reference) are discarded, with relevant information summarised in the model's parameters. For simplicity, quality is shown as a function of grind setting, though in reality other factors are also relevant. **c.** Structure learning entails a discrete learning problem where qualitatively different hypotheses are evaluated based on a measure of congruence with observations. The structure (here understood as probabilistic causal bayesian networks[47] but see Box 2 for alternatives) determines the set of relevant variables and their connections. **d.** Episodes contain numerous variables, with the semantic model determining which of these are deemed relevant. During model updates, only the information from relevant variables is retained, while irrelevant information is discarded. However, this process carries the risk of misclassifying a relevant variable (such as water temperature) as irrelevant, potentially leading to the systematic loss of information essential for future model updates.

A key property behind the failure of the learning process outlined above is that beyond *parameter estimation* (i.e., refining a known parametric relationship between known variables), it also features an additional problem of *structure learning*[48,49] (Box 2). Structure learning involves, for example, identifying causal variables in a given

environment (e.g., weather, clothing, grind setting), as well as determining how they affect each other (e.g., weather may directly affect mood, but not the grind setting). In terms of compression, a known model structure allows for highly efficient use of memory resources by only encoding information relevant to the parameters. While this means that parameter estimation may often be accomplished in an online way, for structure learning, online updates require tracking and updating each possible hypothesis in parallel. Unfortunately, this quickly becomes impractical, as even in the case of a toy problem with just four variables, there are 543 possible hypotheses about the causal structure, with this number growing to over 29,000 with a single additional variable. Such a combinatorial explosion of the hypothesis space is typical for structure learning problems, and due to this proliferation of hypotheses, maintaining the relevant information for each candidate structure is as challenging as storing all past episodes directly.

Therefore, we face a conundrum: On the one hand, limited memory resources require us to store experiences in a compressed format, which is supported by a learned semantic model of the environment. On the other hand, learning and maintaining a semantic model requires access to details of previous episodes that may not have been considered relevant under the current model. How, then, does the brain thread the needle between a combinatorial explosion of hypotheses and the risk of discarding key information?

We propose that the brain uses an approximation relying on the combination of two interlinked memory systems (Fig. 2). *Semantic memory* builds a model of environmental regularities to facilitate compression and due to computational and memory constraints, tracks only a restricted set of hypotheses over structure (perhaps only the single most likely hypothesis[50–53]). However, restricting the set of tracked hypotheses risks being stuck in a dead-end, where information that is necessary for further improvement of the model has already been selectively discarded (as in our coffee brewing example). Therefore, *episodic memory* retains a relatively raw and uncompressed encoding of novel and surprising episodes (i.e., those most likely to be misinterpreted under the current hypothesis), offering some insurance against inherent failure modes of online structure learning.

In the following, we propose a new integration of semantic and episodic memory that solves the dual theoretical problems of *learning to remember* (building a semantic compression model) and *remembering to learn* (storing relevant episodes for future model updates), while explaining a range of different empirical findings. Section 3 describes the process of building the semantic model, drawing on the analogy of Neurath's ship as a metaphor for structure learning under bounded rationality[50]. Here, the need for approximate inference implies a distinct sensitivity to the order in which stimuli are encountered (i.e., curriculum effects). In Section 4, we show how episodic memory can effectively complement semantic memory. By preserving surprising events in a high-fidelity format, episodic memory serves as a "life-raft" enabling future updates to the semantic model by retaining seemingly irrelevant details. We then review two families of replay algorithms proposed in prior research (prioritised replay and generative replay) in light of our framework, arguing that our proposal suggests a combination that draws on the benefits of both.

3. Learning to remember

In earlier sections, we argued that the efficient allocation of memory resources necessitates the construction and continual updating of a generative model of the environment (i.e., semantic memory) based on observations. Specifically, we have highlighted the challenge of identifying the correct causal structure, given the combinatorial explosion of the space of hypotheses capturing potential relationships between relevant variables. Although causal learning is a specific instance of the broader domain of structure learning that our argument applies to (see Box 2), both face a common challenge: vast and difficult-to-navigate hypothesis spaces, necessitating the use of approximations.

One of the most common methods for approximate structure learning is to use Monte Carlo sampling, tracking a selected set of hypotheses instead of the full distribution. Converging evidence from multiple learning paradigms suggests that the brain may also be limited to tracking a restricted set or even a single structural hypothesis[50–53]. In the context of learning global causal structure, this reflects the intuition that it is challenging to maintain parallel interpretations of experience under vastly different hypotheses. A

compelling proposal[50] has likened this learning process to the metaphor of Neurath's ship, originally introduced in the philosophy of science[54,55] to illustrate the gradual and continuous development of scientific theories:

"We [theorists] are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction."

Applied to the brain, the ship represents an individual's evolving understanding of the structure of the world in their semantic memory: a dynamic and evolving hypothesis that informs perception, decision-making, and — as we propose — memory. Neurath's metaphor underscores the locality of the changes made to the ship, reflecting the idea that updates to the brain's model of the world are not wholesale replacements but rather incremental modifications. Just as sailors replace individual planks or beams while keeping the rest of the ship seaworthy, the brain updates its hypotheses by adding, removing, or adjusting elements to make local changes to the current model without compromising its ability to function effectively within its environment (Fig 4a; for details, see Box 2).

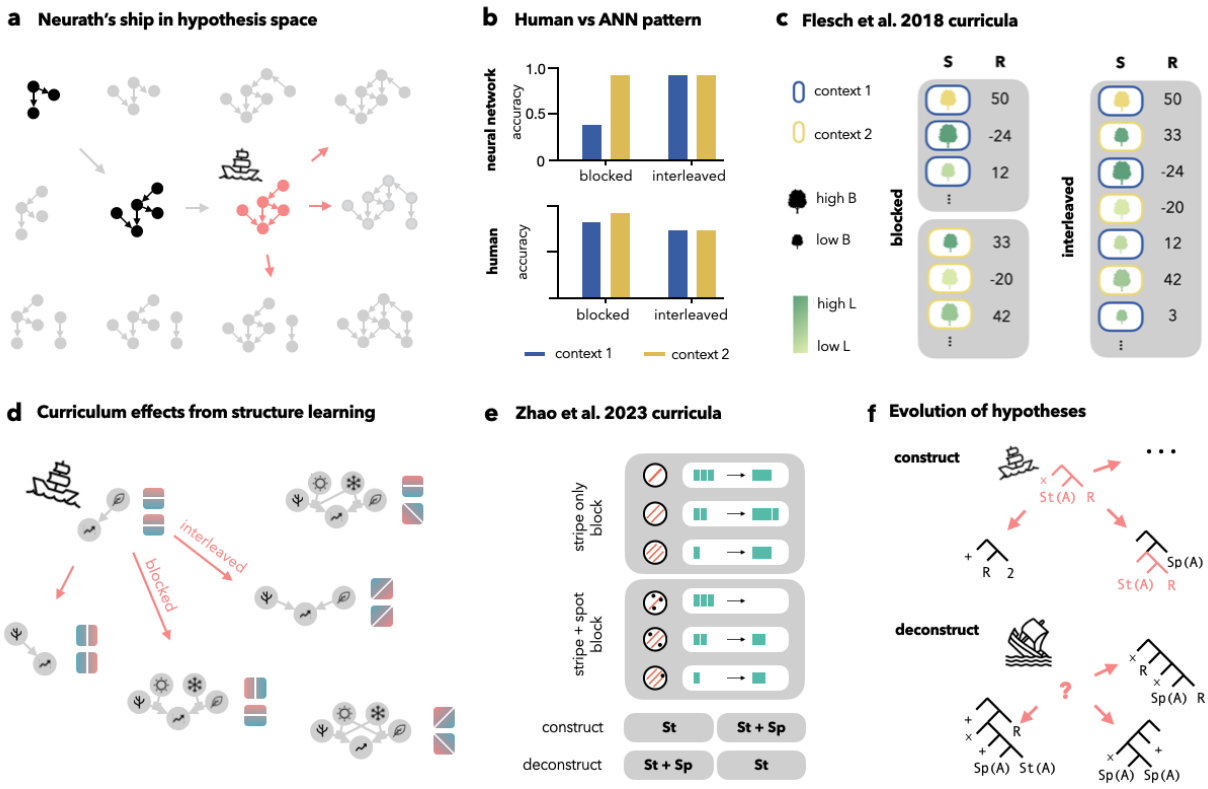


Fig 4. Neurath's ship analogy for human learning. **a.** Neurath's ship represents the evolution of structural hypotheses during learning. **b.** Comparison of curriculum dependence in ANNs versus human learners. *Blue* and *yellow* bars indicate the first and second tasks, respectively. ANNs suffer from catastrophic interference in the blocked setting, while humans do not. However, humans tend to show better performance in blocked settings. **C.** Schematic of blocked versus interleaved curricula in a decision-making task from Flesch et al. (2018). Each observation presents a tree characterised by two features—leafiness and branchiness (depicted as colour and size, respectively)—and a context indicated by the background (represented by outline colour). Participants learn through trial and error which plants thrive (i.e. are rewarded) in specific contexts. **d.** An example of possible hypothesis evolution in the Flesch et al. task. Nodes depict the relevant variables given a particular hypothesis. Coloured plates show the hypothesised dependence of reward magnitude on the strength of one or the other feature (horizontal and vertical axes). Interleaved training prevents the discovery of the regularities governing the two distinct contexts. **e.** Stylised depiction of the data used in the Zhao et al. task[56]. The object in the left column (referred to as a "magic egg" in the experiment) possesses two features—number of stripes (*St*) and number of spots (*Sp*)—which determine its effect on the stack of rectangles (*Re*) upon collision. One group of participants is presented with the top block first, followed by the bottom block, while the other group sees the reverse order. **f.** Illustration of hypothesis evolution in the Zhao et al. (2023) task. The top panel depicts a learner in the "construct" curriculum, who has successfully identified the primitive structure (*St* x *Re*) in Block I, and can therefore use this primitive to constrain the search problem in Block II. In contrast, the learner in the bottom panel, having encountered a more challenging structure inference (*St* x *Re* – *Sp*) in Block I, and having failed to identify a useful structure for retaining the relevant information from Block I, by Block II the discovery of the same primitive comes too late.

According to our compression perspective, the ship represents a single structural hypothesis in semantic memory, determining how new experiences are interpreted and therefore what information is retained. In an experimental setting, the subject forms this structural hypothesis on the basis of earlier observations. If they succeed in discovering the correct structure, then further congruent observations can be integrated quickly and efficiently[46,57,58], with semantic memory determining which aspects of the observations are safe to discard. However, an incorrect structural hypothesis can lead to two key failure modes in learning: First, it may compromise the interpretation of future observations, leading to erroneous parameter updates. Second, even if the subject eventually realises that their hypothesis is flawed, alternatives are evaluated on the basis of past data, which was compressed on the basis of an incorrect hypothesis. As a result, supporting evidence for the correct structure may have been mistaken for noise and systematically discarded, leaving the learner stranded in a dead-end hypothesis. This entanglement of learning and compression in the Neurath's ship approximation leads to learning dynamics with specific patterns of order dependence[39,59,60], which have been observed in human data, but diverge from learning dynamics characteristic of alternative accounts of semantic learning using Artificial Neural Networks (ANNs), such as CLS theory.

CLS models the acquisition of the semantic model as the gradual updating of an ANN, integrating information across multiple experiences over time. Similar to human learning, ANNs also display robust curriculum effects, but often in an opposite pattern to what humans tend to exhibit. In a study by Flesch and colleagues[61], both humans and artificial neural networks were given the same context-dependent decision-making task (Fig. 4b) in either blocked or interleaved curriculum. When ANNs were presented with different tasks or learning contexts in a *blocked* manner, the different blocks tended to overwrite one another, in a well-studied phenomenon known as "catastrophic forgetting"[62,63]. But when the training data was *interleaved*, with a shuffled ordering of the same data, ANNs could learn reliably[64]. CLS theory[37,41], one of the most influential proposals for why an episodic memory system is required, was concerned with exactly the challenge of mitigating catastrophic forgetting. The utility of episodic memory in CLS is that interleaving older episodes with current observations protects older knowledge from being overwritten. In contrast to ANNs, humans performed better in blocked settings, but were hindered by interleaved curricula, with a stronger effect as the

complexity of the task was increased (i.e., by adding more features for each context). Other empirical studies have also found similar effects of blocked curricula leading to better performance for humans in tasks with structural uncertainty[65–67]. Note that some studies have also found benefits to interleaved curricula in different settings, especially when generalisation between stimuli was beneficial to the performance of the task[68], or in discrimination tasks where the immediate juxtaposition of exemplars from different classes seems to highlight the differences between them[69–71].

From the perspective of online structure learning, blocked data is ideal, since consecutive trials from a single context allow the learner to focus on a subset of the complete structure[39,60,67], creating a more manageable hypothesis space to be searched. Once the structure has been discovered, semantic memory can be used to efficiently compress further observations from the same context, allowing the learner to fine-tune the parameters, similar to our coffee example (Fig. 3b). In the case of interleaved training, the initial hypothesis space to be considered is much larger, making it difficult to form an initial hypothesis. Furthermore, without the interpretative structure provided by an effective hypothesis, the useful information in the observations cannot be selectively retained, preventing the accumulation of evidence for the correct structure. One possible outcome of this is that certain subjects might fail to retain the context accurately or arrive at an overly simplified structure that merges the contexts together. A key distinction between CLS and our framework is that in CLS, semantic memory relies on interleaved training, with episodic memory mitigating the adverse effects of blocking. In contrast, our approach suggests a benefit to blocked training for semantic memory, while episodic memory is essential for counteracting failure modes caused by interleaved training.

A more direct connection between curriculum dependence in human learning and the problem of structure discovery was established in a recent study by Zhao and colleagues[56]. The study focuses exclusively on blocked curricula, where the content of the blocks themselves are manipulated (Fig. 4c), as participants learn a causal relationship between the features of a dragon egg (stripes and spots) and the length of a magic wand. Under the “construct” curriculum, the first block only contains examples where one feature is present (either stripes or spots), while the second block introduces the second feature (stripes + spots). In contrast, the “deconstruct” curriculum reverses the order, presenting a more challenging structure inference problem in the first block,

where both features are varied simultaneously. Subjects in the experiment were allowed to revisit previous examples within the same block, mitigating the demands on their memory. The study found that more participants discovered the correct structure – here conceptualised as a program[72,73] (see Box 2) – under the "construct" than the "deconstruct" curriculum. This is because in the "construct" curriculum, a correct *partial* rule (i.e., only incorporating the first feature) could be easily inferred from the simple Block I, and then extended to also incorporate the second feature in Block II. In contrast, the reversed order in the "deconstruct" curriculum made the initial hypothesis space much more complex in Block I, and even the comparatively simple stimuli in Block II did not guide them to the correct solution. Even though the simpler Block II allowed a significant proportion to identify the correct structural primitive, they were unable to retrospectively apply this knowledge to the observations from the first block, consistent with the hypothesis that they were unable to effectively compress its information content due to the lack of a suitable interpretative structure.

In summary, we argued that efficient compression requires the brain to iteratively construct a generative model of the environment in semantic memory, while simultaneously applying the same model to compress observations. We proposed that the brain relies on an approximate solution to this problem of online structure learning that can be likened to Neurath's ship. This approximation leads to a characteristic path-dependence in learning, where the ability to compress is critically reliant on the success of structure discovery. The resulting curriculum effects align with empirical data, but contrast with the dynamics observed in ANNs, which are commonly used as models of human learning. A key insight of our framework is that online structure learning via approximate inference implies a trade-off between efficient compression and robust structure learning: efficient compression involves using semantic memory to discard irrelevant information, whereas learning the underlying structure requires holding onto seemingly irrelevant aspects of experience in episodic memory in order to evaluate alternative hypotheses. Next, we focus on this latter component of our proposed framework, exploring how episodic memory may support the acquisition of semantic knowledge.

4. Remembering to learn

Thus far, we have focused on how people learn to remember—how we build a model of the environment in semantic memory that allows us to selectively retain information necessary for further adaptation of our predictive model. In this section we explore the use of *remembering* as opposed to simply *knowing*, that is, an ability to create rich reconstructions of prior experience including haphazard details. We have already pointed out that when the current structural hypothesis used to interpret and compress incoming experiences is flawed, this can result in erroneous model updates and an inability to evaluate alternative models. Thus, our framework is based on the insight that the only generally applicable way to mitigate these failure modes of semantic learning is to retain information that might seem irrelevant in the context of the current hypothesis, but relevant for evaluating potential alternatives. Therefore, the ability to remember is crucial to ensure that future learning remains possible.

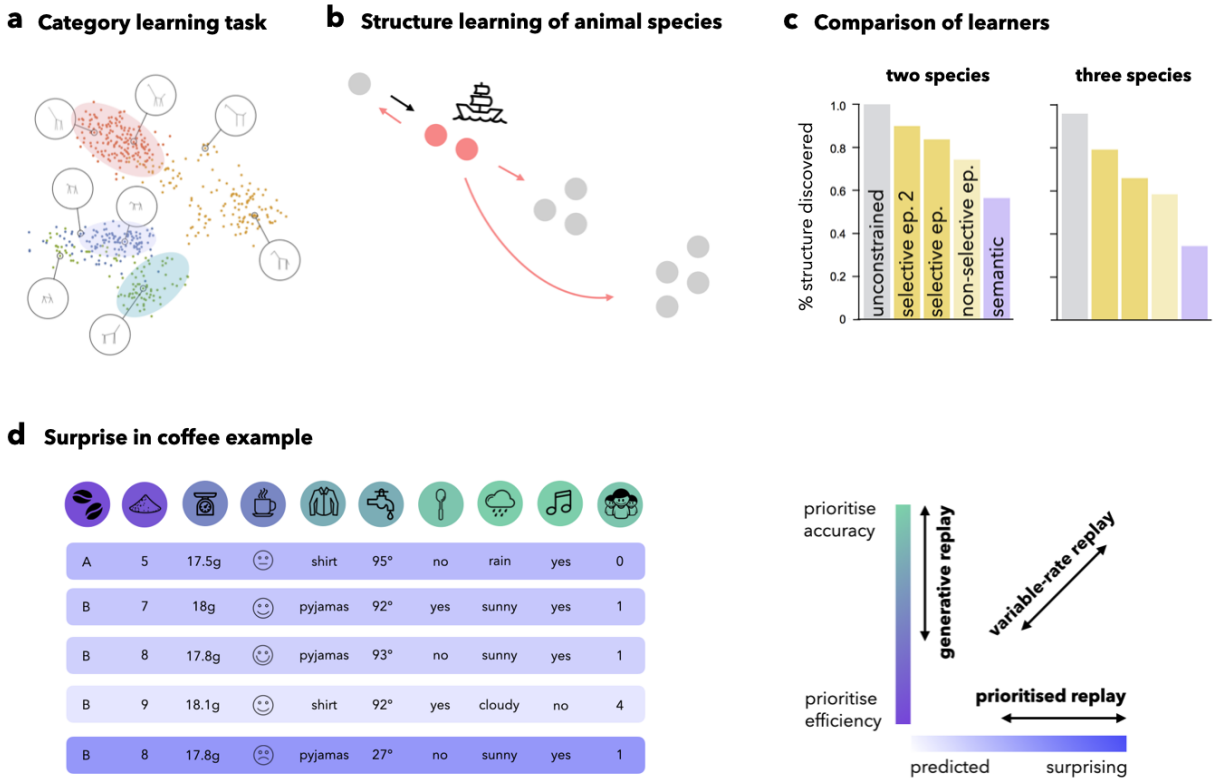


Fig 5. The episodic "life raft". a. In a category learning task classes disjunct classes are inferred from data, illustrated by a 2D embedding of stick figure animals[74]. The segment lengths and angles cluster around species but individual species are characterised by considerable variance. **b.** Structure learning

entails the discovery of the number and property of classes. While multiple hypotheses about the number of classes can be compatible with the data, the inventory of possible hypotheses needs to be navigated eventually based on a single hypothesis at a time. **c.** An unconstrained learner tracks all possible hypotheses, while the semantic learner only relies on a single hypothesis, learning about the current classes but ignoring alternatives. The episodic learner integrates multiple experiences by selectively or non-selectively retaining raw past experiences, prioritising experiences by the degree of surprise under the current hypothesis. **d.** Variable-rate encoding. Increasing the resources dedicated to encoding surprising experiences makes recall more accurate for these episodes (e.g., retaining more variables). From the perspective of replay, using surprise to define the rate of encoding couples the priority of an episode to the level of detail generated by the model. This mechanism combines the benefits of prioritised and generative replay by retaining informative episodes with higher fidelity while conserving resources where they are unlikely to hinder further learning (see Box 3).

In our coffee making example (Fig. 3), the omission of water temperature from the set of relevant variables meant that it was unclear how to update the model after an unexpectedly poor outcome. If the unpleasant episode were encoded in episodic memory, it might include details that are irrelevant in the context of the current hypothesis, such as the water being drawn from a cold tap. Although the significance of these details may not be immediately apparent, a subsequent episode—where the water comes from a preheated kettle and produces a much better tasting result—could retrospectively reveal water temperature as a relevant contextual variable. Thus, extending the Neurath's ship metaphor with an "episodic life-raft" (Fig. 5a), allows a learner to make greater overhauls to model structure by preserving experiences in a detail-rich and relatively unrefined format.

A study by Nagy and Orban[39] provides an intuitive demonstration of the value of an "episodic life-raft" in a simple category learning task. In this study, learning agents had to iteratively learn categories through sequential observations, akin to categorising unknown animals into species based solely on their observed features (Fig. 5a). Here, *structure learning* simply requires the learner to determine the number of categories (e.g. species; Fig. 5b), whereas *parameter estimation* requires refining the feature distributions for each category. Consistent with our current proposal, the study found that a "semantic-only" learner often failed at this structure learning task, by systematically underestimating the number of categories unless the observations were carefully ordered (i.e., blocked curricula). However, endowing the learner with episodic memory, even with severely limited capacity, greatly improved successful structure discovery (Fig. 5c). Since this episodic memory stored a small subset of past experiences,

they could be *replayed* when considering alternative structures, thus augmenting the information contained in the parameters of the current hypothesis. This process can be viewed as an analogue to replay for memory consolidation[75], where episodic memories are integrated with the knowledge maintained in semantic memory.

While an episodic life-raft is clearly advantageous for structure learning, episodic memory is a costly memory format, with this cost stemming from the very source of its utility. This raises the question of which experiences are most beneficial to allocating scarce resources to, or rather, which episodes should be prioritised? First, we consider a simplified context where the only limitation is on the number of episodes that can be stored, but assume each of these episodes can be recalled perfectly. Then we extend this idea to more realistic scenarios where episodes themselves are stored in compressed form. We have argued that, to avoid losing relevant information, the most critical experiences to capture accurately are those that seem incongruent with the current model. Either because they are surprising, or because they relate to a new, previously unexplored area or aspect of the environment, such incongruent experiences are expected to be the most useful for the goal of structure learning.

Consistent with this reasoning, simulated learners with limited episodic memory capacity that selectively prioritised experiences with high Bayesian surprise[76–78] performed better in online structure learning[39] (Fig. 5c). This approach is similar to earlier proposals in category learning that learned exceptions to general rules, sharing the need to store anomalous events[79,80]. An advantage of Bayesian surprise is that it distinguishes between surprise due to mere noise and surprise that warrants model change[76,81], although alternative formalizations of surprise or novelty could be explored in the future[82]. Surprise can also signal environmental change, and underlie the detection of new types of events, thus contributing to event segmentation[83,84].

The idea that incongruent and novel information is selectively prioritised in memory has a long history in psychology[45] and is supported by extensive empirical findings[42–44,46,85]. Similarly, in neuroscience, the idea that the hippocampal formation (associated with episodic memory) serves to retain novel information has been extensively explored[46], particularly in the context of experience replay and memory consolidation[38,41,86,87]. The normative question of how episodes should be prioritised is typically referred to as *prioritised replay*[88,41,87]. Since constraints on memory resources are not a primary

concern in these approaches, prioritisation refers not to which episodes should be retained, but to how frequently they should be replayed. Note that in the limit, lowering the probability of replaying an episode corresponds to discarding it. Prioritising episodes based on their associated reward prediction error has been instrumental in recent machine learning advancements[88–90]. In reinforcement learning, it has also been argued that the utility of retaining episodes is greatest in the early stages of encountering a novel environment, before a sufficiently accurate semantic model can be established[35,91,92].

After considering the simplified normative problem where selected episodes could be recalled perfectly (i.e. *exact replay*), we now turn to more realistic scenarios, where memory resources are more limited. The RDT perspective suggests that memory resources can be decreased if episodes are compressed lossily, with missing details filled in by a generative model maintained in semantic memory (Box 1). In the replay literature, the solution of replaying what are essentially compressed episodes during the training of ANNs, called *generative replay*, has been shown to place significantly lower demands on memory resources compared to exact replay while still protecting against catastrophic forgetting[93,94].

Although compressing episodes using semantic memory enables significant savings in memory resources[94], it appears to be directly at odds with our proposed role for episodic memory. If episodes are stored to preserve seemingly irrelevant details for later reinterpretation in case the current model is incorrect, how can the same model be relied upon to compress these experiences? A key insight of RDT, discussed in Section 1, may be crucial in resolving this tension: episodes can be compressed at varying levels of detail, reflecting different trade-offs between allocated resources (rate) and distortion (Fig. 1c). Rather than a binary choice between storing an exact copy of an event in episodic memory or merely updating the model's parameters, RDT allows for a continuum of choices regarding the desired fidelity of the reconstruction.

We propose that this *variable-rate encoding* of sensory experience underlies the brain's ability to balance the competing goals of conserving resources and maintaining robustness to novelty (Box 3). Specifically, we suggest that the rate of encoding — or equivalently, the desired accuracy of recall — should be determined by a measure of surprise or novelty associated with each observation (Fig. 5d). Under this approach,

most experiences would leave a trace in episodic memory, but well-predicted episodes would be stored in a highly compressed form relative to surprising ones. This makes them prone to increasing degrees of gist-like distortions as a result of increasing levels of compression, as demonstrated in previous work on the RDT perspective on human memory[11,12]. However, when semantic memory has lower confidence in its interpretation, due either to violated predictions or novel situations, additional memory resources could be allocated to encode the episode in greater detail. This mechanism could prioritise information in episodes in a graded fashion, and explain why surprising experiences are often recalled with more episodic detail[42–44,46,77,85]. Altogether, our proposal provides a computational account of memory distortions through the interaction between episodic and semantic memory. However, the impact of variable-rate compression on episodic memory's ability to prevent structure-learning failures remains theoretically unexplored and awaits empirical validation.

5. Conclusion

We argued that RDT, in its current form, faces a fundamental challenge when applied as a normative framework for human memory. While RDT-based approaches describe how episodes might be compressed using a semantic model, they overlook how semantic knowledge is acquired in the first place—from the same experiences that the model interprets and compresses. We highlighted how this omission results in qualitative discrepancies from the empirical phenomena of human memory and argued that addressing them requires a rethinking of the fundamental assumptions of RDT.

Thus, we proposed a revised normative framework, where *semantic memory* tracks a limited approximation of the environmental structure, based on the analogy of Neurath's ship. Since interpreting observations under a single structural hypothesis can result in systematic loss of essential information, we argued that Neurath's ship also requires an *episodic life-raft*, recruiting additional memory resources to encode novel and surprising observations in a relatively uncompressed format. By accounting for the role of episodic memory in safeguarding against learning the wrong generative model, we arrive at a normative explanation for why surprising stimuli are often remembered

with higher fidelity. However, for experiences that are congruent with the current structural hypothesis, our framework produces typical RDT-like distortions. Ultimately, our perspective on the interplay between episodic and semantic memory systems offers a parsimonious explanation for a wide range of phenomena in human learning and memory, while also providing new insights into several ongoing challenges in the field.

A key focus of our Perspective are the consequences that an evolving compression model (i.e., semantic memory) has for memory distortions. The standard RDT approach has served as a unifying framework for classical gist-based memory distortions, such as intrusions of semantically related items[95] or label-consistent distortions in memory[96] for sketches[11,14]. However, if we allow the compression model to evolve over time, updates driven by new observations will influence the encoding of subsequent ones – indeed the defining property of curriculum effects. Conversely, updating the model after a new experience (such as in post-event misinformation[97] or hindsight bias[98]) changes the decoder, and thus alters how we reconstruct past experiences[23], potentially also including associative memory errors[99].

In the study of both human and machine learning, stimuli are typically presented in randomised fashion, with simple or non-existent dependencies between trials. This has clear benefits for eliminating experimental confounds. However, it stands in stark contrast with the rich, multi-scale sequential structure that characterises natural environments. In this Perspective, we have focused on coarse-grained structure, neglecting the temporal breadth of episodes, and consequently the issue of how to segment continuous sensory inputs[83,84,100–102]. However, this fine-grained temporal structure and its interactions with structure learning are likely to be important in a more nuanced understanding of curriculum effects[103], and an integration of these approaches may explain a larger variety of curriculum effects[69–71] under a unified framework. A more refined understanding of path-dependencies in learning, also crucial for educational applications, will require improved theories about how semantic knowledge is represented and organised. One intriguing approach is to view semantic memory as a library of concepts, often formalised in a program induction framework[104–106], where a goal of curriculum design is to induce widely applicable and composable conceptual modules that further learning can build on[56,60,107].

Although absent from existing RDT accounts, emotional salience is empirically one of the strongest factors influencing memory[108–111]. However, our framework offers two promising directions. First, emotionally relevant aspects of experience can be prioritised by the RDT distortion function. This aligns with how rewards have been integrated with generative models in reinforcement learning contexts[112,113], with emotions mediating reward-related computations[114]. Second, we proposed novelty and surprise to determine resource allocation in variable-rate encoding. Since emotional salience is indicative of whether the episode is expected to be retrieved in the future[115], high salience implies an increased rate of the encoding. Accordingly, traumatic experiences may be understood as an extreme case of encoding primarily uninterpreted sensory features, which aligns with qualitative properties of PTSD[116,117]. More broadly, the combination of Neurath's ship with the episodic life raft may prove fertile ground for a deeper, computational understanding of traumatic events on memory and their long-term effects on development.

While our Perspective focuses on how the combination of episodic and semantic memory support learning an effective model of the environment, these are unlikely to be the only learning systems an intelligent agent needs[41], just as there are multiple memory systems[1]. While RDT helps illuminate some of these differences[12], additional computational considerations — such as trade-offs in computational cost[35,60] and the path-dependent co-evolution of these systems[118]—are also likely to play a crucial role.

Display items

Box 1 - RDT and human memory

Rate distortion theory (RDT) provides a normative framework for how to optimally encode information in settings where resource limitations make it impossible to have lossless reconstruction[17]. Here, consistent regularities in the environment can be exploited to remove redundant information, with a fundamental trade-off that balances the degree of compression (i.e., rate) with encoding accuracy (i.e., distortion), producing a continuum of compression strategies (Fig. 1c). However, RDT has only recently emerged as a normative framework for human memory[11,12,119,120], supported by the development of generative machine learning models known as variational

autoencoders[25,27] (VAEs). Generative models can learn to generate new stimuli consistent with their training data, often by "encoding" the stimuli into a latent representation and then "decoding" it to produce a (typically imperfect) reconstruction of the original stimulus. Intuitively, this process resembles the encoding and decoding of a memory trace (Fig. 1b). Indeed it has been shown that VAEs, and specifically an extended version called a beta-VAE[26] can be interpreted as an approximate solution to RDT[26,28,29] and the internal generative model in the brain, either couched explicitly in the normative framework of RDT[11,12] or relying on a qualitative match to human data[13,14,32]. Altogether, RDT provides three principles for memory, based on prior knowledge, capacity limits and task-dependency.

While this provides an appealing framework for human memory, investigations have been hampered by the inadequacy of methods for learning generative models of natural environments. Thus, engineered compression algorithms typically produce compression artefacts or "memory distortions" (e.g., blocky artefacts in images) that are qualitatively different from what we observe in human experiments (exemplified in Fig. 1e). Modern machine learning methods, and in particular the application of deep generative models[25–27] to compression[28–30] have drastically changed this picture, enabling RDT-based models of memory dynamics that are directly applicable in complex naturalistic domains, such as human drawings, text, and even natural images[11,12,31].

The most straightforward application of RDT in the context of human memory concerns the influence of prior knowledge on recall. It follows naturally from principles of compression that domain expertise leads to more accurate recall, but only for stimuli congruent with the statistics of past observations. This has been demonstrated in studies of memory for synthetic words[121] and chess configurations[11,122]. According to RDT, when specific details of an experience are forgotten, they are reconstructed using the generative model based on a high-level interpretation ("gist") of the stimulus. The resulting distortions, such as the appearance of a monocle on the Monopoly Man[5], are known in the memory literature as gist-based distortions[1,6,8]. A well-known example is the Deese-Roediger-McDermott (DRM) effect[95], where recalling lists of semantically related words often leads to the recall of a strongly related but non-presented "lure" item with nearly the same probability as presented items. The influence of gist-based distortions also implies that when the interpretations of ambiguous stimuli are manipulated (e.g., via contextual cues), both recall accuracy and the nature of

distortions are affected[96], with both phenomena capable of being reproduced using RDT[11,14].

Additionally, RDT can also naturally account for the effect of varying resource constraints. Influential theoretical analyses of memory suggest that the likelihood of a piece of information being needed decreases with time [123–125]. This can be incorporated into RDT by varying the targeted point on the rate-distortion curve (Fig. 1c), which increases the extent of gist-based distortions as a function of delay before recall[11,12], and is a robust feature of human memory[126–128].

Lastly, RDT also accounts for how memory is affected by goals and the task an individual is performing. RDT can incorporate these factors through the distortion function, for example, by overweighting errors related to danger or reward[12,129]. This degree of freedom in RDT can also be exploited to optimise for the goal of prediction, which can be shown to imply a need for updating parameters rather than a precise reconstruction of stimuli[130,131].

Box 2 - Structure learning and Neurath's ship

Structure learning refers to a class of learning problems in which competing hypotheses differ not only in the precise numerical values of parameters, but also qualitatively, such as in the number of parameters, choice of variables, forms of relationships, and even the fundamental building blocks used to specify the model. Normative theories of learning often decompose learning problems into these two processes: determining the high-level structure of the model, and fine-tuning the parameters while keeping the structure fixed[132,133]. Some approaches further distinguish between *structure* and *form*, where a transition in the form of the model is a rare but fundamental shift, such as a child deciding to organise animal species into a tree structure rather than separate clusters[48]. However, for simplicity, we use structure learning here in the broader sense, encompassing both structure and form.

Two properties of structure learning make it fundamentally more challenging than parameter estimation. First, just as causal graphs are constructed from nodes and directed edges, other structure learning problems are often defined by specifying

primitive components along with rules for their composition. These composition rules are typically open-ended, allowing arbitrarily complex structures to be "grown" over the course of learning[134]. While compositionality enables such models to construct genuinely novel explanations, it also results in inconceivably vast hypothesis spaces[135,136]. The second difficulty lies in navigating these spaces. To illustrate, imagine the "learning landscape", with the horizon spanned by possible configurations of the model, and height of the terrain defined by the "goodness of fit" for that particular configuration (e.g., Fig. 2c). In parameter estimation, this landscape is typically smooth and continuous, with small changes in parameters resulting in small changes in model predictions. However, in structure learning, the possible configurations are typically discrete, and neighbouring points may sometimes correspond to dramatically different predictions, making the terrain rugged and treacherous[107].

These difficulties in searching over structures during individual learning mirror those encountered in the development of scientific theories, making the Neurath's ship analogy, originally proposed in the latter context, applicable to the former as well[50]. Beyond offering an evocative analogy, the iterative rebuilding of Neurath's ship can be precisely formalised as a specific type of approximate structure learning within the framework of hierarchical Bayesian inference. The Bayesian solution for uncertainty involves keeping track of all possibilities, summarising them in the posterior distribution. Ideally, hierarchical Bayesian inference prescribes computing the posterior distribution over all structural hypotheses, updating each, in parallel, with incoming observations. However, in practice, Monte Carlo approximations are commonly used, where on the highest levels of the hierarchy, only a restricted set (or even a single hypothesis) of "particles" are tracked. By keeping the high-level structural hypothesis fixed, the posteriors over parameters are much less resource intensive to maintain. In this class of Monte Carlo algorithms, the process for updating the model structure is encoded in the *proposal distribution*. This distribution specifies, for each hypothesis, what alternative hypotheses may be considered in a single update. The iterative replacement of the planks and beams in Neurath's ship implies making this proposal distribution local, for example by only allowing the addition or removal of a single causal edge.

Our prototypical example has been that of causal learning[50]. However, structure learning problems are ubiquitous in natural environments, encompassing contextual learning[137], the identification of underlying structural forms within data[48], and learning

visual[105,138,139105,138] or abstract concepts[56,60,104,106,107,136,140]. Structure learning has also been implicated in event segmentation[100–102], where the temporal structure of visual or auditory information stream needs to be discovered[83,84,141]. At the most general level, the composable building blocks for theories may define components of a programming language, making learning akin to program induction[104–107].

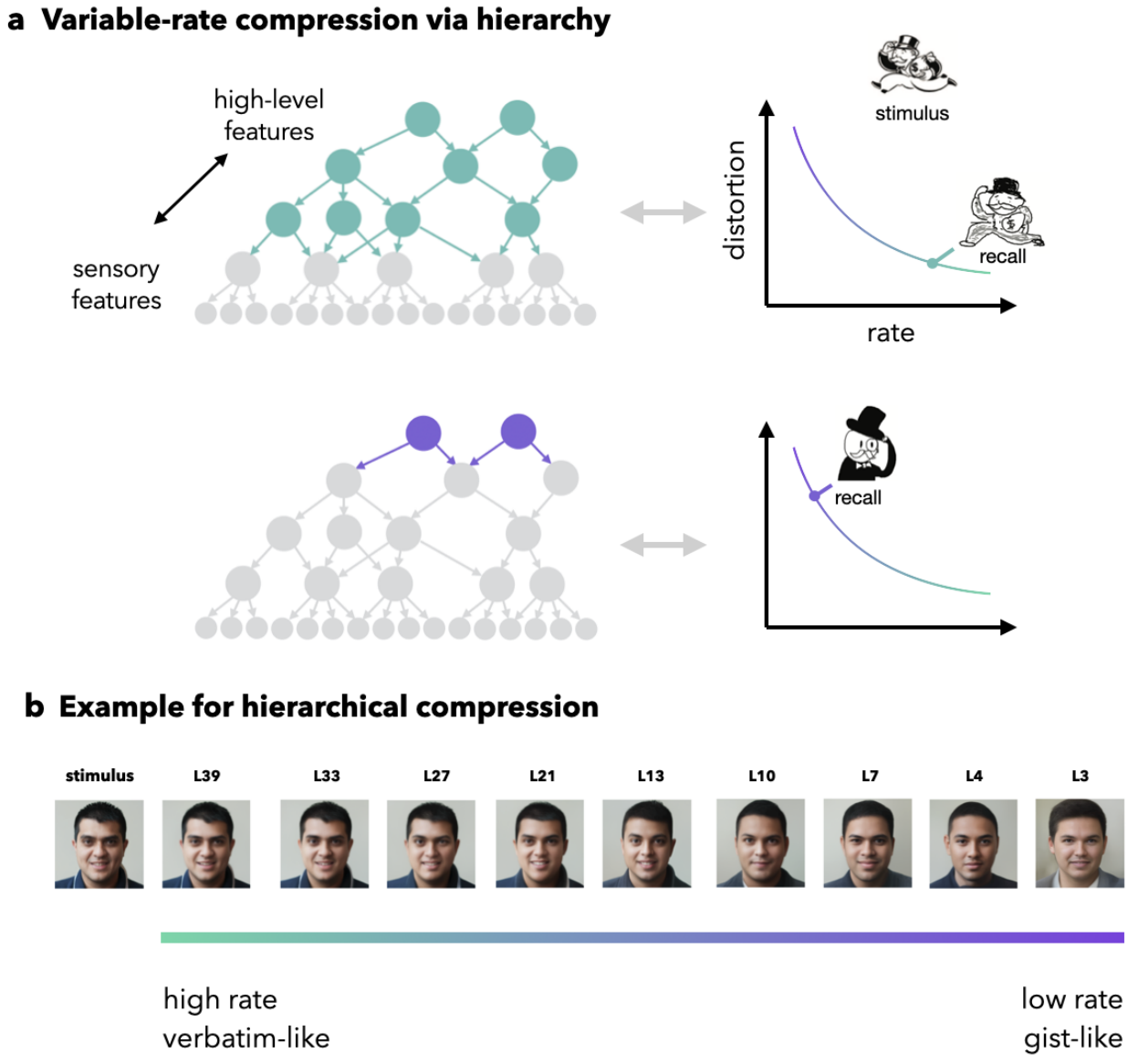
Box 3 - Implementation of variable-rate compression in the brain

The idea of variable-rate encoding integrates both components of our proposed framework for the interactions of the semantic and episodic memory systems (Fig. 2). In this view, sensory experience is compressed via semantic memory, producing episodic memory traces with model-congruent distortions. This increases the effective capacity of episodic memory, but may also hinder its proposed role in retaining seemingly irrelevant details crucial for structure learning. The role of the variable-rate mechanism is to enable the system to selectively retain precise details when surprise or novelty suggests the compression model may be unreliable.

Two key challenges stand out in bridging the computations of variable-rate compression with a neural implementation in the brain. First, while RDT allows encoding at multiple compression levels (Fig. 1c), each level relies on a distinct encoder-decoder pair optimized for a specific rate-distortion trade-off. This implies the maintenance of multiple semantic models in parallel, contradicting the principle behind Neurath's ship. Second, RDT provides no mechanism for converting from detailed to more compressed memory traces after the initial encoding (e.g., forgetting over time; Box 1).

Multiple approaches have been proposed in machine learning for achieving variable-rate compression while avoiding multiple encoder-decoder pairs[142–144]. A promising idea for how the brain might implement variable-rate compression is through a hierarchical generative model[145–148], with layers deeper in hierarchy corresponding to progressively stronger levels of compression (top panel of figure). Similar to these models, the visual cortex is thought to represent sensory information hierarchically: early layers respond to basic features like edges, intermediate layers detect textures, and deeper layers

integrate these features into representations of objects and scenes. Hierarchical models offer a straightforward way to reduce memory capacity by selectively discarding information over time, starting with lower-level details (bottom panel of figure). An intriguing possibility is to associate certain layers of the cortical hierarchy with layers of variables in the generative model. Retaining a subset of these activations in hippocampal regions and reinstating them in the relevant cortical layers during recall has been proposed as a mechanism for memory storage and retrieval[149]. Memory resources could then be reduced by sequentially discarding the activations of increasingly deep layers in the memory trace, such that traces with more episodic details rely on earlier layers of the sensory hierarchy. Similar hypotheses have been proposed in the context of the visual hierarchy[14,32,150,151], and behavioural evidence suggests that this framework may extend to other modalities as well[152].



Acknowledgements

We thank the anonymous reviewers, the editor and Neil Bramley for their insightful comments and suggestions, which helped improve this manuscript. We are also grateful to Peter Dayan for valuable discussions and extensive comments on earlier drafts. Additionally, we appreciate Csenge Frater, Ryutaro Uchiyama and Mihaly Banyai for their helpful feedback on the manuscript. This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645, and funded by the DFG under Germany's Excellence Strategy – EXC 2117 – 422037984. G.O. was supported by a grant from the Human Frontiers

Science Program, and by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory in Hungary.

References

1. Baddeley, A., Eysenck, M. W. & Anderson, M. C. *Memory*. (Routledge, London, England, 2020).
2. Nickerson, R. S. & Adams, M. J. Long-term memory for a common object. *Cogn. Psychol.* **11**, 287–307 (1979).
3. Martin, M. & Jones, G. V. Generalizing everyday memory: signs and handedness. *Mem. Cognit.* **26**, 193–200 (1998).
4. Blake, A. B., Nazarian, M. & Castel, A. D. Rapid Communication: The Apple of the mind's eye: Everyday attention, metamemory, and reconstructive memory for the Apple logo. *Q. J. Exp. Psychol.* **68**, 858–865 (2015).
5. Prasad, D. & Bainbridge, W. A. The visual Mandela effect as evidence for shared and specific false memories across people. *Psychol. Sci.* **33**, 1971–1988 (2022).
6. Schacter, D. L., Guerin, S. A. & St Jacques, P. L. Memory distortion: an adaptive perspective. *Trends Cogn. Sci.* **15**, 467–474 (2011).
7. Wu, C. M., Meder, B. & Schulz, E. Unifying principles of generalization: Past, present, and future. *Annu. Rev. Psychol.* (2024) doi:10.1146/annurev-psych-021524-110810.
8. Reyna, V. F. & Brainerd, C. J. Fuzzy-trace theory: An interim synthesis. *Learn. Individ. Differ.* **7**, 1–75 (1995).
9. Reyna, V. F., Corbin, J. C., Weldon, R. B. & Brainerd, C. J. How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *J. Appl. Res. Mem. Cogn.* **5**, 1–9 (2016).
10. Bartlett, F. C. Remembering: A Study in experimental and Social Psychology. (1932) doi:10.1111/j.2044-8279.1933.tb02913.x.
11. Nagy, D. G., Török, B. & Orbán, G. Optimal forgetting: Semantic compression of episodic memories.

PLoS Comput. Biol. **16**, e1008367 (2020).

12. Bates, C. J. & Jacobs, R. A. Efficient data compression in perception and perceptual memory. *Psychol. Rev.* **127**, 891–917 (2020).
13. Fayyaz, Z., Altamimi, A., Cheng, S. & Wiskott, L. A model of semantic completion in generative episodic memory. *arXiv [q-bio.NC]* (2021).
14. Spens, E. & Burgess, N. A generative model of memory construction and consolidation. *Nat Hum Behav* **8**, 526–543 (2024).
15. Tompary, A. & Thompson-Schill, S. L. Semantic influences on episodic memory distortions. *J. Exp. Psychol. Gen.* **150**, 1800–1824 (2021).
16. Tandoc, M. C., Dong, C. V. & Schapiro, A. C. Object feature memory is distorted by category structure. *Open Mind* **8**, 1348–1368 (2024).
17. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).
18. Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec* (1959).
19. Barlow, H. B. Possible principles underlying the transformations of sensory messages. in *Sensory Communication* 216–234 (The MIT Press, 1961).
20. Barlow, H. Redundancy reduction revisited. *Network* **12**, 241–253 (2001).
21. Zhaoping, L. Theoretical understanding of the early visual processes by data compression and data selection. *Network* **17**, 301–334 (2006).
22. Craik, K. J. W. *The Nature of Explanation*. (Cambridge University Press, Cambridge, England, 1967).
23. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. Neurosci.* **7**, 286–294 (2004).
24. Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331**, 83–87 (2011).

25. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).

26. Higgins, I. *et al.* Beta-VAE: Learning basic visual concepts with a constrained variational framework. *Int Conf Learn Represent* (2016).

27. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv [stat.ML]* (2014).

28. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep Variational Information Bottleneck. *arXiv [cs.LG]* (2016).

29. Alemi, A. *et al.* Fixing a Broken ELBO. in *Proceedings of the 35th International Conference on Machine Learning* (eds. Dy, J. & Krause, A.) vol. 80 159–168 (PMLR, 10--15 Jul 2018).

30. Ballé, J., Laparra, V. & Simoncelli, E. P. End-to-end Optimized Image Compression. *arXiv [cs.CV]* (2016).

31. Bates, C. J., Alvarez, G. A. & Gershman, S. J. Scaling models of visual working memory to natural images. *Communications Psychology* **2**, 1–8 (2024).

32. Hedayati, S., O'Donnell, R. E. & Wyble, B. A model of working memory for latent representations. *Nat Hum Behav* **6**, 709–719 (2022).

33. Martin-Ordas, G. & Easton, A. Elements of episodic memory: lessons from 40 years of research. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **379**, 20230395 (2024).

34. Nicholas, J. & Mattar, M. G. Humans use episodic memory to access features of past experience for flexible decision making. *Proceedings of the Annual Meeting of the Cognitive Science Society* **46**, (2024).

35. Lengyel, M. & Dayan, P. Hippocampal contributions to control: The third way. *Adv. Neural Inf. Process. Syst.* 889–896 (2007).

36. Mahr, J. & Csibra, G. Why do we remember? The communicative function of episodic memory. *Behav. Brain Sci.* **41**, 1–93 (2017).

37. 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).

38. Moscovitch, M. The hippocampus as a 'stupid,' domain-specific module: Implications for theories of recent and remote memory, and of imagination. *Canadian journal of experimental psychology = Revue canadienne de psychologie experimentale* **62**, 62–79 (2008).

39. Nagy, D. G. & Orban, G. Episodic memory as a prerequisite for online updates of model structure. *CogSci* (2016).

40. Lu, Q., Hummos, A. & Norman, K. A. Episodic memory supports the acquisition of structured task representations. *bioRxiv* (2024) doi:10.1101/2024.05.06.592749.

41. 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).

42. Antony, J. W., Van Dam, J., Massey, J. R., Barnett, A. J. & Bennion, K. A. Long-term, multi-event surprise correlates with enhanced autobiographical memory. *Nat Hum Behav* **7**, 2152–2168 (2023).

43. Lin, Q., Li, Z., Lafferty, J. & Yildirim, I. Images with harder-to-reconstruct visual representations leave stronger memory traces. *Nat Hum Behav* **8**, 1309–1320 (2024).

44. Rouhani, N. & Niv, Y. Signed and unsigned reward prediction errors dynamically enhance learning and memory. *Elife* **10**, (2021).

45. von Restorff, H. Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychol. Forsch.* **18**, 299–342 (1933).

46. van Kesteren, M. T. R., Ruiter, D. J., Fernández, G. & Henson, R. N. How schema and novelty augment memory formation. *Trends Neurosci.* **35**, 211–219 (2012).

47. Pearl, J. *Causality*. (Cambridge University Press, Cambridge, England, 2013).

48. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* **105**,

10687–10692 (2008).

49. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
50. Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychol. Rev.* **124**, 301–338 (2017).
51. Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. *Cogn. Sci.* **38**, 599–637 (2014).
52. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* **117**, 1144–1167 (2010).
53. Courville, A. C. & Daw, N. The rat as particle filter. *Advances in neural information processing systems* **20**, (2007).
54. Neurath, O., Neurath, M. & Cohen, R. S. Empiricism and sociology. **1**, (1973).
55. Van Orman Quine, W. Word and Object. *MIT Press* https://mitpress.mit.edu/9780262670012/word-and-object/ (1960).
56. Zhao, B., Lucas, C. G. & Bramley, N. R. A model of conceptual bootstrapping in human cognition. *Nat. Hum. Behav.* **8**, 125–136 (2024).
57. Tse, D. *et al.* Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
58. Tse, D. *et al.* Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
59. Abbott, J. T. & Thomas, L. Exploring influence particle filter parameters order effects causal learning. *Proceedings annual meeting cognitive science society* **33**, (2011).
60. Zhou, H., Nagy, D. G. & Wu, C. M. Harmonizing program induction with Rate-Distortion Theory. in *Proceedings of the 46th Annual Conference of the Cognitive Science Society* (ed. Frank, SL and Toneva, M and Mackey, A and Hazeltine, E) 2511–2518 (Cognitive Science Society, 2024).

61. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10313–E10322 (2018).

62. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. in *Psychology of Learning and Motivation* vol. 24 109–165 (Elsevier, 1989).

63. French, R. M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).

64. Hadsell, R., Rao, D., Rusu, A. A. & Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends Cogn. Sci.* **24**, 1028–1040 (2020).

65. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2205582119 (2022).

66. Beukers, A. O. *et al.* Blocked training facilitates learning of multiple schemas. *Commun Psychol* **2**, 28 (2024).

67. Gong, T., Gerstenberg, T., Mayrhofer, R. & Bramley, N. R. Active causal structure learning in continuous time. *Cogn. Psychol.* **140**, 101542 (2023).

68. Zhou, Z., Singh, D., Tandoc, M. C. & Schapiro, A. C. Building integrated representations through interleaved learning. *J. Exp. Psychol. Gen.* **152**, 2666–2684 (2023).

69. Birnbaum, M. S., Kornell, N., Bjork, E. L. & Bjork, R. A. Why interleaving enhances inductive learning: the roles of discrimination and retrieval. *Mem. Cognit.* **41**, 392–402 (2013).

70. Zulkiply, N. & Burt, J. S. The exemplar interleaving effect in inductive learning: moderation by the difficulty of category discriminations. *Mem. Cognit.* **41**, 16–27 (2013).

71. Carvalho, P. F. & Goldstone, R. L. Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Mem. Cognit.* **42**, 481–495 (2014).

72. Chater, N. & Oaksford, M. Programs as causal models: speculations on mental programs and

mental representation. *Cogn. Sci.* **37**, 1171–1191 (2013).

73. Icard, T. F. From programs to causal models ∗. *Proceedings of the 21st Amsterdam colloquium* (2017).
74. Sanborn, A. & Griffiths, T. Markov chain Monte Carlo with people. *Adv. Neural Inf. Process. Syst.* **20**, (2007).
75. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* **23**, R764–73 (2013).
76. Baldi, P. & Itti, L. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw.* **23**, 649–666 (2010).
77. Koch, C., Zika, O., Bruckner, R. & Schuck, N. W. Influence of surprise on reinforcement learning in younger and older adults. *PLoS Comput. Biol.* **20**, e1012331 (2024).
78. Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2**, 230–247 (2010).
79. Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychol. Rev.* **101**, 53–79 (1994).
80. Nosofsky, R. M. & Palmeri, T. J. A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychon. Bull. Rev.* **5**, 345–369 (1998).
81. Piray, P. & Daw, N. D. A model for learning based on the joint estimation of stochasticity and volatility. *Nat. Commun.* **12**, 6587 (2021).
82. Modirshanechi, A., Brea, J. & Gerstner, W. A taxonomy of surprise definitions. *J. Math. Psychol.* **110**, 102712 (2022).
83. Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M. & Gershman, S. J. Structured Event Memory: A neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327–361 (2020).
84. Gershman, S. J., Radulescu, A., Norman, K. A. & Niv, Y. Statistical computations underlying the

dynamics of memory updating. *PLoS Comput. Biol.* **10**, e1003939 (2014).

85. Rouhani, N., Norman, K. A. & Niv, Y. Dissociable effects of surprising rewards on learning and memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1430–1443 (2018).
86. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
87. 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).
88. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. *arXiv [cs.LG]* (2015).
89. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
90. Aljundi, R. *et al.* Online continual learning with maximally interfered retrieval. *ArXiv* **abs/1908.04742**, (2019).
91. Botvinick, M. *et al.* Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
92. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
93. Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with Deep Generative Replay. *arXiv [cs.AI]* (2017).
94. 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).
95. Roediger, H. L. & McDermott, K. B. Creating false memories: Remembering words not presented in lists. *J. Exp. Psychol. Learn. Mem. Cogn.* **21**, 803–814 (1995).
96. 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).
97. Loftus, E. F. Planting misinformation in the human mind: a 30-year investigation of the malleability

of memory. *Learn. Mem.* **12**, 361–366 (2005).

98. Roese, N. J. & Vohs, K. D. Hindsight bias. *Perspect. Psychol. Sci.* **7**, 411–426 (2012).
99. Carpenter, A. C. & Schacter, D. L. Flexible retrieval: When true inferences produce false memories. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 335–349 (2017).
100. Baldassano, C. *et al.* Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
101. Davachi, L. & DuBrow, S. How the hippocampus preserves order: the role of prediction and context. *Trends Cogn. Sci.* **19**, 92–99 (2015).
102. Zheng, J. *et al.* Neurons detect cognitive boundaries to structure episodic memories in humans. *Nat. Neurosci.* **25**, 358–368 (2022).
103. Flesch, T., Nagy, D. G., Saxe, A. & Summerfield, C. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLoS Comput. Biol.* **19**, e1010808 (2023).
104. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The Child as Hacker. *Trends Cogn. Sci.* **24**, 900–915 (2020).
105. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
106. Ellis, K., Wong, C., Nye, M. & Sablé-Meyer, M. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the* (2021).
107. Ullman, T. D., Goodman, N. D. & Tenenbaum, J. B. Theory learning as stochastic search in the language of thought. *Cogn. Dev.* **27**, 455–480 (2012).
108. Rouhani, N., Niv, Y., Frank, M. J. & Schwabe, L. Multiple routes to enhanced memory for emotionally relevant events. *Trends Cogn. Sci.* **27**, 867–882 (2023).
109. Kalbe, F. & Schwabe, L. Beyond arousal: Prediction error related to aversive events promotes

episodic memory formation. *J. Exp. Psychol. Learn. Mem. Cogn.* **46**, 234–246 (2020).

110. Laney, C. & Loftus, E. F. Emotional content of true and false memories. *Memory* **16**, 500–516 (2008).
111. Christianson, S.-Å. Emotional stress and eyewitness memory: A critical review. *Psychol. Bull.* **112**, 284–309 (1992).
112. Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse domains through world models. *arXiv [cs.AI]* (2023).
113. Wayne, G. *et al.* Unsupervised predictive memory in a goal-directed agent. *arXiv [cs.LG]* (2018).
114. Emanuel, A. & Eldar, E. Emotions as computations. *Neurosci. Biobehav. Rev.* **144**, 104977 (2023).
115. Gagne, C., Dayan, P. & Bishop, S. J. When planning to survive goes wrong: predicting the future and replaying the past in anxiety and PTSD. *Curr. Opin. Behav. Sci.* **24**, 89–95 (2018).
116. van der Kolk, B. A. & Fisler, R. Dissociation and the fragmentary nature of traumatic memories: overview and exploratory study. *J. Trauma. Stress* **8**, 505–525 (1995).
117. Ehlers, A. & Clark, D. M. A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* **38**, 319–345 (2000).
118. Bramley, N. R., Zhao, B., Quillien, T. & Lucas, C. G. Local search and the evolution of world models. *Top. Cogn. Sci.* (2023) doi:10.1111/tops.12703.
119. Sims, C. R., Jacobs, R. A. & Knill, D. C. An ideal observer analysis of visual working memory. *Psychol. Rev.* **119**, 807–830 (2012).
120. Gershman, S. J. The rational analysis of memory. in *The Oxford Handbook of Human Memory, Two Volume Pack* 1505–1520 (Oxford University Press, 2024).
121. Baddeley, A. D. Language habits, acoustic confusability, and immediate memory for redundant letter sequences. *Psychon. Sci.* **22**, 120–121 (1971).
122. Gobet, F. & Simon, H. A. Recall of rapidly presented random chess positions is a function of skill.

Psychon. Bull. Rev. **3**, 159–163 (1996).

123. Anderson, J. R. & Milson, R. Human memory: An adaptive perspective. *Psychol. Rev.* **96**, 703–719 (1989).

124. Anderson, J. R. & Schooler, L. J. Reflections of the environment in memory. *Psychol. Sci.* **2**, 396–408 (1991).

125. Roediger, H. L., 3rd & DeSoto, K. A. Cognitive psychology. Forgetting the presidents. *Science* **346**, 1106–1109 (2014).

126. Hanawalt, N. G. & Demarest, I. H. The effect of verbal suggestion in the recall period upon the reproduction of visually perceived forms. *J. Exp. Psychol.* **25**, 159–174 (1939).

127. Toglia, M. P., Neuschatz, J. S. & Goodwin, K. A. Recall accuracy and illusory memories: when more is less. *Memory* **7**, 233–256 (1999).

128. Seamon, J. G. *et al.* Are false memories more difficult to forget than accurate memories? The effect of retention interval on recall and recognition. *Mem. Cognit.* **30**, 1054–1064 (2002).

129. Sims, C., Ma, Z., Allred, S. R., Lerch, R. & Flombaum, J. I. Exploring the cost function in color perception and memory: An information-theoretic model of categorical effects in color matching. *CogSci* **38**, (2016).

130. Alemi, A. A. Variational Predictive Information Bottleneck. *arXiv [cs.LG]* (2019).

131. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).

132. Gelman, A. *et al. Bayesian Data Analysis*. (Chapman and Hall/CRC, 2013).

133. Grunwald, P. D. *The Minimum Description Length Principle*. (MIT Press, London, England, 2007).

134. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).

135. Fränken, J.-P., Theodoropoulos, N. C. & Bramley, N. R. Algorithms of adaptation in inductive

inference. *Cogn. Psychol.* **137**, 101506 (2022).

136. Rubino, V., Hamidi, M., Dayan, P. & Wu, C. M. Compositionality Under Time Pressure. in *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (eds. Goldwater, M., Anggoro, F., Hayes, B. & Ong, D.) (Cognitive Science Society, 2023). doi:10.31234/osf.io/z2648.
137. Heald, J. B., Lengyel, M. & Wolpert, D. M. Contextual inference in learning and memory. *Trends Cogn. Sci.* **27**, 43–64 (2023).
138. Orbán, G., Fiser, J., Aslin, R. N. & Lengyel, M. Bayesian learning of visual chunks by human observers. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2745–2750 (2008).
139. Austerweil, J. L., Sanborn, S. & Griffiths, T. L. Learning How to Generalize. *Cogn. Sci.* **43**, e12777 (2019).
140. Piantadosi, S. T., Tenenbaum, J. B. & Goodman, N. D. Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition* **123**, 199–217 (2012).
141. Shin, Y. S. & DuBrow, S. Structuring memory through inference-based event segmentation. *Top. Cogn. Sci.* **13**, 106–127 (2021).
142. Choi, Y., El-Khamy, M. & Lee, J. Variable rate deep image compression with a conditional autoencoder. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019). doi:10.1109/iccv.2019.00324.
143. Yang, Y., Bamler, R. & Mandt, S. Variable-bitrate neural compression via Bayesian arithmetic coding. *ICML* **abs/2002.08158**, (2020).
144. Bae, J. *et al.* Multi-rate VAE: Train once, get the full rate-distortion curve. *arXiv [cs.LG]* (2022).
145. Gregor, K., Besse, F., Rezende, D. J., Danihelka, I. & Wierstra, D. Towards Conceptual Compression. *arXiv [stat.ML]* (2016).
146. Maaløe, L., Fraccaro, M., Liévin, V. & Winther, O. BIVA: A very deep hierarchy of latent variables for generative modeling. *arXiv [stat.ML]* (2019).

147. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *arXiv [cs.NE]* (2017).

148. Child, R. Very deep VAEs generalize autoregressive models and can outperform them on images. *arXiv [cs.LG]* (2020).

149. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus* **17**, 1158–1169 (2007).

150. Bányai, M., Nagy, D. G. & Orbán, G. Hierarchical semantic compression predicts texture selectivity in early vision. in *2019 Conference on Cognitive Computational Neuroscience* (Cognitive Computational Neuroscience, Brentwood, Tennessee, USA, 2019). doi:10.32470/ccn.2019.1092-0.

151. Brady, T. F., Robinson, M. M. & Williams, J. R. Noisy and hierarchical visual memory across timescales. *Nat. Rev. Psychol.* **3**, 147–163 (2024).

152. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).