Iteratively learned compression defines the interplay between episodic and semantic memory

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] Department of Computational Neuroscience, 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. Recently, Rate Distortion Theory (RDT) has been proposed to formulate this process as lossy compression, accounting for a wealth of experimental observations. Despite its successes, RDT has a glaring problem: it assumes that observation statistics are known and unchanging, dismissing surprising experiences as noise. However, the brain's model of environmental statistics (semantic memory) must be continually learned and refined, during which surprising events play a pivotal role. In this Perspective, we argue that such iteratively learned compression produces a characteristic sensitivity to training curricula; a focus of recent research on learning. We suggest that this iterative process provides novel insights into the role of episodic memory: preserving experiences in 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[10]. Memory can be strikingly inaccurate even for frequently

encountered stimuli such as coins[11], traffic signs[12], corporate logos[13], or icons from popular culture[14]. Rather than being random, many of these memory distortions and biases are systematic[15] 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[14].

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 such as prediction, generalisation, decision-making and creativity[15]. Yet, this leaves the question: what computational principles underlie the way memory encodes past experiences in service of these goals?

A recently emerging normative perspective on this question is that of *compression* – the efficient use of inevitably bounded memory resources. If the encoding of sensory stimuli into memory traces (Fig. 1a) is indeed shaped by a need for efficient compression, then the normative framework developed for achieving such an outcome in engineered systems, Rate Distortion Theory (RDT), should provide specific insights into the dynamics and use of memory[1,2,16]. Indeed, recent parallel lines of research have demonstrated that many distortion effects may be parsimoniously explained using RDT as a framework for memory[1,2] (Fig. 1b; Box 1).

RDT starts from the observation that consistent regularities in the environment can be exploited to remove redundant information[17]. This idea has been influential in neuroscience since the 1960's under the rubric of the Efficient Coding Hypothesis[18–20]. However, when resources are insufficient for perfect reconstruction, RDT allows even further compression by strategically discarding information with as minimal an impact as possible on the reconstructed stimuli. In fact, RDT presents a continuum of compression strategies (Fig. 1c) that balances the degree of compression (i.e., *rate*) with encoding accuracy (i.e., *distortion*). When information is discarded from memory, previously observed regularities can be used to reconstruct it, at the cost of distortions that often

make the recalled stimuli more prototypical (such as adding a monocle; Fig. 1d). This aligns with one of the earliest and most influential theories of memory distortions, proposing that they reflect the influence of pre-existing knowledge through structures called memory schemas[21], a concept which served as a rich source of insights ever since[22].

In memory research, the two forms of memory we've discussed map onto *episodic memory*, containing direct (albeit distorted) representations of prior sensory experience, and *semantic memory*, containing representations of previously observed environmental regularities[23]. The distinction has since extensively been shown to be supported by distinct neural processes[24,25]. *Semantic memory* is thought to accumulate general knowledge about the environment, with strong arguments supporting the normative perspective that these regularities are represented in the form of an internal *generative model* of the environment, which enables prediction and interpretation of ongoing experience[26–29]. In contrast, *episodic memory* retains traces of specific events and, as opposed to the "summary" provided by semantic memory, shows signatures of accessing rather more raw sensory experience[30]. However, the normative role of this memory system has historically been seen as more elusive, and the subject of various proposals[31–37].

Recent proposals applying RDT to memory distortions build on the distinction between memory systems by arguing that semantic memory provides the encoding framework for the efficient compression of episodes[1,28,38]. While this provides an appealing framework for human memory, investigations have been hampered by the inadequacy of methods for learning a generative model of regularities characteristic 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[39–41] to compression[42–44] 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[1,2,4].

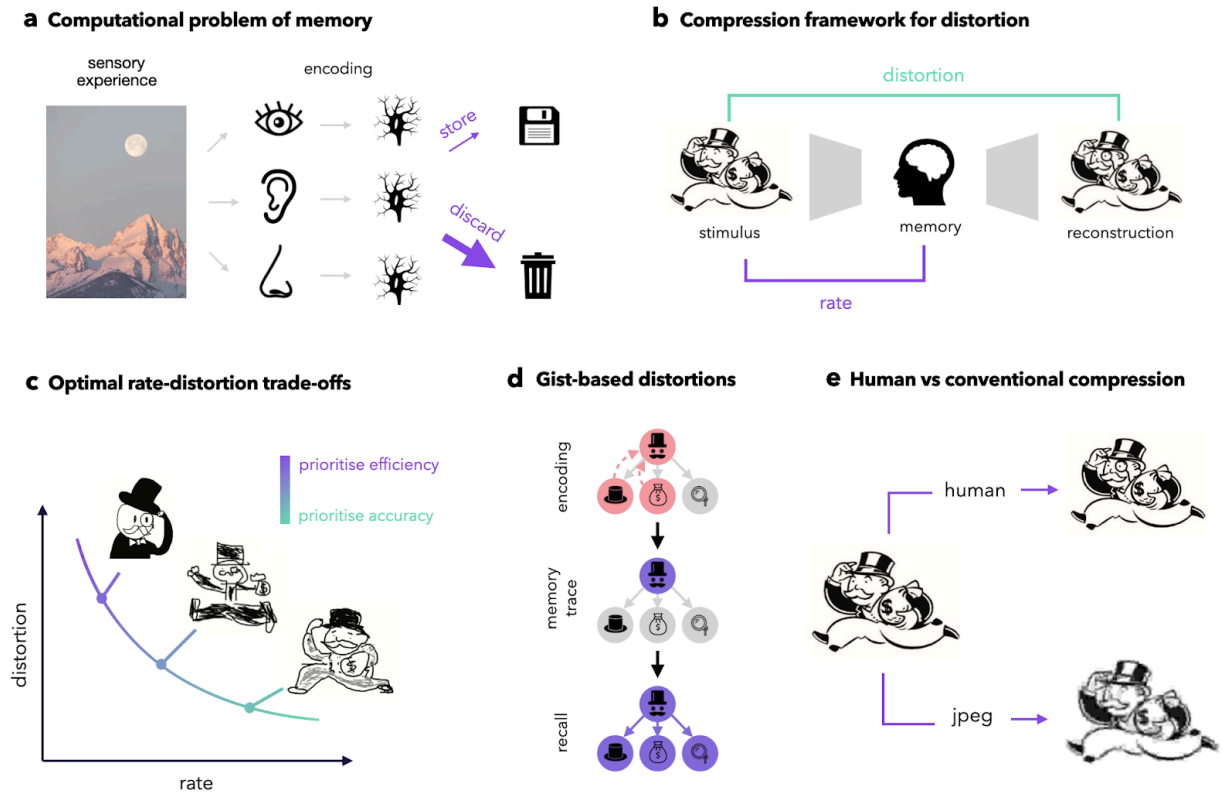

Fig 1. Compression in memory. a. The core computational challenge of memory is determining how to adapt the brain based on incoming experiences. Viewed as a lossy compression process, the critical question is which information to retain and what can be safely discarded to conserve memory resources. **b.** 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)[14]. **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, RDT as a normative framework for human memory (Box 1) suffers from a glaring issue. While it provides a unifying explanation for a wide variety of memory distortions and biases, RDT does not consider a key piece of the challenge for memory: learning and updating the 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 online learned compression and offer a resolution through the combination of semantic and episodic memory systems. We then review literature on curriculum

sensitivity in human learning, supporting our view of the updating of the internal generative model as online structure learning under a limited approximation. This allows us to contrast the predictions with that of Complementary Learning Systems (CLS), an influential account of the interaction between episodic and semantic memory[33,45]. Next, we turn to the question of what is stored in episodic memory and how, interpreting recent theoretical and empirical investigations into prioritisation in memory and experience replay in 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.

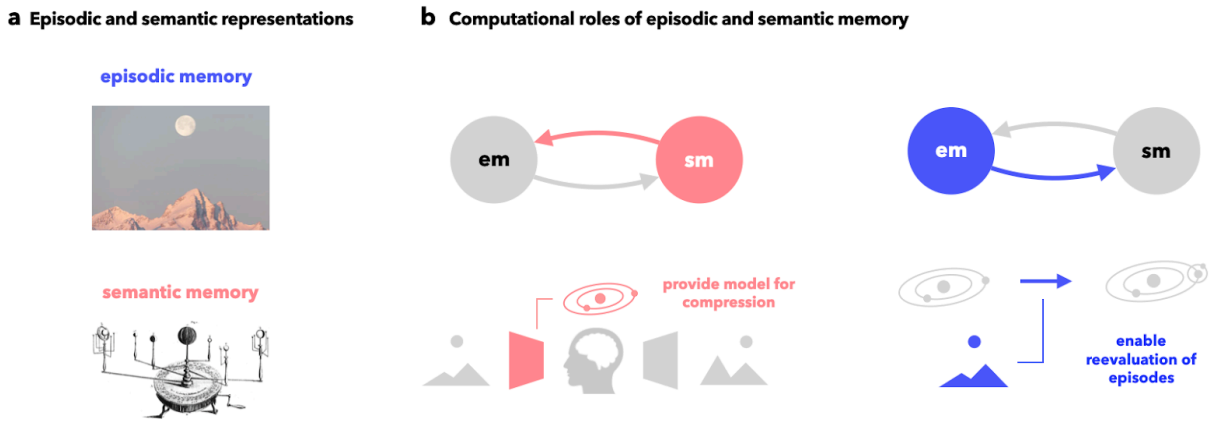

Fig 2. Interactions between semantic and episodic memory. a. We conceptualise episodic memory as retaining traces of individual experiences, and semantic memory as a simplified internal model of the environment, formalised as a probabilistic generative model over experiences. **b.** Our framework offers 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[1,2,4,28,38], 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 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”[19]. This divergence in assumptions also leads to a divergence in predictions: for RDT, the only available interpretation for surprising aspects of experience is that they are the result of noise, 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 more accurately[22,46–49]

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 updates to the model. 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 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 way of telling what went wrong, or even worse, mistakenly attributing the failure to the wrong variable and therefore making erroneous parameter 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* (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

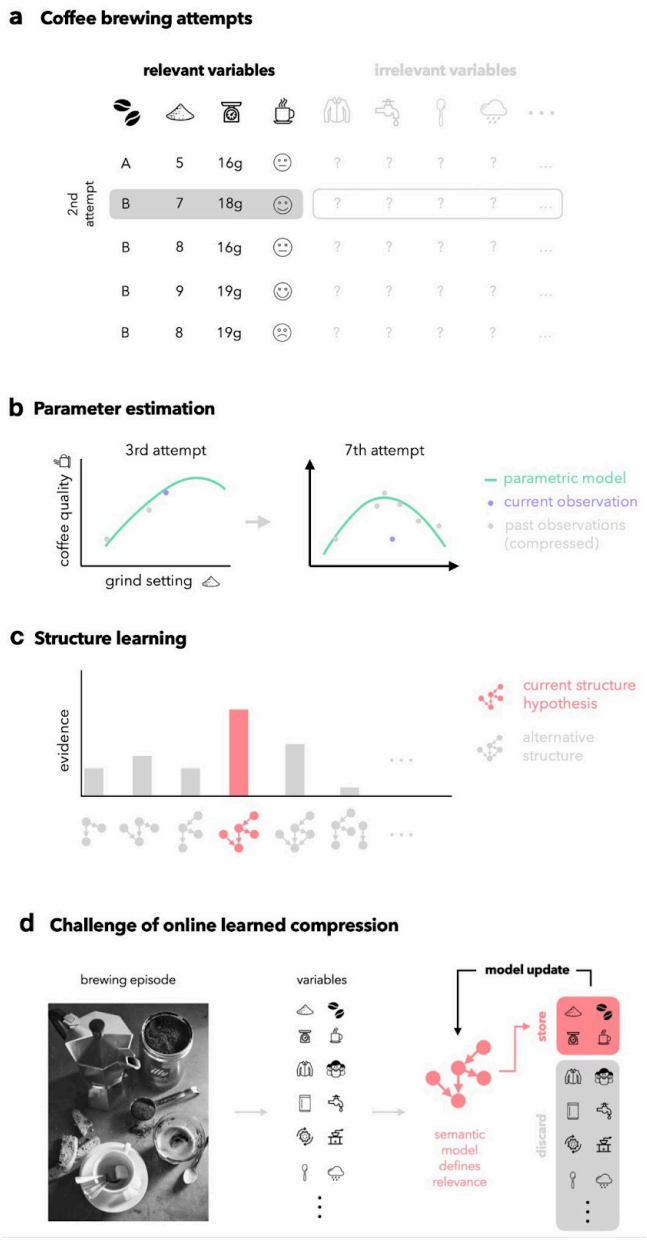

Fig 3. Online structure learning in the coffee brewing example. a. Each episode consists of values for relevant (black) and irrelevant (grey) variables. In online learning, the information content of irrelevant variables are not retained in the parameters, preventing learning about the regularities governed 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. **c.** Structure learning entails a discrete learning problem where qualitatively different hypotheses are evaluated based on some measure of congruence with observations.in Bayesian inference relies on calculating the model evidence for the different potential structures. The structure determines the set of relevant variables and their connections, and iterative learning means that, through evaluating the evidence, we find structures that discover the relevance of hitherto irrelevant variables. **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.

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), which is consistent with recent empirical results and serves as a contrast to the predictions of CLS. 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 efficient allocation of memory resources necessitates the development of a generative model of the environment (i.e., semantic memory), a process inherently linked to structure learning (see Box 2). 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 (see Box 2), these 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, by 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 interpret experience under vastly different hypotheses about the causal structure of the world. 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 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 (for details, see Box 2). 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 their environment (Fig 4a).

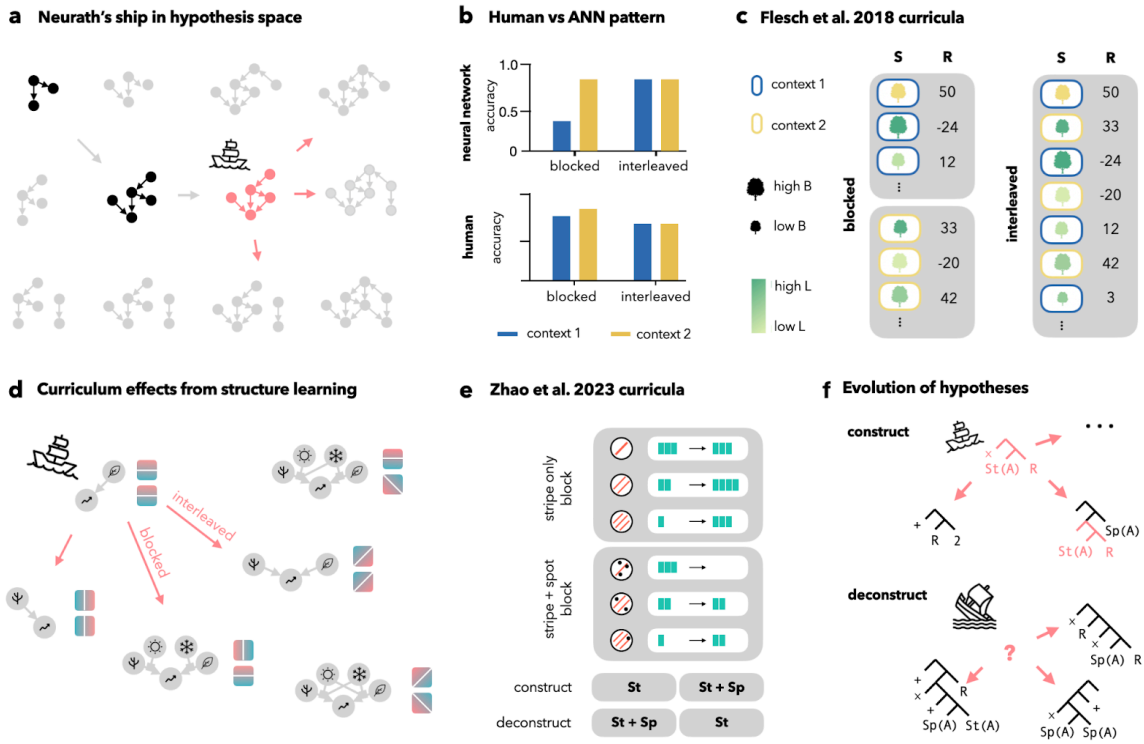

Fig 4. Neurath's ship analogy for human learning **a.** Neurath's ship represents the evolution of structural hypotheses during learning. **b.** Comparison of curriculum dependence in ANNs versus human learners. *Blue* and *yellow* bars indicate the first and second tasks, respectively. ANNs suffer from 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[8]. 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[22,56,57], 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 learning stranded with a dead-end hypothesis. This entanglement of learning and compression in the Neurath's ship approximation leads to a learning dynamics with specific patterns of order dependence, which are consistent with those observed in human data[9,35,58] but distinct from CLS and the Artificial Neural Networks (ANNs) upon which it depends.

Artificial neural networks (ANNs) also display curriculum effects, but in an opposite pattern to humans. In a study by Flesch and colleagues[5], 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"[59,60]. But when the training data was *interleaved*, with a shuffled ordering of the same data, ANNs could learn reliably[61]. CLS theory[33,45], one of the most influential proposals for why an episodic memory system is required, was concerned with exactly the challenge of mitigating catastrophic forgetting. By retaining episodes and interleaving them with current observations, CLS offers a solution for protecting older knowledge from being overwritten. In contrast with 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 in humans[6,7].

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. This creates a smaller 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 the formation of an initial hypothesis difficult . Furthermore, without the interpretative structure provided by an effective hypothesis in semantic memory, 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 recent study by Zhao and colleagues[8] establishes a more direct connection between curriculum dependence in human learning and the problem of structure discovery. 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 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 for the brain requires it 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. We proposed that this approximation leads to 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 artificial neural networks (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, however, learning the underlying structure requires holding onto seemingly aspects of experience 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 attend to information necessary for further adaptation of their 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 observed 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. 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 a potential alternative hypothesis. Therefore, the ability to remember is crucial to ensure that future learning remains possible.

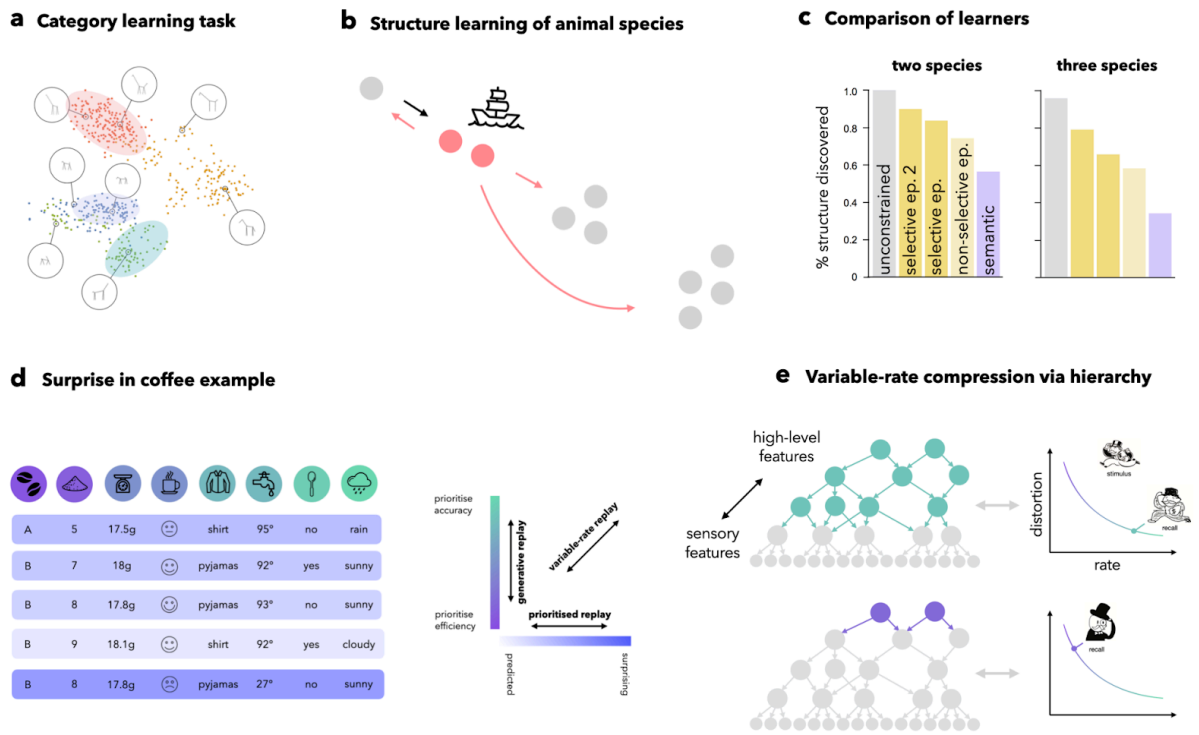

Fig 5. The episodic "life raft". a. In a category learning task classes disjunct classes are inferred from data, illustrated by a 2D embedding of stick figure animals[62]. The segment lengths and angles cluster around species but individual species are characterised by considerable variance. **b.** Structure 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. It is the selective episodic learner which assesses the surprise of experiences under the current hypothesis. **d.** Variable-rate encoding of experiences. A growing set of variables permits an increasing level of detail for encoding an experience, corresponding to dedicating more resources to encoding. The level of resources will thus determine the level of accuracy of the recalled experience, which then could contribute to learning the structure of the problem at hand. Assessment of the level of surprise can help in determining the balance between prioritising the resources or encoding fidelity. **e.** Compression at variable rates using a single hierarchical model. The depth of latent variables used to encode an experience maps onto the amount of resources allocated to memory. A richer representation in the hierarchical generative model ensures more faithful recall of the experience, which can be translated to a shift along the RDT curve.

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[35] provides an intuitive demonstration of the value of such a "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, by acting as an effective safeguard against incorrectly discarding relevant information (Fig. 5c).

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 this question in a simplified context where episodes can be recalled perfectly, then extend it to also consider how these detail-rich episodes may be stored by the brain. 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, the computational study on category learning found that prioritising experiences with high Bayesian surprise[63,64] enhances the benefit of episodic memory in online structure learning (Fig. 5c). An intriguing possibility suggested by our framework is that, among various forms of surprise, it may be most beneficial to retain experiences that are informative about the model structure. This criterion would prioritise experiences that seem surprising under the current hypothesis but well-explained by an alternative[65].

The idea that incongruent and novel information is selectively prioritised in memory has a long history in psychology[49] and is supported by extensive empirical findings[22,46–48]. Similarly, in neuroscience, the idea that the hippocampal formation (associated with episodic memory) serves to retain novel information has been extensively explored[22], particularly in the context of experience replay and memory consolidation[34,45,66,67]. The normative question of how episodes should be prioritised is typically referred to as *prioritised replay*[68], and is also applicable to CLS theory[45,67]. Since these approaches are typically unconcerned with constraints on memory resources, prioritisation in this context refers not to which episodes should be retained, but to how frequently they should be replayed. Prioritising episodes based on their associated reward prediction error has been instrumental in recent machine learning advancements[68–70]. 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[31,71,72].

So far, in our discussion of prioritisation, we have considered a simplified normative problem, where selected episodes could be recalled perfectly. In this simplified context, the question has been limited to which experiences should be selectively retained or preferentially reactivated, corresponding to *prioritised* rather than *uniform replay*. However, perfect recall is likely inconsistent with resource constraints for the brain, as well as empirical findings[10,14,21] on human memory distortions. Interpreted from the perspective of RDT, this literature suggests that episodes are compressed, with missing details filled in by a generative model maintained in semantic memory. Consistent with these findings, *generative replay*[72,73] offers a potential solution by simulating episodes from a generative model, thus placing significantly lower demands on memory resources compared to storing and replaying episodes under *exact replay*. Indeed, a recent

extension of CLS[28] has used a similar approach to better reconcile it with the memory distortion literature.

While compressing episodes using semantic memory enables significant savings in memory resources[74], 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 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. 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 episodes would be stored in a highly compressed form, making them prone to model-congruent distortions, consistent with the RDT perspective[1,2,38]. 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 explain why such experiences are often recalled more accurately[22,46–48,64].

A variable-rate encoding integrates both components of our proposed framework for the interactions of the semantic and episodic memory systems (Fig. 2). It leverages the benefits of both prioritised and generative replay by retaining important episodes with higher fidelity while conserving resources by generating details where they are unlikely to hinder further learning (Fig. 6). A promising idea for how the brain might implement such variable-rate compression is through a hierarchical generative model, with layers corresponding to progressively stronger levels of compression (Fig. 5e). For example, 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. 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[75]. 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 (Fig. 5f). Similar hypotheses have been put forward in the context of the visual hierarchy[3,28,76,77], and behavioural evidence suggests that this framework may extend to other modalities as well[78].

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. The approximations required for online structure learning predict characteristic patterns of curriculum sensitivity that are consistent with human learning and in contrast to traditional CLS accounts. Additionally, 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 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.

In the study of both human and machine learning, it is typical to present stimuli in a randomised fashion, with simple or nonexistent 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 and tasks. While a significant portion of the "natural curriculum" for human learners is self-generated[79], it is also augmented by direct instruction explicitly designed with the goal of transmitting knowledge[80]. Therefore, experiments investigating curriculum sensitivity in human learning are essential not only to study behaviour in ecologically relevant contexts, but also for uncovering normative principles for education. While the experiments we reviewed in Section 3 are important steps in this direction, future research could target a more fine-grained understanding of path-dependencies in learning. This will require both theoretical and empirical advances focusing on 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[81–83], where a goal of curriculum design is to induce widely applicable and composable conceptual modules that further learning can build on[8,9,84].

As discussed in Section 3, curriculum effects in standard ANNs do not mirror patterns in human learning in important respects[5] Thus, it is important to continue exploring how to reconcile the learning dynamics of artificial models with human empirical data[85]. Notably, deep reinforcement learnings often also exhibit a primacy bias similar to what is implied by our framework, where an over-reliance on the model formed on the basis of earlier experiences compromises further learning[86]. The factors underlying this behaviour are not well understood, and it is recommended that learned representations are periodically "reset"[86]. However, it may be fruitful to explore whether our proposal offers a practical alternative for reinforcement learning..

Classical memory distortion experiments such as those explained by RDT typically investigate the effects of semantic knowledge that was acquired by the subject prior to the experiment. A convincing demonstration of our understanding of these effects would be if they were shown to be inducible in controlled settings, using synthetic stimuli and statistics that subjects learned during the course of the experiment.

Empirical evidence suggests that one of the strongest factors influencing memory—both in terms of distortions and overall memory strength—is emotional salience[87–90]. RDT-based investigations have not yet focused on this aspect, however our framework offers two promising directions. First, the compression of episodes via RDT allows the distortion function to differentially preserve emotionally meaningful aspects of the experience, even when memory resources are severely constrained. This aligns with how rewards have been integrated with generative models in a reinforcement learning context[91,92], with emotions mediating reward-related computations[93]. Second, we proposed that episodes are encoded at variable rates, with the rate selected based on novelty or surprise. Beyond these factors, emotional salience is a key factor in estimating the likelihood of needing to revisit an episode in the future[94], and therefore implies an increased rate of the encoding. Traumatic experiences may be understood as extreme cases in this respect, implying an encoding based primarily on sensory features as opposed to high-level interpretation, which aligns with qualitative features of re-experiencing traumatic events in PTSD[95,96]. More broadly, our framework built on the Neurath's ship analogy and its interplay 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.

Ultimately, our proposal rectifies important gaps in how RDT and the metaphor of compression has been applied to the study of human memory. By shining light on the approximations we use to iteratively learn a semantic model of the world—one plank at a time, as in Neurath's ship—we motivate the need to also keep an "episodic life raft" containing high-fidelity representations of surprising events, in case our ship ever runs aground.

Display items

Box 1 - RDT and human memory

RDT was developed in the 1950s as an extension of information theory[17,97], intended for applications where some loss of information is acceptable (i.e., lossy compression). An intuitive example is how a streaming video gets grainy when the connection is unreliable, with better compression algorithms achieving higher image fidelity for a given speed. Specifically, RDT characterises the fundamental trade-off, that reducing

distortion (measured between the sensory input and what can be reconstructed based on a memory trace), results in an increase in the *rate* of information required (typically defined by the mutual information; see [17]). A common requirement is that the rate does not exceed some capacity limit (i.e., an information bottleneck[98]). Thus, RDT describes the optimal level of compression achievable under this constraint

However, RDT has only recently emerged as a normative framework for human memory[1,2,16,99], supported by the development of generative machine learning models known as variational autoencoders[39,41] (VAEs). Generative models can generate new stimuli consistent with their training data. A subset of generative models, including VAEs, are also capable of "encoding" the stimuli into a latent representation, which can then be "decoded" to generate 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 beta-VAE[40] can be interpreted as approximate solutions to RDT[40,42,43]. This theoretical insight has been foundational in recent work using VAEs as approximations to the internal generative model in the brain, either couched explicitly in the normative framework of RDT[1,2] or relying on a qualitative match to human data[3,28,38]. Altogether, RDT provides three principles for memory, based on prior knowledge, capacity limits and task-dependency.

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[100] and chess configurations[1,101]. 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[14], are known in the memory literature as gist-based distortions[10,15,102]. A well-known example is the Deese-Roediger-McDermott (DRM) effect[103], 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[104], with both phenomena capable of being reproduced using RDT[1,28].

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 [105–107]. 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[1,2], a robust feature of human memory[108–110].

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[2,111]. 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[112,113].

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, 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[114,115]. 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 [116]. 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[117]. While compositionality enables such models to construct genuinely novel explanations, it also results in inconceivably vast hypothesis spaces[118,119]. 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[84].

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 may be 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[120], the identification of underlying structural forms within data[116] and learning

visual[82,121,12282,121] or abstract concepts[8,9,81,83,84,119,123]. At the most general level, the composable building blocks for theories may define components of a programming language, making learning akin to program induction[81–84].

Co-authors

- **David G. Nagy**, University of Tübingen and Max Planck Institute for Biological Cybernetics
- **Gergo Orban**, CEU Centre for Cognitive Computations and HUN-REN Wigner Research Centre for Physics
- **Charley M. Wu**, University of Tübingen and Max Planck Institute for Biological Cybernetics

References

1. Nagy, D. G., Török, B. & Orbán, G. Optimal forgetting: Semantic compression of episodic memories. *PLoS Comput. Biol.* **16**, e1008367 (2020).
2. Bates, C. J. & Jacobs, R. A. Efficient data compression in perception and perceptual memory. *Psychol. Rev.* **127**, 891–917 (2020).
3. Hedayati, S., O'Donnell, R. E. & Wyble, B. A model of working memory for latent representations. *Nat Hum Behav* **6**, 709–719 (2022).
4. Bates, C. J., Alvarez, G. A. & Gershman, S. J. Scaling models of visual working memory to natural images. *Communications Psychology* **2**, 1–8 (2024).
5. 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).
6. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2205582119 (2022).
7. Beukers, A. O. *et al.* Blocked training facilitates learning of multiple schemas. *Commun Psychol* **2**, 28 (2024).
8. Zhao, B., Lucas, C. G. & Bramley, N. R. A model of conceptual bootstrapping in human cognition. *Nat. Hum. Behav.* **8**, 125–136 (2024).
9. 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).
10. Baddeley, A., Eysenck, M. W. & Anderson, M. C. *Memory*. (Routledge, London, England, 2020).
11. Nickerson, R. S. & Adams, M. J. Long-term memory for a common object. *Cogn. Psychol.* **11**, 287–307 (1979).
12. Martin, M. & Jones, G. V. Generalizing everyday memory: signs and handedness. *Mem. Cognit.* **26**, 193–200 (1998).
13. 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).
14. 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).
15. Schacter, D. L., Guerin, S. A. & St Jacques, P. L. Memory distortion: an adaptive perspective. *Trends*

Cogn. Sci. **15**, 467–474 (2011).
16. Sims, C. R., Jacobs, R. A. & Knill, D. C. An ideal observer analysis of visual working memory. *Psychol. Rev.* **119**, 807–830 (2012).
17. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).
18. Barlow, H. B. Possible principles underlying the transformations of sensory messages. in *Sensory Communication* 216–234 (The MIT Press, 1961).
19. Barlow, H. Redundancy reduction revisited. *Network* **12**, 241–253 (2001).
20. Zhaoping, L. Theoretical understanding of the early visual processes by data compression and data selection. *Network* **17**, 301–334 (2006).
21. Bartlett, F. C. Remembering: A Study in experimental and Social Psychology. (1932) doi:10.1111/j.2044-8279.1933.tb02913.x.
22. 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).
23. Tulving, E. & Donaldson, W. Organization of memory: episodic and semantic memory. (1972).
24. Prince, S. E., Tsukiura, T. & Cabeza, R. Distinguishing the neural correlates of episodic memory encoding and semantic memory retrieval. *Psychol. Sci.* **18**, 144–151 (2007).
25. Wiggs, C. L., Weisberg, J. & Martin, A. Neural correlates of semantic and episodic memory retrieval. *Neuropsychologia* **37**, 103–118 (1999).
26. Craik, K. J. W. *The Nature of Explanation*. (Cambridge University Press, Cambridge, England, 1967).
27. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. Neurosci.* **7**, 286–294 (2004).
28. Spens, E. & Burgess, N. A generative model of memory construction and consolidation. *Nat Hum Behav* **8**, 526–543 (2024).
29. 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).
30. 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).
31. Lengyel, M. & Dayan, P. Hippocampal contributions to control: The third way. *Adv. Neural Inf. Process. Syst.* 889–896 (2007).
32. Mahr, J. & Csibra, G. Why do we remember? The communicative function of episodic memory. *Behav. Brain Sci.* **41**, 1–93 (2017).
33. 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).
34. 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).
35. Nagy, D. G. & Orban, G. Episodic memory as a prerequisite for online updates of model structure. *CogSci* (2016).
36. 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.
37. 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).
38. Fayyaz, Z., Altamimi, A., Cheng, S. & Wiskott, L. A model of semantic completion in generative episodic memory. *arXiv [q-bio.NC]* (2021).
39. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
40. Higgins, I. *et al.* Beta-VAE: Learning basic visual concepts with a constrained variational framework.

Int Conf Learn Represent (2016).
41. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv [stat.ML]* (2014).
42. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep Variational Information Bottleneck. *arXiv [cs.LG]* (2016).
43. 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).
44. Ballé, J., Laparra, V. & Simoncelli, E. P. End-to-end Optimized Image Compression. *arXiv [cs.CV]* (2016).
45. 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).
46. 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).
47. 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).
48. Rouhani, N. & Niv, Y. Signed and unsigned reward prediction errors dynamically enhance learning and memory. *Elife* **10**, (2021).
49. von Restorff, H. Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychol. Forsch.* **18**, 299–342 (1933).
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. Tse, D. *et al.* Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
57. Tse, D. *et al.* Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
58. Abbott, J. T. & Thomas, L. Exploring influence particle filter parameters order effects causal learning. *Proceedings annual meeting cognitive science society* **33**, (2011).
59. 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).
60. French, R. M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).
61. Hadsell, R., Rao, D., Rusu, A. A. & Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends Cogn. Sci.* **24**, 1028–1040 (2020).
62. Sanborn, A. & Griffiths, T. Markov chain Monte Carlo with people. *Adv. Neural Inf. Process. Syst.* **20**, (2007).
63. Baldi, P. & Itti, L. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw.* **23**, 649–666 (2010).
64. 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).
65. Griffiths, T. L. & Tenenbaum, J. B. From mere coincidences to meaningful discoveries. *Cognition* **103**,

180–226 (2007).
66. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
67. 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).
68. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. *arXiv [cs.LG]* (2015).
69. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
70. Aljundi, R. *et al.* Online continual learning with maximally interfered retrieval. *ArXiv* **abs/1908.04742**, (2019).
71. Botvinick, M. *et al.* Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
72. 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).
73. Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with Deep Generative Replay. *arXiv [cs.AI]* (2017).
74. 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).
75. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus* **17**, 1158–1169 (2007).
76. 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.
77. Brady, T. F., Robinson, M. M. & Williams, J. R. Noisy and hierarchical visual memory across timescales. *Nat. Rev. Psychol.* **3**, 147–163 (2024).
78. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).
79. Smith, L. B., Jayaraman, S., Clerkin, E. & Yu, C. The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.* **22**, 325–336 (2018).
80. Csibra, G. & Gergely, G. Natural pedagogy. *Trends Cogn. Sci.* **13**, 148–153 (2009).
81. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The Child as Hacker. *Trends Cogn. Sci.* **24**, 900–915 (2020).
82. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
83. Ellis, K., Wong, C., Nye, M. & Sablé-Meyer, M. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the* (2021).
84. 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).
85. 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).
86. Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L. & Courville, A. C. The primacy bias in deep reinforcement learning. *ICML* **162**, 16828–16847 (2022).
87. 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).
88. 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).
89. Laney, C. & Loftus, E. F. Emotional content of true and false memories. *Memory* **16**, 500–516 (2008).
90. Christianson, S.-Å. Emotional stress and eyewitness memory: A critical review. *Psychol. Bull.* **112**,

284–309 (1992).
91. Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse domains through world models. *arXiv [cs.AI]* (2023).
92. Wayne, G. *et al.* Unsupervised predictive memory in a goal-directed agent. *arXiv [cs.LG]* (2018).
93. Emanuel, A. & Eldar, E. Emotions as computations. *Neurosci. Biobehav. Rev.* **144**, 104977 (2023).
94. 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).
95. 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).
96. Ehlers, A. & Clark, D. M. A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* **38**, 319–345 (2000).
97. Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec* (1959).
98. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. *arXiv [physics.data-an]* (2000).
99. Gershman, S. J. The rational analysis of memory. in *The Oxford Handbook of Human Memory, Two Volume Pack* 1505–1520 (Oxford University Press, 2024).
100. Baddeley, A. D. Language habits, acoustic confusability, and immediate memory for redundant letter sequences. *Psychon. Sci.* **22**, 120–121 (1971).
101. Gobet, F. & Simon, H. A. Recall of rapidly presented random chess positions is a function of skill. *Psychon. Bull. Rev.* **3**, 159–163 (1996).
102. Reyna, V. F. & Brainerd, C. J. Fuzzy-trace theory: An interim synthesis. *Learn. Individ. Differ.* **7**, 1–75 (1995).
103. 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).
104. 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).
105. Anderson, J. R. & Milson, R. Human memory: An adaptive perspective. *Psychol. Rev.* **96**, 703–719 (1989).
106. Anderson, J. R. & Schooler, L. J. Reflections of the environment in memory. *Psychol. Sci.* **2**, 396–408 (1991).
107. Roediger, H. L., 3rd & DeSoto, K. A. Cognitive psychology. Forgetting the presidents. *Science* **346**, 1106–1109 (2014).
108. 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).
109. Toglia, M. P., Neuschatz, J. S. & Goodwin, K. A. Recall accuracy and illusory memories: when more is less. *Memory* **7**, 233–256 (1999).
110. 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).
111. 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).
112. Alemi, A. A. Variational Predictive Information Bottleneck. *arXiv [cs.LG]* (2019).
113. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
114. Gelman, A. *et al. Bayesian Data Analysis*. (Chapman and Hall/CRC, 2013).
115. Grunwald, P. D. *The Minimum Description Length Principle*. (MIT Press, London, England, 2007).
116. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* **105**,

10687–10692 (2008).
117. 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).
118. Fränken, J.-P., Theodoropoulos, N. C. & Bramley, N. R. Algorithms of adaptation in inductive inference. *Cogn. Psychol.* **137**, 101506 (2022).
119. 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.
120. Heald, J. B., Lengyel, M. & Wolpert, D. M. Contextual inference in learning and memory. *Trends Cogn. Sci.* **27**, 43–64 (2023).
121. 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).
122. Austerweil, J. L., Sanborn, S. & Griffiths, T. L. Learning How to Generalize. *Cogn. Sci.* **43**, e12777 (2019).
123. 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).