nature reviews psychology

https://doi.org/10.1038/s44159-025-00458-6

Perspective

Check for updates

Adaptive compression as a unifying framework for episodic and semantic memory

David G. Nagy [1,2] ✉, Gergő Orbán [3,4,5] & Charley M. Wu [1,2,5]

Abstract

Sensory experiences are encoded as memories, not as verbatim copies, but through interpretation and transformation. Rate distortion theory frames this process as compression in which irrelevant details are discarded. Despite the successes of approaches based on rate–distortion theory in aligning with empirical findings, these approaches assume that environmental regularities are known and unchanging and that surprising experiences are dismissed. However, the brain's model of environmental regularities (semantic memory) is continually learned and refined, and surprising events have a pivotal role in this learning. In this Perspective, we offer a normative framework that addresses the interplay between semantic and episodic memory in the context of this computational problem that encompasses memory distortions, curriculum effects and prioritized replay. We propose to consider memory as solving an online structure learning problem, with semantic and episodic memory each having a role. We argue that semantic memory must learn the regularities that enable the efficient encoding of experience and that episodic memory supports this process by preserving surprising experiences in a relatively raw format for later interpretation. This framework opens up avenues towards understanding how adaptive compression and surprise shape the trajectory of learning and memory distortions.

Sections

[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]Department of Computational Sciences, HUN-REN Wigner Research Centre for Physics, Budapest, Hungary. [4]Center for Cognitive Computation, Central European University, Budapest, Hungary. [5]These authors contributed equally: Gergő Orbán, Charley M. Wu. ✉e-mail: david.nagy@uni-tuebingen.de

Introduction

Human memory does not store a verbatim copy of sensory experience, but rather 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[5]. A visual analogue of this effect is that a majority of participants falsely recognized manipulated versions of visual cultural iconography, such as a monocle-wearing version of the Monopoly man, even when they were presented alongside the original[5].

The extent of these memory inaccuracies might seem surprising and could be perceived as fundamental flaws of human memory. However, the primary purpose of memory is generally recognized as not merely to accurately recall past experience, but rather to support other cognitive functions[6,7] such as prediction, generalization, decision-making and creativity (Fig. 1a). For example, many of these errors fall under the category of gist-based distortions, in which 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 prioritize information most relevant for anticipating future events and guiding behaviour by interpreting experience in light of prior knowledge and expectations[10–16]. However, it is still unclear what computational principles underlie the ways in which multiple memory systems (including semantic and episodic memory) encode past experiences in service of these cognitive goals.

A normative perspective on this question that has recently gained traction in memory research highlights the importance of compression. Specifically, the mathematical framework of rate–distortion theory, which originated in the 1950s as an extension of information theory[17,18], asks how to optimally encode an input so that it fits within the available capacity budget (the rate), while taking the goals of the system into account (Box 1 and Fig. 1b). With a sufficiently large budget, perfect reconstruction (lossless compression) is possible, but the theory extends to the more general case in which even the best possible encoding leads to distortion in the reconstructed input (lossy compression). Rate–distortion theory derives a fundamental trade-off, showing that a reduction in rate leads to a corresponding increase in the minimum achievable distortion (Fig. 1c). For example, when streaming video with a poor connection, video quality is reduced to maintain smooth playback. In rate–distortion terms, the system lowers the rate of the encoding to match the available budget, which increases the distortion (in this case, the visual degradation of the image).

A key insight in rate–distortion theory is that regularities in the environment can be exploited by the encoder to remove redundant

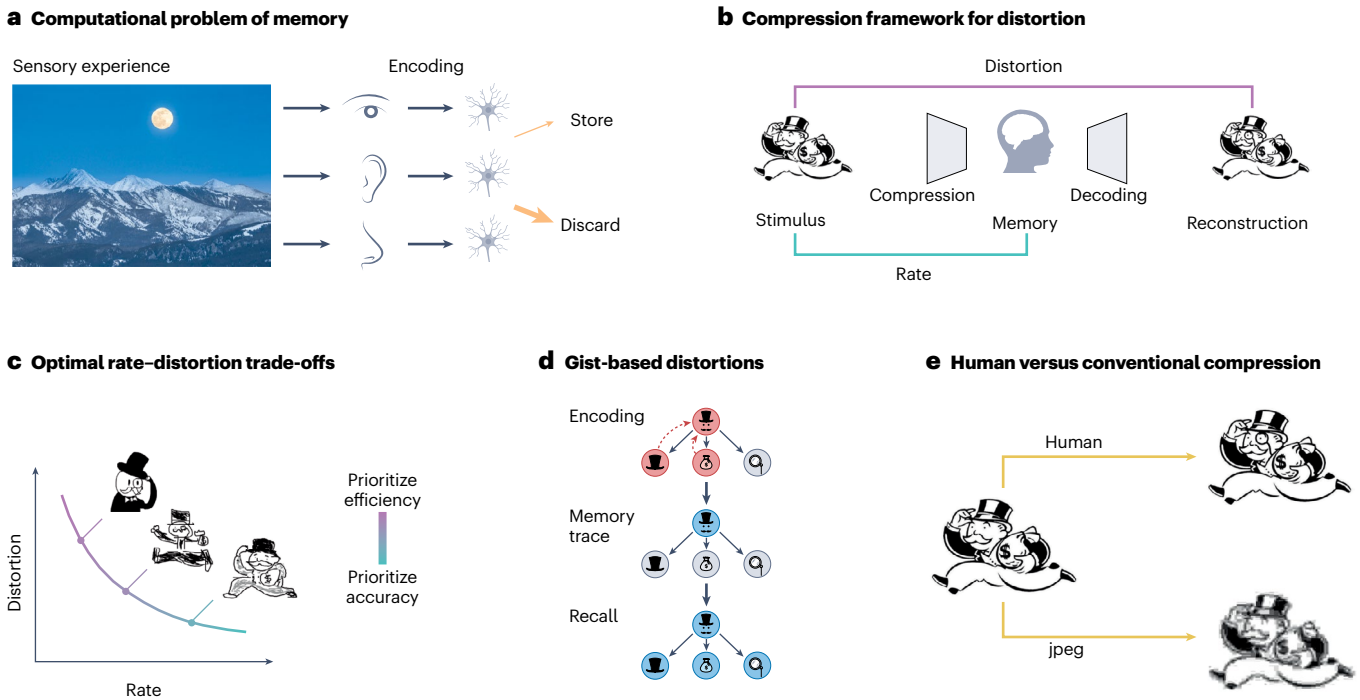


Fig. 1 | Compression in memory. a, Viewed as a lossy compression process, the core computational problem of memory is what sensory information to retain and to discard to conserve memory resources. **b**, Rate distortion theory characterizes compression algorithms using two key measures: distortion (how much the reconstructed experience differs from the original) and rate (the average amount of information preserved in the memory trace). **c**, According to rate–distortion theory, there is a continuum of trade-offs between rate and distortion. Higher rates allow more accurate recall on average, whereas lower rates result in greater distortions. **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 block compression artefacts produced by classical algorithms, such as block distortions in .jpeg files. Panel **a** (mountain photo) credit: john lambing/Alamy Stock Photo. Parts **b**,**c** and **e** adapted with permission from ref. 5, Sage.

Box 1 | Rate–distortion theory and human memory

Although rate–distortion theory provides an appealing framework for human memory[126,127], investigations of their alignment have been hampered by the difficulty of learning accurate generative models of naturalistic stimuli. Thus, engineered compression algorithms typically produce compression artefacts or 'memory distortions' that are qualitatively different from what is observed in human experiments. Modern machine learning methods, and in particular the application of deep generative models[25–27] such as variational autoencoders for compression[28–30], have drastically changed this picture. Variational autoencoders have enabled models of memory dynamics based on rate–distortion theory that are directly applicable in complex naturalistic domains, such as human drawings, text and even natural images[11,12,31].

Generative models, such as variational autoencoders, 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. Indeed, variational autoencoders — specifically, an extended version called a beta-variational autoencoder[26] — can be interpreted as an approximate implementation of rate–distortion theory[26,28,29]. These autoencoders can also be considered an analogue to the internal generative model in the brain, either couched explicitly in the normative framework of rate–distortion theory[11,12] or relying on a qualitative match to human data[13,14,32]. Altogether, rate–distortion theory provides three principles for memory, which we detail here: prior knowledge, capacity limits and task dependency.

The most straightforward application of rate–distortion theory in the context of human memory concerns the influence of prior knowledge on recall. If a learned model of environmental regularities provides the basis for an efficient encoding of memory traces, it follows that domain experts have more accurate recall than do novices, as they have a more accurate model of the domain. However, this benefit for domain experts holds only for model-congruent stimuli. This pattern has been demonstrated in studies of memory for synthetic words[128] and chess configurations[11,129]. Using a learned generative model for compressing experiences accounts not just for varying accuracy of recall with expertise, but also the kinds of error introduced. The process of encoding a stimulus can be seen as interpreting it in terms of the internal variables of the generative model; this high-level interpretation constitutes the 'gist'. When specific details of an experience are discarded from the memory trace, they are generated from the latent representation of the model. 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 effect[96], in which recalling lists of semantically related words often leads to the recall of a strongly related but non-presented item with nearly the same probability as presented items. Approaches based on rate–distortion theory using variational autoencoders to learn a generative model of natural language have been used to show that the intrusion of non-presented items can be explained by reconstructing the word list from the latent representation of the model. The influence of gist-based distortions also implies that when the interpretations of ambiguous stimuli are manipulated (such as via contextual cues), both recall accuracy and the nature of distortions should be affected[97], and this effect can also be reproduced using rate–distortion theory[11,14].

Rate–distortion theory also naturally accounts for the effect of varying resource constraints. Theoretical analyses of memory suggest that the likelihood of a piece of information being needed decreases with time[130–132]. Consistent with this pattern, human forgetting curves seem to be adapted to such declining need probabilities[131,132], with recalled stimuli showing increased gist-based distortions as a function of delay before recall[133–135]. In rate–distortion theory, the amount of resources allocated to a memory trace corresponds to the targeted point on the rate–distortion curve (Fig. 1c), which also modulates the amount of model-congruent distortions in the reconstructed stimuli[11,12].

Last, rate–distortion theory also accounts for how memory is shaped by task demands and behavioural goals. Human memory and perception consistently show sensitivity to the cost of confusing stimuli within a given task[2,12,136–138]. For instance, in category-learning tasks, human memory becomes increasingly accurate for features that are relevant to the learned category, while accuracy for irrelevant features declines[12,136] Rate–distortion theory can incorporate these factors through the distortion function, for example, by overweighting errors related to danger or reward[12,136]. This degree of freedom in rate–distortion theory can also be exploited to optimize for the goal of prediction, which can be shown to imply a need for updating parameters rather than a precise reconstruction of stimuli[139,140].

information[17] from the encoding. This redundancy reduction enables compression even in the lossless case, inspiring the efficient coding hypothesis in neuroscience[19–21]. When resources are insufficient for perfect reconstruction, rate–distortion theory enables further compression by strategically discarding non-essential information and later trying to reconstruct it on the basis of known regularities. However, this reconstructive process often introduces distortions that align the recalled stimuli better with previously observed regularities.

Applying rate–distortion theory as a normative framework for human memory, previously observed regularities are a form of knowledge typically thought to belong to the domain of semantic memory. These regularities can be formalized as an internal generative model of the environment, enabling the interpretation and prediction of ongoing experience[14,22–24]. The process of compression using a generative model maintained in semantic memory introduces distortions in the encoding–decoding process — such as adding a monocle to the Monopoly Man (Fig. 1d). This explanation aligns with gist-based distortions and with early theories of memory distortions that attributed similar errors to the influence of pre-existing knowledge structures (memory schemas)[10]. Although classical compression algorithms produce qualitatively different memory distortions from human memory (Fig. 1e), advances in machine learning — particularly in applying deep generative models[25–27] to compression[28–30] — have enabled models based on rate–distortion theory to capture memory phenomena in complex domains, such as human drawings, text and natural images[11,12,31]. These findings have been used to demonstrate

that rate–distortion theory can be adapted as a unifying framework for parsimoniously explaining how prior knowledge (maintained in semantic memory) 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 been the subject of numerous proposals[34–40]. Work applying rate–distortion theory 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].

By describing how pre-existing knowledge affects the encoding and reconstruction of sensory experience, approaches based on rate–distortion theory successfully account for a wide variety of memory distortions and biases[11–14,31]. However, we argue that rate–distortion theory neglects a key challenge for memory: the need to learn and update an internal generative model on the basis of continually accumulating experiences.

In this Perspective, we propose an augmented framing of the computational problem of memory as iteratively learned compression, achieved through the combination of semantic and episodic memory systems. We suggest 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 the literature on curriculum sensitivity in human learning and contrast our predictions with predictions of the complementary learning systems theory[37,41], 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 theoretical and empirical results regarding memory prioritization 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.

The computational problem of memory

A fundamental problem with rate–distortion theory as a unifying explanation for human memory is that it assumes a known and unchanging set of environmental regularities, abstracted into an internal generative model. In reality, the brain must construct this generative model (semantic memory) over a lifetime, adapting it constantly in the light of new experiences. Notably, this assumption of known and unchanging regularities was later acknowledged as a limitation of the efficient coding framework by its originator, who observed that what is redundant today is not necessarily what was redundant yesterday[20]. The assumption also leads to misaligned predictions about human memory. Because 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. Therefore, 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].

To address this issue, we propose to consider two additional factors beyond how to efficiently compress experiences under a known generative model. First, the internal generative model needs to be learned. Second, the learning of the generative model must proceed in an online, iterative manner, in which 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, because while it is optimizing 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 example. Imagine learning how to brew good coffee with an unfamiliar machine (such as a stovetop moka pot) using trial and error, by figuring out how different variables influence the taste of the coffee. Each 'episode' of brewing a cup (Fig. 2a) involves both relevant variables (such as the bean type or quantity) and irrelevant variables (such as the weather or background music). In this situation, a generative model of coffee brewing could be created by observing how the relevant input variables influence the taste of the coffee and capturing these relations in the parameters of a generative model. According to normative theories of learning, a generative model of coffee brewing can be acquired without specifically remembering any individual episodes. Rather, all relevant information can be captured by iteratively updating the parameters of the generative model and discarding the raw experiences (Fig. 2b).

Imagine that after many brewing episodes, you have figured out a configuration of variables that consistently produces tasty coffee. Yet today, it tastes inexplicably terrible. If you had direct access to all past episodes, you could readily determine the cause: although 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 the water was heated to a boil. However, because initial water temperature was previously considered irrelevant, its value in past episodes has been discarded. Therefore, you are surprised and have no clear indication of what went wrong or how to adjust for the next attempt.

A key property behind the failure of the learning process outlined above is that beyond parameter estimation (refining a known parametric relationship between known variables), it also features an additional problem of structure learning[47,48] (Box 2). Structure learning involves identifying causal variables in a given environment (for instance, the bean type, weather or background music) and how they affect each other (for instance, weather might affect mood, but not bean type) (Fig. 2c). In terms of compression, a known model structure enables highly efficient use of memory resources by only encoding information relevant to the parameters and often enables them to be estimated online. By contrast, for structure learning, online updates require each possible hypothesis to be tracked and updated in parallel. Online tracking and updating quickly becomes impractical, as even for a toy problem with just four variables, there are 543 possible hypotheses about the causal structure (and with a single additional variable it becomes 29,000). Such a combinatorial explosion of the hypothesis space is typical for structure-learning problems. This proliferation of structural hypotheses means that maintaining the relevant information for each candidate structure is as challenging as storing all past episodes directly.

On the one hand, limited human memory resources require experiences to be stored in a compressed format, which is supported by a learned generative model of the environment. On the other hand, learning and maintaining a generative model requires access to details of previous episodes that might not have been considered relevant under the current model structure. Thus, there must be a balance between a combinatorial explosion of structural hypotheses (considering all possible model structures) and the risk of discarding key information (considering interpretations under only a single structure) (Fig. 2d).

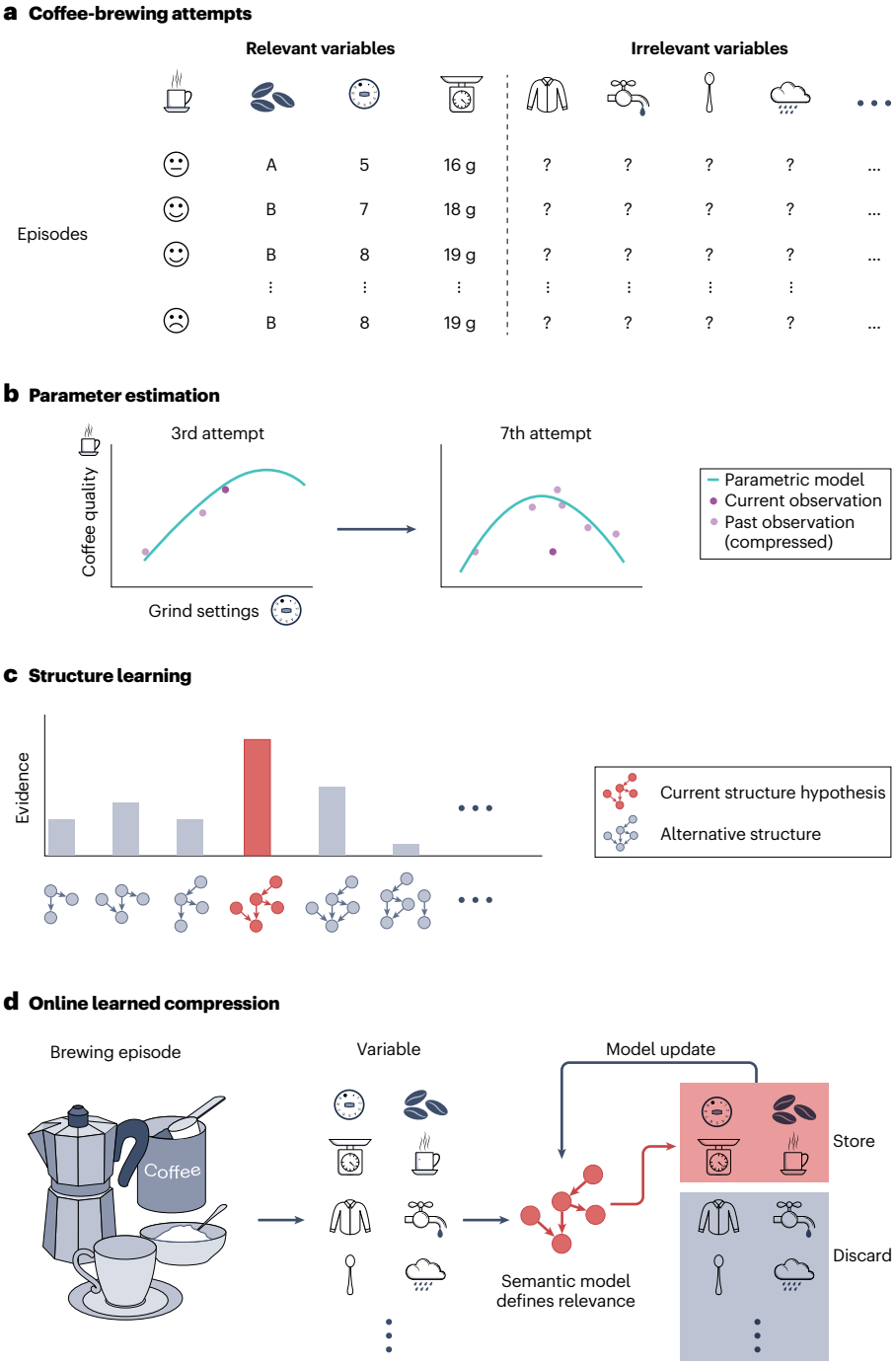


Fig. 2 | Online structure learning in the coffee-brewing example. a, Each episode consists of values for relevant (left) and irrelevant (right) variables. In online learning, the information content of irrelevant variables is discarded, preventing learning about the regularities governed by them. **b**, Online learning involves parameter adjustments, such as how grind setting influences coffee quality. Previous data points (light purple) are discarded, with relevant information summarized in the model's parameters (green). For simplicity, quality is shown as only a function of grind setting. The parameters summarize an increasingly large number of observations (compare the 3rd and 7th attempts); at each point in time, only the parameters and the current observation (dark purple) are available for making decisions. **c**, In structure learning, qualitatively different hypotheses are evaluated using a measure of congruence with observations (model evidence). Each hypothesis determines the set of relevant variables. The hypothesis with the highest evidence (red bar) is selected as the basis for relevance judgements, while alternative structures (grey bars) are not tracked. **d**, During online model updates, only the information from relevant variables is retained, whereas irrelevant information is discarded. This process carries the risk of misclassifying a relevant variable as irrelevant, potentially leading to the systematic loss of information that is essential for future model updates.

We propose that the brain uses an approximation that relies on the combination of two interlinked memory systems (Fig. 3). Semantic memory builds a generative model that captures environmental regularities and facilitates compression. Thanks to computational and memory constraints, semantic memory tracks only a single working hypothesis about the overarching causal structure of the environment. Although our argument can be extended to cases in which multiple structural hypotheses are tracked within restricted local domains, for simplicity we assume a single hypothesis. We refer to the generative model based on this single structural hypothesis, stored in semantic memory, as the semantic model. However, restricting the set of tracked hypotheses risks being stuck in a dead end, where information that is necessary for further improvement of the semantic model has already been selectively discarded (as in the coffee example above). Thus, episodic memory retains a relatively raw and uncompressed encoding of novel and surprising episodes (episodes most likely to be

Box 2 | Structure learning and Neurath's ship

Structure learning refers to a class of learning problems in which competing models differ not only in the precise numerical values of model parameters (parameter estimation), but also in the number of parameters, choice of variables, forms of relationships, or even the fundamental building blocks used to specify the model. Normative theories of structure learning often decompose learning problems into determining the high-level structure of the model (the structural hypothesis) and fine-tuning the parameters while keeping the structure fixed[141,142]. Some approaches further distinguish between structure and form, where form refers to the general class of graph (such as tree or grid) and structure defines the exact set of edges and nodes. A change in structure means a local modification (such as adding or removing an edge), whereas a transition in the form of the model is a rare but fundamental shift, such as a child deciding to organize animal species into a tree structure rather than separate clusters[47]. 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, structure-learning problems are often defined by specifying primitive components along with rules for their composition (for instance, causal graphs are constructed from nodes and directed edges). These composition rules are typically open-ended, enabling arbitrarily complex structures to be 'grown' over the course of learning[143]. Although compositionality enables such models to construct genuinely novel explanations, it also results in inconceivably vast hypothesis spaces[144,145]. Second, navigating hypothesis spaces is quite difficult. To illustrate, imagine a 'learning landscape', with the horizon spanned by possible configurations of the model, and the height of the terrain defined by the goodness of fit for that configuration (as in Fig. 3c). 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 can sometimes correspond to dramatically different predictions, making the terrain rugged and treacherous[113].

The Neurath's ship analogy was originally proposed in the context of difficulties encountered in revising scientific theories. However, these difficulties mirror those encountered during learning, making the Neurath's ship analogy applicable to this context as well[53]. The iterative rebuilding of Neurath's ship can be precisely formalized 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, summarizing them in the posterior distribution. Ideally, hierarchical Bayesian inference prescribes computing the posterior distribution over all structural hypotheses, updating them in parallel with incoming observations. In practice, Monte Carlo approximations are usually used, in which on the highest levels of the hierarchy, only a restricted set of hypotheses are tracked (or even just a single one). Maintaining the posteriors over parameters for this restricted set is much less resource-intensive than maintaining the full set. In this class of Monte Carlo algorithms, the process for updating the model structure is encoded in the proposal distribution, which specifies what alternative hypotheses can be considered in a single update for each hypothesis. Following Neurath's ship, this proposal distribution favours local changes, for example, allowing the addition or removal of only a single causal edge.

Causal learning is a prototypical example of structure learning[53]. However, structure-learning problems are ubiquitous in natural environments, encompassing contextual learning[50], the identification of underlying structural forms within data[47], and learning visual[111,146,147] or abstract concepts[61,64,110,112,113,145,148]. Structure learning has also been implicated in event segmentation[101–105], in which the temporal structure of visual or auditory information stream needs to be discovered[84,85,149]. At the most general level, the composable building blocks for theories can define components of a programming language, making learning akin to program induction[110–113].

misinterpreted under the current hypothesis), offering some insurance against an incorrect structural hypothesis.

This proposed integration of semantic and episodic memory solves the dual theoretical problems of learning to remember (building a semantic model that enables compression) and remembering to learn (storing relevant episodes for future model updates). It also explains a range of empirical findings about human behaviour, including memory distortions and curriculum effects in learning.

Learning to remember

The efficient allocation of memory resources necessitates the construction and continual updating of a generative model of the environment based on observations, which we posit to be the role of semantic memory. Owing to the unknown causal structure of the environment, the hypothesis space available for such a model is vast and difficult to navigate, making approximations necessary. An often-used method for approximate structure learning is to track a selected set of hypotheses instead of the full distribution, known as Monte Carlo sampling[49–52]. Converging evidence from multiple learning paradigms suggests that the brain might also be limited to tracking a restricted set or even a single structural hypothesis[49,53–55]. A proposal has likened this learning process to the metaphor of Neurath's ship[53], originally introduced in the philosophy of science[56,57] to illustrate the gradual and continuous development of scientific theories. The Neurath's ship metaphor likens theorists to sailors attempting to reconstruct a ship while sailing on it, gradually replacing pieces of the ship but never wholly starting afresh (given that then the ship would sink).

Applied to the brain, the ship in the metaphor represents an individual's evolving understanding of the structure of the world in their semantic memory. In our proposal, semantic memory maintains a dynamic and evolving structural hypothesis that informs perception and decision-making. 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. Local changes to the model structure can therefore be made without compromising its ability to function effectively within its environment (Fig. 4a).

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, a participant 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[46,58,59], 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 can compromise the interpretation of future observations, leading to erroneous parameter updates. Second, even if the participant realises that their hypothesis is flawed, alternatives are evaluated on the basis of past data, which were compressed on the basis of an incorrect hypothesis. Consequently, supporting evidence for the correct structure might 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 a distinct sensitivity to the order in which stimuli are encountered, generally referred to as curriculum effects. These effects include primacy effects in which stimuli experienced early in an experiment determine the influence of what is experienced later[39,60,61] Such effects have been observed in human behaviour using tasks such as reward learning[62,63] and causal learning[64,65].

These human learning patterns diverge from the learning dynamics that are characteristic of alternative accounts of semantic learning using artificial neural networks. Artificial neural networks also display robust curriculum effects, but often in an opposite pattern to what humans tend to exhibit. For instance, in one study humans and artificial neural networks were given the same context-dependent decision-making task[62] (Fig. 4b). When artificial neural networks were presented with different tasks or learning contexts in a blocked manner, the different blocks tended to overwrite one another and result in poor performance, a well studied phenomenon known as 'catastrophic forgetting'[66,67]. But when the training data were interleaved, artificial neural networks could learn reliably[68]. In contrast to artificial neural networks, humans performed better in blocked settings and were hindered by interleaved curricula, with a stronger effect as the complexity of the task was increased (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[63,65,69]. We note that some studies on human learning have also found benefits to interleaved curricula in different settings, especially when generalization between stimuli was beneficial to the performance of the task[70], or in discrimination tasks where the immediate juxtaposition of exemplars from different classes seems to highlight the differences between them[71–73].

Complementary learning systems 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. Complementary learning systems theory views the acquisition of the semantic model as the gradual updating of an artificial neural network, integrating information across multiple experiences over time. According to this view, the utility of episodic memory is that interleaving older episodes with current observations protects older knowledge from being overwritten.

From the perspective of online structure learning, blocked data is ideal because consecutive trials from a single context enable the learner to focus on a subset of the complete structure[39,61,65], 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, enabling the learner to fine-tune the parameters. This situation is similar to the coffee example described above. By contrast, with 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 situation is that certain learners might fail to retain the context accurately or arrive at an overly simplified structure that merges the contexts together. In complementary learning systems, semantic memory relies on interleaved training, with episodic memory mitigating the adverse effects of blocking. By contrast, our approach suggests that semantic memory learns most effectively under blocked training, whereas episodic memory is essential for counteracting the failure modes caused by interleaved training.

A more direct connection between curriculum dependence in human learning and the problem of structure discovery was established

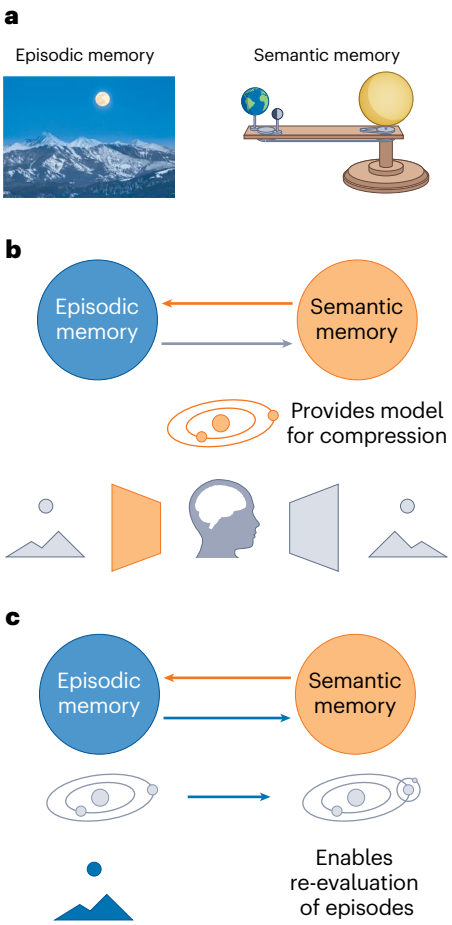


Fig. 3 | Interactions between semantic and episodic memory. a, We conceptualize episodic memory as retaining traces of individual experiences, and semantic memory as a simplified internal model of the environment, formalized as a probabilistic generative model over experiences. **b**, Semantic memory facilitates the efficient compression of episodic memories by providing the statistical model required for compression. **c**, The iterative refinement of the model through learning critically relies on encoding surprising and novel episodes in a less compressed format in episodic memory, to preserve them for later re-evaluation if the current model proves to be inaccurate or incomplete. Panel **a** (mountain photo) credit: john lambing/Alamy Stock Photo.

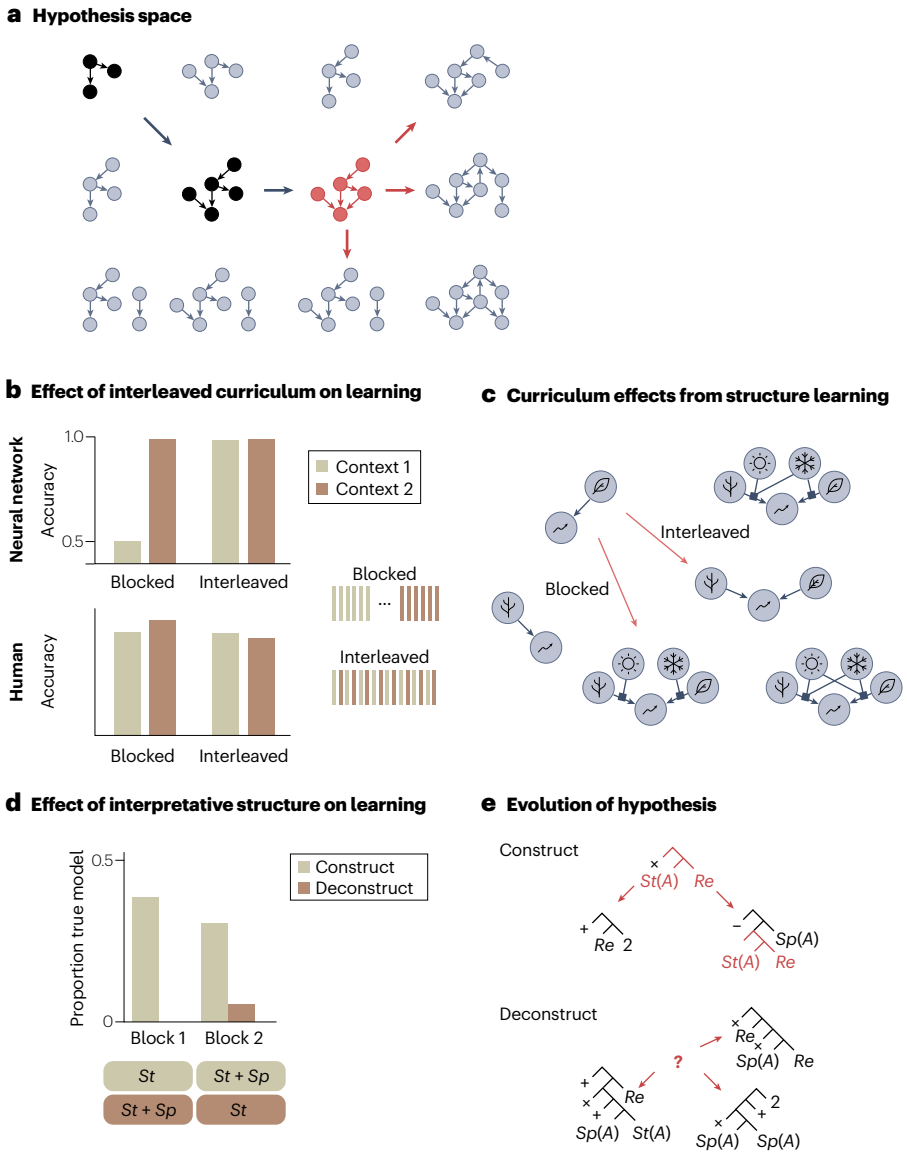


Fig. 4 | The Neurath's ship analogy for human learning. a, The evolution of structural hypotheses during learning: the structure estimate is not discarded entirely but can be altered while moving through hypothesis space. The current hypothesis (red) is progressively altered relative to previous hypotheses (black). **b**, Neural networks suffer from catastrophic interference under blocked curricula (top), whereas humans do not (bottom). However, humans tend to show better performance under blocked compared to interleaved curricula. Presenting fewer stimulus features in Block 1 and then the full set in Block 2 ('Construct', light brown; panel **d**) results in higher accuracy than presenting the full feature set in Block 1 and a simplified set in Block 2 ('Deconstruct', dark brown; panel **d**). **c**, Possible hypothesis evolution during learning for learners under blocked versus interleaved curricula (red arrows) from panel **b**. Nodes depict the relevant variables given a particular hypothesis. The difficulty of the structure inference task in earlier blocks determines whether the correct structure is successfully discovered in later blocks for human learners. **d**, Humans learn better when they can learn a useful structure early. In this experiment, number of stripes (*St*) and number of spots (*Sp*) influence the length of a 'stick', represented by a stack of rectangles (*Re*). **e**, A learner in the 'Construct' curriculum who has successfully identified the primitive structure ($St \times Re$) (number of stripes on 'magic egg' A multiplies the number of rectangles) in Block I can use this primitive to constrain the search problem in Block II (for example, eliminate $Re + 2$) and discover the correct full structure ($St \times Re - Sp$) (top). A learner in the 'Deconstruct' curriculum, who failed to identify a useful structure for retaining the relevant information from Block I, and in Block II the discovery of the ($St \times Re$) primitive comes too late (bottom). Data plotted in part **b** are adapted from ref. 62, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **d** adapted from ref. 64, Springer Nature Limited.

in a study that used blocked curricula and manipulated the contents of the blocks[64]. Participants learned a causal relationship between the features of a 'magic egg' and the length of a 'stick' (Fig. 4c). Under the 'construct' curriculum, the first block contained only examples with one feature present (either stripes or spots), whereas the second block introduced the second feature. By contrast, the 'deconstruct' curriculum presented both features in the first block (a more challenging structure inference problem). Participants were allowed to revisit previous examples within the same block, mitigating the demands on their memory. More participants discovered the correct structure — here conceptualized as a program[74,75] — under the construct curriculum than under the deconstruct curriculum. In the construct curriculum, a correct partial rule (incorporating only one feature) could be easily inferred from the first block and then extended to incorporate the second feature. By contrast, the reversed order in the deconstruct curriculum made the initial hypothesis space much more complex. Although the simpler second block allowed a substantial proportion of participants to identify the correct structural primitive (the rule that the number of stripes has a multiplicative effect on the length of the stick), 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 owing to the lack of a suitable interpretative structure.

In summary, we argue 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 propose that the brain relies on an approximate solution to this problem of online structure learning, which leads to a characteristic path-dependence in learning: 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. 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 seemingly irrelevant aspects of experience to be retained in episodic memory in order to evaluate alternative hypotheses. Next, we focus on this latter component of our proposed framework, exploring how episodic memory can support the acquisition of semantic knowledge.

Remembering to learn

In this section, we explore the use of memory to create rich reconstructions of prior experience, including idiosyncratic and potentially irrelevant details. Our framework is based on the insight that to mitigate the failure modes of semantic learning that occur when the current structural hypothesis is flawed, the only generally applicable method is to retain information that might seem irrelevant in the context of the current hypothesis but is relevant for evaluating potential alternatives. Thus, the ability to remember is crucial to ensure that future learning remains possible.

In our coffee-making example, 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 that episode were encoded in episodic memory, it might include details that are irrelevant in the context of the current structural hypothesis, such as the water being drawn from a cold tap. A subsequent episode in which the water comes from a preheated kettle and produces a much better-tasting result could retrospectively reveal that water temperature is a relevant contextual variable. More generally, we posit that learners preserve experiences in a detail-rich and relatively unrefined format, which enables them to make greater overhauls to model structure. Extending the Neurath's ship metaphor from the last section, we refer to this aspect of memory as an 'episodic life raft' (Fig. 5a).

The results of a simple category-learning task demonstrate the value of an episodic life raft[39]. In this study, artificial learners had to iteratively learn categories through sequential observations, akin to discovering the species of unknown animals on the basis solely of their visible features (unsupervised clustering, Fig. 5a). In this context, structure learning requires the learner to determine the number of categories, whereas parameter estimation requires the feature distributions to be refined for each category (Fig. 5b). The study found that a semantic-only learner, who uses the Neurath's ship approximation to iteratively update the structure estimate and parameters of its semantic model but retains no explicit representation of individual episodes, often failed at the structure-learning task. Specifically, the semantic-only learner tended to systematically underestimate the number of categories, unless the observations were carefully ordered (using a blocked curricula). However, a learner with semantic and episodic memory, even with severely limited episodic capacity, had greatly improved structure learning (Fig. 5c). The learner's episodic memory stored a small subset of past experiences that could be replayed when considering alternative structures, augmenting the information contained in the parameters of the current hypothesis. This replay process can be viewed as an analogue to replay for memory consolidation[76], in which episodic memories are integrated with the knowledge maintained in semantic memory.

Although an episodic life raft is clearly advantageous for structure learning, episodic memory is a costly memory format. The cost of episodic memory stems from the very source of its utility: retaining idiosyncratic details is expensive. This trade-off between the cost of memory and the benefit to learning raises the question of where to allocate scarce resources and which episodes to prioritize remembering. To answer this question, we first consider a simplified context in which the only storage limitation is on the number of episodes that can be stored, but each episode can be recalled perfectly.

Consistent with the reasoning that episodic memory needs to be applied selectively because of its cost, simulated learners with limited episodic memory capacity performed better in online structure learning when they selectively prioritized experiences with high Bayesian surprise[77–79], compared to learners that applied the same capacity indiscriminately[39] (Fig. 5c). This approach is similar to earlier proposals in category learning that added exceptions to general rules, sharing the need to store anomalous events[80,81]. An advantage of Bayesian surprise is that it distinguishes between mere noise and surprise that warrants model change[77,82], although alternative formalizations of surprise or novelty could be explored in the future[83]. Surprise can also signal environmental change and underlie the detection of new types of event, thus contributing to event segmentation[84,85].

The idea that incongruent and novel information is selectively prioritized in memory has a long history in psychology[45] and is supported by extensive empirical findings[42–44,46,86]. Similarly, in neuroscience, the idea that the hippocampal formation (associated with episodic memory) retains novel information has been extensively explored[46], particularly in the context of experience replay and memory consolidation[38,41,87,88].

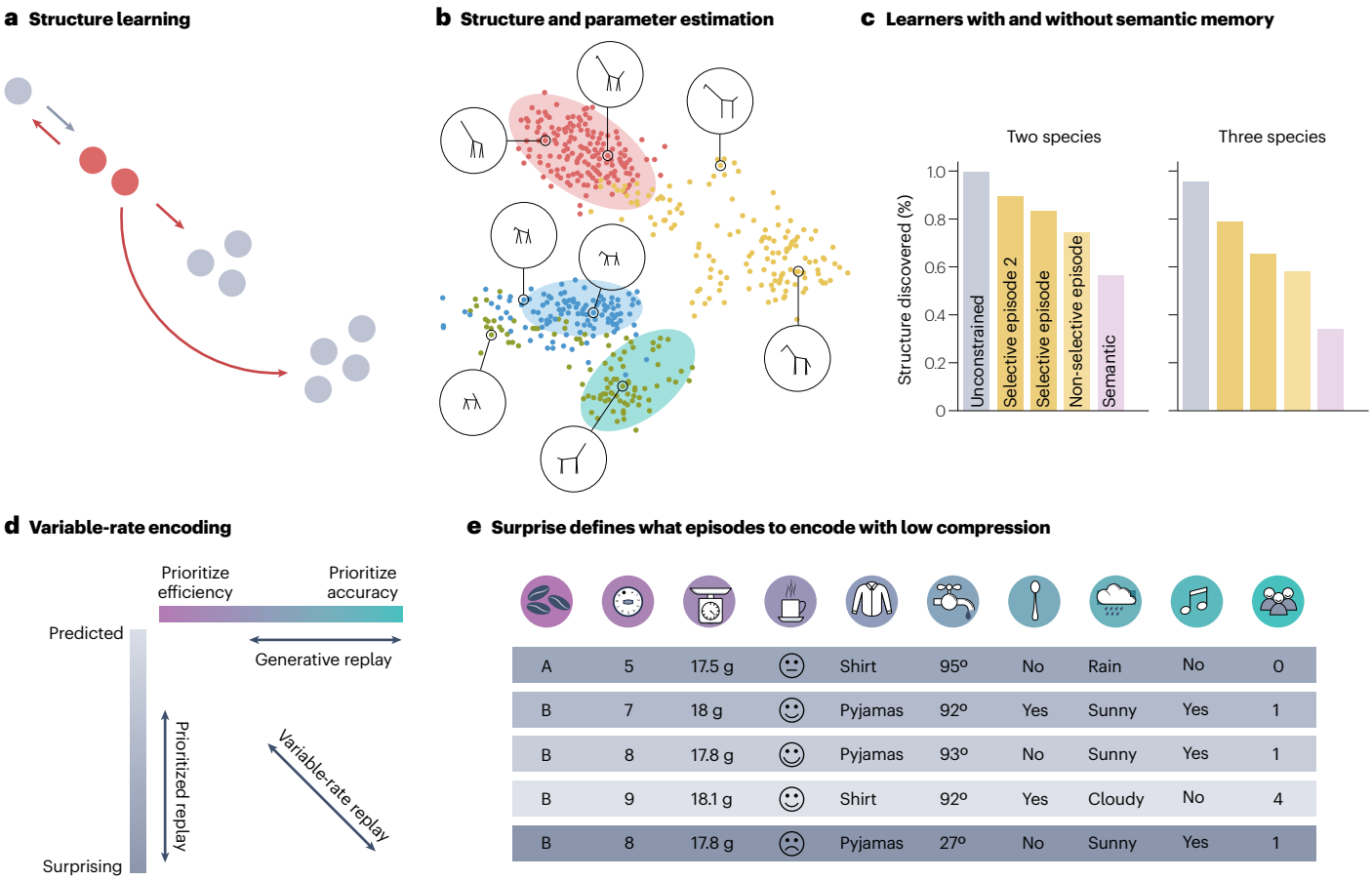
The question of how episodes should be prioritized is typically referred to as prioritized replay[41,88,89]. However, because constraints on memory resources are not a primary concern in these approaches, prioritization refers not to which episodes should be retained, but to how frequently they should be replayed. (We note that in the limit, lowering the probability of replaying an episode corresponds to discarding it.) Prioritizing episodes according to their associated reward prediction error has been instrumental in machine learning advances such as achieving human-level performance in Atari games via reinforcement learning[89–91]. Consistent with the idea of prioritizing novel information, it has been argued that the utility of retaining episodes in the reinforcement learning setting is greatest in the early stages of encountering a novel environment, before a sufficiently accurate semantic model can be established[35,92,93]. However, this argument emphasizes the direct utility of episodes for decision-making, instead of their potential role in supporting the construction of a semantic model.

We now turn to more realistic scenarios in which memory resources are more limited and episodes cannot be recalled perfectly. The rate–distortion theory perspective suggests that memory resources can be decreased if episodes are compressed lossily, with missing details filled in by a generative model that is maintained in semantic memory (Box 1). Generative replay — replaying what are essentially compressed episodes during the training of artificial neural networks — has been shown to place much lower demands on memory resources compared to exact replay (replaying episodes without any loss of detail) while still protecting against catastrophic forgetting[94,95].

Although compressing episodes using semantic memory enables substantial savings in memory resources relative to the verbatim storage required for exact replay[95], it seems 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, it is unclear how the current model can be relied

upon to compress these experiences. A key insight of rate–distortion theory might 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. Rather than a binary choice between storing an exact copy of an event in episodic memory or merely updating the model's parameters, this structure suggests the possibility of a variable-rate encoding, with a continuum of choices regarding the desired fidelity of the reconstruction.

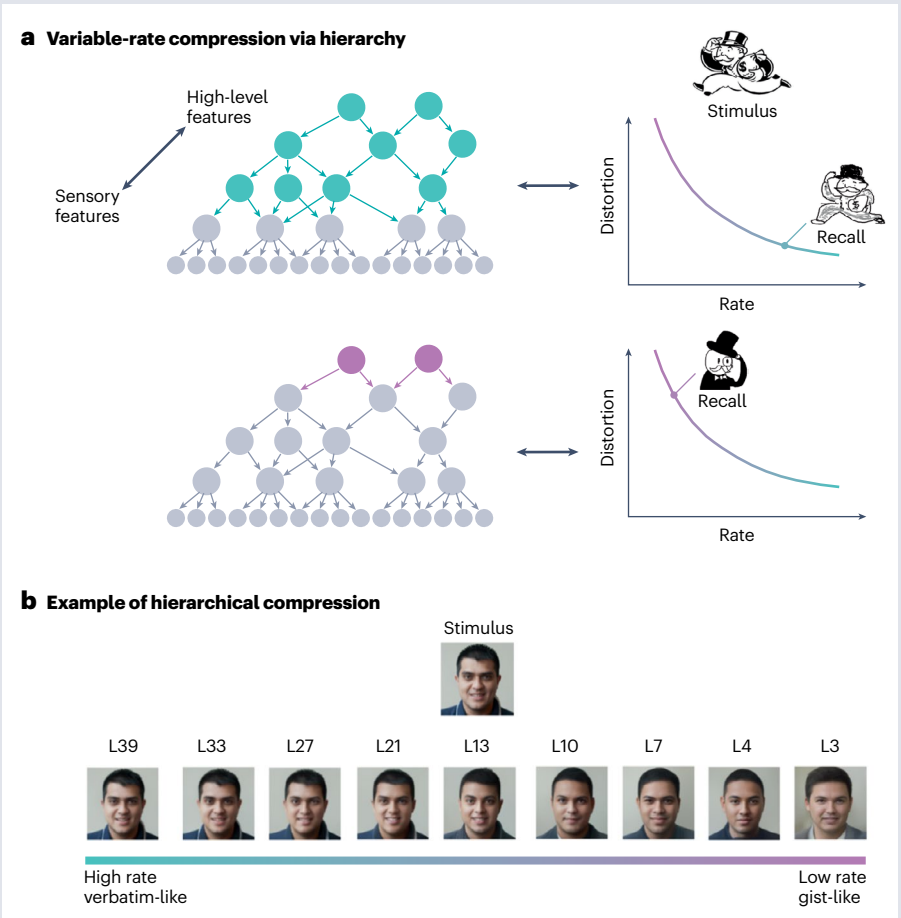
We propose that variable-rate encoding of sensory experience underlies the brain's ability to balance the competing goals of conserving resources and maintaining robustness to novelty (Box 3). Specifically, we suggest that the rate of encoding — or equivalently, the desired accuracy of recall — should 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 and surprising ones would be less compressed. The high compression of well predicted episodes would make them prone to increasing degrees of gist-like distortion, as demonstrated in previous work on human memory[11,12]. However, when semantic memory has lower confidence in its interpretation, owing either to violated predictions or novel situations, additional memory resources could be allocated to encode an episode with less compression and therefore result in more accurate recall of episodic details. This mechanism could prioritize



A	5	17.5 g	😐	Shirt	95°	No	Rain	No	0
B	7	18 g	🙂	Pyjamas	92°	Yes	Sunny	Yes	1
B	8	17.8 g	🙂	Pyjamas	93°	No	Sunny	Yes	1
B	9	18.1 g	🙂	Shirt	92°	Yes	Cloudy	No	4
B	8	17.8 g	☹	Pyjamas	27°	No	Sunny	Yes	1

Fig. 5 | The episodic 'life raft'. a, Structure learning involves discovering the large-scale organization of data, such as the number of classes. Multiple hypotheses about the number of classes can be compatible with the data but the space of possible hypotheses needs to be navigated on the basis of a single hypothesis at a time. **b**, In a category-learning task, the number of classes and their properties are inferred from data, in this case a series of stick figure animals. The segment lengths and angles of the figures cluster by species (coloured ovals) but individual animals (dots) vary considerably within each cluster. **c**, Structure-learning performance of different agent types highlights the value of the episodic life raft. An unconstrained learner (grey) tracks all hypotheses in parallel, setting an upper benchmark. The semantic learner (pink), following the Neurath's ship heuristic, commits to a single best hypothesis, leading to diminished performance. The episodic learner can mitigate the gap between the semantic and the unconstrained learner by augmenting it with an episodic memory that retains raw past experiences either selectively (dark yellow) or non-selectively (light yellow) according to the degree of surprise under the current hypothesis. **d**, Variable-rate encoding entails increasing the resources dedicated to encoding surprising experiences to make recall more accurate for these episodes than for predicted episodes. Using surprise to define the rate of encoding couples the priority of an episode to the level of detail generated by the model. **e**, Variable-rate encoding for the coffee example: prioritized replay modulates which episodes (rows) are stored and replayed, whereas generative replay modulates which variables (columns) are encoded. Variable-rate encoding combines these approaches, enabling more variable values to be accurately recovered from surprising than from predicted episodes. Part **b** adapted with permission from ref. 125, Elsevier. Part **c** is adapted from ref. 39, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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. 3b,c). Sensory experience is compressed via semantic memory, producing episodic memory traces with model-congruent distortions. This compression increases the effective capacity of episodic memory, but might also hinder its proposed role in retaining seemingly irrelevant details that are 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 that the compression model might be unreliable.

Two key challenges stand out when we attempt to bridge the computations that implement variable-rate compression with a neural implementation in the brain. First, although rate–distortion theory enables 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 requirement for separate encoder–decoder pairs implies the maintenance of multiple semantic models in parallel, contradicting the principle behind Neurath's ship. Second, rate–distortion theory provides no mechanism for converting from detailed to more-compressed memory traces after the initial encoding (varying resource constraints; Box 1).

Multiple machine learning approaches have been proposed to achieve variable-rate compression while avoiding multiple encoder–decoder pairs[150–152]. A promising idea for how the brain might implement variable-rate compression is through a hierarchical generative model[153–156], with layers deeper in hierarchy corresponding to progressively stronger levels of compression (figure panel **a**). Similar to these models, the human visual cortex is thought to represent sensory information hierarchically: early layers respond to basic features such as edges, intermediate layers detect textures and deeper layers integrate these features into representations of objects and scenes[157–159]. Hierarchical models offer a straightforward way to reduce memory capacity by selectively discarding information over time, starting with lower-level details (figure panel **b**). 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[160]. 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[14,32,161,162], and behavioural evidence suggests that this framework might extend to other modalities as well[163].

Box 3 figure part **a** is adapted with permission from ref. 5, Sage. Box 3 figure part **b** is adapted from the Flickr-Faces-HQ Dataset (FFHQ)-256 dataset (https://github.com/NVlabs/ffhq-dataset)[164], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

information in episodes in a graded fashion and explain why surprising experiences are often recalled with more episodic detail than expected ones[42–44,46,78,86].

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.

Conclusion

Human memory is prone to systematic biases and distortions, which are widely interpreted as reflections of both its adaptive function and

inherent resource constraints; memory is shaped by the need to interpret and learn from experience. However, this setup raises the dual challenge of retaining useful information while learning how to decide what is useful. We have argued that, in its current form, rate–distortion theory falls short as a normative framework for human memory. Although approaches based on rate–distortion theory offer a normative solution for the first piece of the puzzle — 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 have highlighte how this omission results in qualitative discrepancies from the empirical phenomena of human memory and argued that addressing these phenomena requires us to rethink the fundamental assumptions of rate–distortion theory.

Thus, we have proposed a revised normative framework in which semantic memory tracks a limited approximation of the environmental structure. In this framework, the online construction and updating of the semantic model is analogous to how Neurath's ship is rebuilt at sea. However, because interpreting observations under a single structural hypothesis can result in systematic loss of essential information, we argue for recruiting additional episodic memory resources to encode novel and surprising observations in a relatively uncompressed format. Thus, episodic memory serves as a life raft while rebuilding the ship of semantic memory.

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 insights into several ongoing challenges in the field. By accounting for the role of episodic memory in safeguarding against learning the wrong semantic model, we arrive at a normative explanation for why surprising stimuli are often remembered with higher fidelity than familiar or expected stimuli. By contrast, our framework predicts typical rate–distortion-theory-like distortions for experiences that are congruent with the current structural hypothesis.

A key focus of our Perspective covers the consequences of an evolving semantic model for memory distortions. The standard rate–distortion-theory approach has served as a unifying framework for classical gist-based memory distortions, such as intrusions of semantically related items[96] or label-consistent distortions in memory[97] 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 — which is the defining property of curriculum effects. Conversely, updating the model after a new experience (such as in post-event misinformation[98] or hindsight bias[99]) changes the decoder and therefore alters how past experiences are reconstructed[23], potentially also explaining associative memory errors[100].

In the study of both human and machine learning, stimuli are typically presented in randomized fashion, with simple or non-existent dependencies between trials. This presentation order has clear benefits for eliminating experimental confounds. However, it stands in stark contrast with the rich, multi-scale sequential structure that characterizes natural environments. Here we have focused on coarse-grained structure, neglecting the temporal breadth of episodes, and consequently the issue of how to segment continuous sensory inputs[84,85,101–105]. This fine-grained temporal structure and its interactions with structure learning are likely to be important in a more nuanced understanding of curriculum effects[106], and an integration of these approaches might explain a larger variety of curriculum effects[71–73] under a unified framework. A more refined understanding of path-dependencies in learning — also crucial for educational applications[107–109] — will require improved theories about how semantic knowledge is represented and organized. One intriguing approach is to view semantic memory as a library of concepts, often formalized in a program-induction framework[110–112], in which a goal of curriculum design is to induce widely applicable and composable conceptual modules that further learning can build on[61,64,113].

Although it is absent from existing rate–distortion theory accounts, emotional salience is empirically one of the strongest factors influencing memory[114–117]. Our framework offers two promising ways in which to consider this factor. First, emotionally relevant aspects of experience could be prioritized by the rate–distortion-theory distortion function. This prioritization aligns with how rewards have been integrated with generative models in reinforcement-learning contexts[118,119], with emotions mediating reward-related computations[120]. Second, we propose that novelty and surprise determine resource allocation in variable-rate encoding. Because emotional salience is indicative of whether the episode is expected to be retrieved in the future[121], high salience implies an increased rate of encoding. Accordingly, traumatic experiences might be understood as an extreme case of encoding primarily uninterpreted sensory features, which aligns with qualitative properties of post-traumatic stress disorder[122,123]. More broadly, the combination of Neurath's ship with the episodic life raft might prove fertile ground for a deeper, computational understanding of traumatic events in memory and their long-term effects on development.

Although 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]. Rate–distortion theory helps to illuminate some of these differences[12], but additional computational considerations — such as trade-offs in computational cost[35,61] and the path-dependent co-evolution of these systems[124] — are also likely to have a crucial role.

Published online: 05 June 2025

References

1. Baddeley, A., Eysenck, M. W. & Anderson, M. C. *Memory* (Routledge, 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. Cogn.* **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.* **76**, 275–302 (2025).
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* (Cambridge Univ. Press, 1932).
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. et al. A model of semantic completion in generative episodic memory. *Neural Comput.* **34**, 1841–1870 (2022).
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. in *Claude E. Shannon: Collected Papers* (eds Sloane, N. J. A. & Wyner, A. D.) vol. 7, 325–350 (Wiley/IEEE Press, 1993).
19. Barlow, H. B. in *Sensory Communication* (ed. Rosenblith, W. A.) 217–233 (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 Univ. Press, 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. In *Intl Conf. Learning Representations* https://openreview.net/forum?id=33X9fd2-9FyZd (2014).
26. Higgins, I. et al. Beta-VAE: learning basic visual concepts with a constrained variational framework. In *Int. Conf. Learning Representations* https://openreview.net/forum?id=Sy2fzU9gl (2017).
27. Rezende D. J., Mohamed S. & Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. 31st Intl Conf. Machine Learning, PMLR* **32**, 1278–1286 (2014).
28. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep variational information bottleneck. In *Intl Conf. Learning Representations 2017* https://openreview.net/forum?id=HyxQzBceg (2017).
29. Alemi, A. et al. Fixing a broken ELBO. In *Proc. 35th Intl Conf. Machine Learning* (eds Dy, J. & Krause, A.) **80**, 159–168 (2018).
30. Ballé, J., Laparra, V. & Simoncelli, E. P. End-to-end optimized image compression. In *Intl Conf. Learning Representations 2017* https://openreview.net/forum?id=rJxdQ3jeg (2017).
31. Bates, C. J., Alvarez, G. A. & Gershman, S. J. Scaling models of visual working memory to natural images. *Commun. Psychol.* **2**, 3 (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. *Phil. Trans. R. Soc. Lond. B* **379**, 20230395 (2024).
34. Nicholas, J. & Mattar, M. G. Humans use episodic memory to access features of past experience for flexible decision making. In *Proc. 46th Ann. Conf. Cognitive Science Society* (eds. Samuelson K. et al.) 2754–2760 (Cognitive Science Society, 2024).
35. Lengyel, M. & Dayan, P. Hippocampal contributions to control: the third way. In *Advances in Neural Information Processing Systems 20* (eds. Platt, J., Koller, D., Singer, Z. & Roweis S.) 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. *Can. J. Exp. Psychol.* **62**, 62–79 (2008).
39. Nagy, D. G. & Orban, G. Episodic memory as a prerequisite for online updates of model structure. In *Proc. 38th Ann. Conf. Cognitive Science Society* (eds. Papafragou, A. et al.) 2699–2704 (Cognitive Science Society, 2016).
40. Lu, Q., Hummos, A. & Norman, K. A. Episodic memory supports the acquisition of structured task representations. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.06.592749 (2024).
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**, e61077 (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. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl. Acad. Sci. USA.* **105**, 10687–10692 (2008).
48. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
49. 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).
50. Heald, J. B., Lengyel, M. & Wolpert, D. M. Contextual inference in learning and memory. *Trends Cogn. Sci.* **27**, 43–64 (2023).
51. Dasgupta, I., Schulz, E. & Gershman, S. J. Where do hypotheses come from? *Cogn. Psychol.* **96**, 1–25 (2017).
52. Sanborn, A. N. & Chater, N. Bayesian brains without probabilities. *Trends Cogn. Sci.* **20**, 883–893 (2016).
53. 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).
54. 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).
55. Courville, A. C. & Daw, N. The rat as particle filter. In *Advances in Neural Information Processing Systems 20* (eds. Platt, J., Koller, D., Singer, Z., & Roweis S.) 369–376 (2007).
56. Neurath, O. *Empiricism and Sociology* (eds Neurath, M. & Cohen, R. S.) Vienna Circle Collection Vol. 1 (D. Reidel Publishing, 1973).
57. Van Orman Quine, W. *Word and Object* (MIT Press, 1960).
58. Tse, D. et al. Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
59. Tse, D. et al. Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
60. Abbott, J. T. & Thomas, L. Exploring the influence of particle filter parameters on order effects in causal learning. In *Proc. 33rd Ann. Conf. Cognitive Science Society* (eds. Carlson, L., Hoelscher, C. & Shipley, T. F.) 2950–2955 (Cognitive Science Society, 2011).
61. Zhou, H., Nagy, D. G. & Wu, C. M. Harmonizing program induction with rate–distortion theory. In *Proc. 46th Ann. Conf. Cognitive Science Society* (eds Frank, S. L., Toneva, M., Mackey, A. & Hazeltine, E.) 2511–2518 (Cognitive Science Society, 2024).
62. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc. Natl Acad. Sci. USA* **115**, E10313–E10322 (2018).
63. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. *Proc. Natl Acad. Sci. USA* **119**, e2205582119 (2022).
64. Zhao, B., Lucas, C. G. & Bramley, N. R. A model of conceptual bootstrapping in human cognition. *Nat. Hum. Behav.* **8**, 125–136 (2024).
65. Gong, T., Gerstenberg, T., Mayrhofer, R. & Bramley, N. R. Active causal structure learning in continuous time. *Cogn. Psychol.* **140**, 101542 (2023).
66. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learning Motiv.* **24**, 109–165 (1989).
67. French, R. M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).
68. Hadsell, R., Rao, D., Rusu, A. A. & Pascanu, R. Embracing change: continual learning in deep neural networks. *Trends Cogn. Sci.* **24**, 1028–1040 (2020).
69. Beukers, A. O. et al. Blocked training facilitates learning of multiple schemas. *Commun. Psychol.* **2**, 28 (2024).
70. Zhou, Z., Singh, D., Tandoc, M. C. & Schapiro, A. C. Building integrated representations through interleaved learning. *J. Exp. Psychol. Gen.* **152**, 2666–2684 (2023).
71. Birnbaum, M. S., Kornell, N., Bjork, E. L. & Bjork, R. A. Why interleaving enhances inductive learning: the roles of discrimination and retrieval. *Mem. Cogn.* **41**, 392–402 (2013).
72. Zulkiply, N. & Burt, J. S. The exemplar interleaving effect in inductive learning: moderation by the difficulty of category discriminations. *Mem. Cogn.* **41**, 16–27 (2013).
73. 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. Cogn.* **42**, 481–495 (2014).
74. Chater, N. & Oaksford, M. Programs as causal models: speculations on mental programs and mental representation. *Cogn. Sci.* **37**, 1171–1191 (2013).
75. Icard, T. F. From programs to causal models. In *Proc. 21st Amsterdam Colloquium* (eds Cremers, A., van Gessel, T. & Roelofsen, F.) 35–44 (2017).
76. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* **23**, R764–R773 (2013).
77. Baldi, P. & Itti, L. Of bits and wows: a Bayesian theory of surprise with applications to attention. *Neural Netw.* **23**, 649–666 (2010).
78. 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).
79. Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2**, 230–247 (2010).
80. Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychol. Rev.* **101**, 53–79 (1994).
81. 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).
82. Piray, P. & Daw, N. D. A model for learning based on the joint estimation of stochasticity and volatility. *Nat. Commun.* **12**, 6587 (2021).
83. Modirshanechi, A., Brea, J. & Gerstner, W. A taxonomy of surprise definitions. *J. Math. Psychol.* **110**, 102712 (2022).
84. 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).
85. Gershman, S. J., Radulescu, A., Norman, K. A. & Niv, Y. Statistical computations underlying the dynamics of memory updating. *PLoS Comput. Biol.* **10**, e1003939 (2014).
86. 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).

87. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
88. 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).
89. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1511.05952 (2015).
90. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
91. Aljundi, R. et al. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) **32**, 11849–11860 (2019).
92. Botvinick, M. et al. Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
93. 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).
94. Shin, H., Lee, J. K., Kim, J., & Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) **30**, 2990–2999 (2017).
95. 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).
96. 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).
97. 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).
98. Loftus, E. F. Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learn. Mem.* **12**, 361–366 (2005).
99. Roese, N. J. & Vohs, K. D. Hindsight bias. *Persp. Psychol. Sci.* **7**, 411–426 (2012).
100. Carpenter, A. C. & Schacter, D. L. Flexible retrieval: when true inferences produce false memories. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 335–349 (2017).
101. Baldassano, C. et al. Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
102. Davachi, L. & DuBrow, S. How the hippocampus preserves order: the role of prediction and context. *Trends Cogn. Sci.* **19**, 92–99 (2015).
103. Zheng, J. et al. Neurons detect cognitive boundaries to structure episodic memories in humans. *Nat. Neurosci.* **25**, 358–368 (2022).
104. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. Event perception: a mind–brain perspective. *Psychol. Bull.* **133**, 273–293 (2007).
105. Butz, M. V., Achimova, A., Bilkey, D. & Knott, A. Event-predictive cognition: a root for conceptual human thought. *Top. Cogn. Sci.* **13**, 10–24 (2021).
106. 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).
107. Zhou, H., Bamler, R., Wu, C. M. & Tejero-Cantero, Á. Predictive, scalable and interpretable knowledge tracing on structured domains. In *Intl Conf. Learning Representations 2024* https://openreview.net/forum?id=NgaLU2fP5D (2024).
108. Goldwater, M. B. & Schalk, L. Relational categories as a bridge between cognitive and educational research. *Psychol. Bull.* **142**, 729–757 (2016).
109. Denervaud, S., Christensen, A. P., Kenett, Y. N. & Beaty, R. E. Education shapes the structure of semantic memory and impacts creative thinking. *npj Sci. Learn.* **6**, 35 (2021).
110. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The child as hacker. *Trends Cogn. Sci.* **24**, 900–915 (2020).
111. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
112. Ellis, K. et al. Dreamcoder: bootstrapping inductive program synthesis with wake–sleep library learning. *Proc. 42nd ACM SIGPLAN Int. Conf. Programming Language and Implementation (PLDI)* 835–850 (ACM, 2021).
113. 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).
114. 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).
115. 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).
116. Laney, C. & Loftus, E. F. Emotional content of true and false memories. *Memory* **16**, 500–516 (2008).
117. Christianson, S.-Å. Emotional stress and eyewitness memory: a critical review. *Psychol. Bull.* **112**, 284–309 (1992).
118. Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse control tasks through world models. *Nature* **640**, 647–653 (2025).
119. Wayne, G. et al. Unsupervised predictive memory in a goal-directed agent. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1803.10760 (2018).
120. Emanuel, A. & Eldar, E. Emotions as computations. *Neurosci. Biobehav. Rev.* **144**, 104977 (2023).
121. 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).
122. van der Kolk, B. A. & Fisler, R. Dissociation and the fragmentary nature of traumatic memories: overview and exploratory study. *J. Traum. Stress* **8**, 505–525 (1995).
123. Ehlers, A. & Clark, D. M. A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* **38**, 319–345 (2000).
124. Bramley, N. R., Zhao, B., Quillien, T. & Lucas, C. G. Local search and the evolution of world models. *Top. Cogn. Sci.* https://doi.org/10.1111/tops.12703 (2023).
125. Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. Uncovering mental representations with Markov chain Monte Carlo. *Cogn. Psychol.* **60**, 63–106 (2010).
126. Sims, C. R., Jacobs, R. A. & Knill, D. C. An ideal observer analysis of visual working memory. *Psychol. Rev.* **119**, 807–830 (2012).
127. Gershman, S. J. in *The Oxford Handbook of Human Memory* (eds Kahana, M. & Wagner, A.) 1505–1520 (Oxford Univ. Press, 2024).
128. Baddeley, A. D. Language habits, acoustic confusability, and immediate memory for redundant letter sequences. *Psychon. Sci.* **22**, 120–121 (1971).
129. Gobet, F. & Simon, H. A. Recall of rapidly presented random chess positions is a function of skill. *Psychon. Bull. Rev.* **3**, 159–163 (1996).
130. Anderson, J. R. & Milson, R. Human memory: an adaptive perspective. *Psychol. Rev.* **96**, 703–719 (1989).
131. Anderson, J. R. & Schooler, L. J. Reflections of the environment in memory. *Psychol. Sci.* **2**, 396–408 (1991).
132. Roediger, H. L. 3rd & DeSoto, K. A. Cognitive psychology. Forgetting the presidents. *Science* **346**, 1106–1109 (2014).
133. 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).
134. Toglia, M. P., Neuschatz, J. S. & Goodwin, K. A. Recall accuracy and illusory memories: when more is less. *Memory* **7**, 233–256 (1999).
135. 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. Cogn.* **30**, 1054–1064 (2002).
136. 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. In *Proc. 38th Ann. Conf. Cognitive Science Society* (eds. Papafragou, A., et al.) 2273–2278 (Cognitive Science Society, 2016).
137. Bates, C. J., Lerch, R. A., Sims, C. R. & Jacobs, R. A. Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *J. Vis.* **19**, 11 (2019).
138. Yoo, A. H., Klyszejko, Z., Curtis, C. E. & Ma, W. J. Strategic allocation of working memory resource. *Sci. Rep.* **8**, 16162 (2018).
139. Alemi, A. A. Variational predictive information bottleneck. *iProc. Machine Learning Res.* **118**, 1–6 (2020).
140. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
141. Gelman, A. et al. *Bayesian Data Analysis* (Chapman and Hall/CRC, 2013).
142. Grunwald, P. D. *The Minimum Description Length Principle* (MIT Press, 2007).
143. 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).
144. Fränken, J.-P., Theodoropoulos, N. C. & Bramley, N. R. Algorithms of adaptation in inductive inference. *Cogn. Psychol.* **137**, 101506 (2022).
145. Rubino, V., Hamidi, M., Dayan, P. & Wu, C. M. Compositionality under time pressure. In *Proc. 45th Ann. Conf. Cognitive Science Society* (eds Goldwater, M., Anggoro, F., Hayes, B. & Ong, D.) (Cognitive Science Society, 2023).
146. Orbán, G., Fiser, J., Aslin, R. N. & Lengyel, M. Bayesian learning of visual chunks by human observers. *Proc. Natl Acad. Sci. USA* **105**, 2745–2750 (2008).
147. Austerweil, J. L., Sanborn, S. & Griffiths, T. L. Learning how to generalize. *Cogn. Sci.* **43**, e12777 (2019).
148. 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).
149. Shin, Y. S. & DuBrow, S. Structuring memory through inference-based event segmentation. *Top. Cogn. Sci.* **13**, 106–127 (2021).
150. Choi, Y., El-Khamy, M. & Lee, J. Variable rate deep image compression with a conditional autoencoder. In *2019 IEEE/CVF Int. Conf. Computer Vision (ICCV)* (IEEE, 2019).
151. Yang, Y., Bamler, R. & Mandt, S. Variational Bayesian quantization. In *Proc. 37th Intl Conf. Machine Learning (ICML 2020)* **119**, 10670–10680 (2020).
152. Bae, J. et al. Multi-rate VAE: train once, get the full rate-distortion curve. In *Intl Conf. Learning Representations 2023* https://openreview.net/forum?id=OJ8aSjCaMNK (2023).
153. Gregor, K., Besse, F., Rezende, D. J., Danihelka, I. & Wierstra, D. Towards conceptual compression. In *Advances in Neural Information Processing Systems* (eds. Lee, D. et al.) **29**, 3549–3557 (2016).
154. Maaløe, L., Fraccaro, M., Liévin, V. & Winther, O. BIVA: a very deep hierarchy of latent variables for generative modeling. In *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) **32**, 6551–6562 (2019).
155. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *Intl Conf. Learning Representations 2018* https://openreview.net/forum?id=Hk99zCeAb (2018).
156. Child, R. Very deep VAEs generalize autoregressive models and can outperform them on images. In *Intl Conf. Learning Representations 2021* https://openreview.net/forum?id=RLRXCV6DbEJ (2021).
157. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
158. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
159. Freeman, J. & Simoncelli, E. P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).

160. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus* **17**, 1158–1169 (2007).
161. Bányai, M., Nagy, D. G. & Orbán, G. Hierarchical semantic compression predicts texture selectivity in early vision. In *Proc. 2019 Conf. Cognitive Computational Neuroscience* 743–746 (2019).
162. Brady, T. F., Robinson, M. M. & Williams, J. R. Noisy and hierarchical visual memory across timescales. *Nat. Rev. Psychol.* **3**, 147–163 (2024).
163. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).
164. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* 4401–4410 (2019).

Acknowledgements
The authors thank N. R. Bramley for insightful comments and suggestions, which helped to improve this manuscript, and P. Dayan for valuable discussions and extensive comments on earlier drafts. Additionally, the authors thank C. Frater, R. Uchiyama and M. Banyai for helpful feedback on the manuscript and B. Meszena for help with figures. This work is supported by the Humboldt Foundation, the German Federal Ministry of Education and Research (BMBF), by the Tübingen AI Center (FKZ 01IS18039A) funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy (EXC2064/1–390727645), and by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2117 422037984). G.O. was supported by a grant from the National Research, Development and Innovation Office (grant ADVANCED 150361) and by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory in Hungary.

Author contributions
D.G.N. and G.O. researched data for the article. All authors contributed substantially to discussion of the content. D.G.N. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Reviews Psychology* thanks Neil R. Bramley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.